ed68fd4b85b778e0f0b1bfea5dbc56cf4148f1229917e799a70e51517ff6d591

Автономная некоммерческая образовательная организация высшего образования
«Европейский университет в Санкт-Петербурге»

Международная школа искусств и культурного наследия

УТВЕРЖДАЮ:

Ректор_____________________В.В. Волков

«29» мая 2024 г.

Протокол УС № 5 от 29.05 2024 г.

Рабочая программа дисциплины

Инструменты обработки естественного языка

образовательная программа
направление подготовки
51.04.04 Музеология и охрана объектов культурного и природного наследия

направленность (профиль)
«Музейные исследования и кураторские стратегии»
программа подготовки – магистратура

язык обучения – русский
форма обучения – очная

квалификация выпускника
Магистр

Санкт-Петербург

Автор:
Лашманов О.Ю., к.т.н., научный руководитель лаборатории «Искусство и искусственный интеллект» Международной школы искусств и культурного наследия АНООВО «ЕУСПб»

Рецензент
Басс В. Г., кандидат искусствоведения, доцент Международной школы искусств и культурного наследия АНОВО «ЕУСПб»

Рабочая программа дисциплины **«Инструменты обработки естественного языка»**, входящей в состав основной профессиональной образовательной программы высшего образования — программы магистратуры «Музейные исследования и кураторские стратегии», утверждена на заседании Совета Международной школы искусств и культурного наследия.

Протокол заседания № 12 от 14 мая 2024 года.

АННОТАЦИЯ РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ
«Инструменты обработки естественного языка»

Дисциплина **«Инструменты обработки естественного языка»** является дисциплиной по выбору части, формируемой участниками образовательных отношений, образовательной программы, Блока 1 «Дисциплины (модули)» образовательной программы «Музейные исследования и кураторские стратегии» по направлению подготовки 51.04.04 Музеология и охрана объектов культурного и природного наследия.

Курс **«Инструменты обработки естественного языка»** знакомит магистрантов с основными методами и подходами к обработке естественного языка, в ходе изучения дисциплины проводится анализ принципов оценки качества методов обработки естественного языка, а также магистрантам предоставляется возможность овладеть практическими навыками обработки больших коллекций текстов.

Программой дисциплины предусмотрены следующие виды контроля: текущий контроль успеваемости, промежуточный контроль в форме зачета с оценкой (в конце 3 семестра).

Общая трудоемкость освоения дисциплины составляет 2 зачетных единицы, 72 часа.

Содержание

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Цели освоения дисциплины «**Инструменты обработки естественного языка**» - знакомство с основными методами и приложениями автоматической обработки естественного языка (ОЕЯ), получение практических навыков работы с инструментами ОЕЯ.

Задачи:

1. Знакомство с основными методами и подходами к ОЕЯ.
2. Понимание принципов оценки качества методов ОЕЯ.
3. Овладение практическими навыками обработки больших коллекций текстов.

2. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ

В результате изучения учебной дисциплины обучающийся должен овладеть следующими компетенциями: универсальными (УК) и профессиональными (ПК). Планируемые результаты формирования компетенций и индикаторы их достижения в результате освоения дисциплины представлены в Таблице 1.

Таблица 1

Планируемые результаты освоения дисциплины, соотнесенные с индикаторами достижения компетенций обучающихся

Код и наименование компетенции	Индикаторы достижения компетенции	Результаты обучения (знать, уметь, владеть)
УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	ИД.УК-1.1. Знать методологию и методику системного и критического анализа проблемных ситуаций, стратегического управления ИД.УК-1.2. Уметь осуществлять системный и критический анализ проблемных ситуаций, вырабатывать стратегию действий ИД.УК-1.3. Уметь обосновывать, формулировать и решать задачи, возникающие в процессе профессиональной деятельности ИД.УК-1.4. Владеть методами системного и критического анализа, стратегического управления	Знать: методы научного познания, в основе которых лежит рассмотрение объекта как системы: целостного комплекса взаимосвязанных элементов, методы и модели стратегического планирования З (УК-1)
		Уметь: с использованием методов системного подхода анализировать альтернативные варианты решения исследовательских задач, вырабатывать стратегию действий и оценивать эффективность реализации стратегических планов У (УК-1)
		Владеть: целостной системой навыков методологического использования системного подхода при решении проблем, возникающих при выполнении исследовательских работ, навыками отстаивания своей точки зрения при выработке стратегических планов выполнения исследовательских работ В (УК-1)
ПК-3 Способен использовать современные методы обработки и интерпретации информации в профессиональной сфере	ИД.ПК-3.1. Знать современные методы накопления, обработки, передачи, поиска и использования информации о культурном и природном наследии ИД.ПК-3.2. Уметь вести результативный поиск информации с использованием современных информационно-коммуникационных технологий ИД.ПК-3.4. Уметь обрабатывать, анализировать и использовать	Знать: принципы и методы ведения самостоятельных научных исследований в профессиональной области и смежных областях З (ПК-3)
		Уметь: выстраивать последовательную работу с информацией по актуальным проблемам сохранения культурного и природного наследия с использованием современных

Код и наименование компетенции	Индикаторы достижения компетенции	Результаты обучения (знать, уметь, владеть)
	информацию в соответствии с научными и познавательными задачами ИД.ПК-3.5. Владеть приемами использования научной информации в научно-исследовательской и профессиональной деятельности ИД.ПК-3.6. Владеть основными способами поиска и представления информации	информационно-коммуникационных технологий У (ПК-3)
		Владеть: навыками самостоятельного проведения научных исследований в сфере профессиональных интересов В (ПК-3)

В результате освоения дисциплины обучающийся должен:

знать: современные методологические принципы и методические приемы исторического исследования;

уметь: выявлять различия в методологических принципах и методических приемах исторического исследования; использовать на практике различные методики работы;

владеть: навыками применения современных методических приемов исторического исследования.

3. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Дисциплина **«Инструменты обработки естественного языка»** является дисциплиной по выбору части, формируемой участниками образовательных отношений, Блока 1 «Дисциплины (модули)» учебного плана основной профессиональной образовательной программы высшего образования — программы магистратуры «Музейные исследования и кураторские стратегии» по направлению подготовки 51.04.04 Музеология и охрана объектов культурного и природного наследия. Код дисциплины по учебному плану Б1.В.ДВ.01.01.03. Курс читается в третьем семестре, форма промежуточной аттестации — зачет с оценкой.

Для успешного освоения материала данной дисциплины требуются знания, умения и навыки, полученные в ходе изучения бакалавриата/специалитета.

Знания, умения и навыки, полученные при освоении данной дисциплины, применяются магистрантами в процессе выполнения научно-исследовательской работы и подготовки к защите и защиты выпускной квалификационной работы.

4. ОБЪЕМ ДИСЦИПЛИНЫ

Общая трудоемкость освоения дисциплины составляет 2 зачетных единиц, 72 часа.

Таблица 2

Объем дисциплины

Типы учебных занятий и самостоятельная работа		Объем дисциплины				
		Всего	Семестр			
			1	2	3	4
Контактная работа обучающихся с преподавателем в соответствии с УП:		**28**	-	-	28	-
Лекции (Л)		**14**	-	-	14	-
Семинарские занятия (СЗ)		**14**	-	-	14	-
Самостоятельная работа (СР)		**44**	-	-	44	-
Промежуточная аттестация	**форма**	**Зачет с оценкой**	-	-	Зачет с оценкой	-
	час.	**-**	-	-	-	-
Общая трудоемкость дисциплины (час./з.е.)		**72/2**	-	-	72/2	-

5. СОДЕРЖАНИЕ И СТРУКТУРА ДИСЦИПЛИНЫ

Содержание дисциплины соотносится с планируемыми результатами обучения по дисциплине: через задачи, формируемые компетенции и их компоненты (знания, умения, навыки – далее ЗУВ) посредством индикаторов достижения компетенций в соответствии с Таблицей 3.

5.1 Содержание дисциплины

Таблица 3

Содержание дисциплины

№ п/п	**Наименование тем (разделов)**	**Содержание тем (разделов)**	**Коды компетенций**	**Индикаторы компетенций** (в соотв. с Таблицей 1)	**Коды ЗУВ** (в соответствии с Таблицей 1)
1	Введение	Краткая история дисциплины, основные инструменты и приложения	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)
2	Морфологический анализ	Основные подходы, данные, инструменты.	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)
3	Языковые модели	Приложения, данные, оценки качества. Языковые модели на основе n-грамм, сглаживание. Нейронные языковые модели.	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)
4	Анализ тональности	Постановка задачи, приложения, данные, методы решения	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)

№ п/п	Наименование тем (разделов)	Содержание тем (разделов)	Коды компетенций	Индикаторы компетенций (в соотв. с Таблицей 1)	Коды ЗУВ (в соответствии с Таблицей 1)
5	Синтаксический анализ	Формализмы для представления синтаксической структуры, данные, подходы к решению, оценка качества.	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)
6	Извлечение информации	Именованные сущности, данные для обучения и тестирования. Методы на основе машинного обучения; методы на основе рекуррентных нейронных сетей.	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)
7	Автоматическое реферирование	Различные постановки задачи и методы, данные для обучения, автоматическое тестирование на основе метрики ROUGE, нейронные сети для автоматического реферирования.	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)
8	Вычислительная семантика	Семантические ресурсы (WordNet), дистрибутивные семантические модели, контекстуализированные векторные представления.	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)
9	Ответы на вопросы	Понимание прочитанного (reading comprehension) и ответы по базе знаний. Методы и данные.	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)

№ п/п	Наименование тем (разделов)	Содержание тем (разделов)	Коды компетенций	Индикаторы компетенций (в соотв. с Таблицей 1)	Коды ЗУВ (в соответствии с Таблицей 1)
10	Машинный перевод (МП)	На основе правил, статистический МП, нейронный МП. Оценка качества (BLEU и другие метрики).	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)

5.2 Структура дисциплины

Таблица 4

Структура дисциплины

№ п/п	Наименование тем (разделов)	Объем дисциплины, час.				Форма текущего контроля успеваемости*, промежуточной аттестации
		Всего	Контактная работа обучающихся с преподавателем по типам учебных занятий в соответствии с УП		СР	
			Л	СЗ		
Очная форма обучения						
Тема 1	Введение	7	1	2	4	О, Д
Тема 2	Морфологический анализ	7	1	2	4	КЗ, О
Тема 3	Языковые модели	7	2	1	4	КЗ
Тема 4	Анализ тональности	7	2	1	4	ПЗ
Тема 5	Синтаксический анализ	7	2	1	4	КЗ, ПЗ
Тема 6	Извлечение информации	7	1	2	4	ПЗ
Тема 7	Автоматическое реферирование	7	1	2	4	КЗ
Тема 8	Вычислительная семантика	6	1	1	4	ПЗ, Д
Тема 9	Ответы на вопросы	6	1	1	4	ПЗ
Тема 10	Машинный перевод (МП)	11	2	1	8	КЗ
Промежуточная аттестация		**-**	**-**	**-**	**-**	**Зачет с оценкой**
ИТОГО:		**72/2**	**14**	**14**	**44**	**-**

**Примечание: формы текущего контроля успеваемости: диспут (Д), опрос (О), практическое задание (ПЗ), контрольное задание (КЗ).*

6. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ

6.1 Общие положения

Знания и навыки, полученные в результате лекций и семинарских занятий, закрепляются и развиваются в результате повторения материала, усвоенного в аудитории, путем чтения текстов и исследовательской литературы (из списков основной и дополнительной литературы) и их анализа.

Самостоятельная работа является важнейшей частью процесса высшего образования. Ее следует осознанно организовать, выделив для этого необходимое время и соответственным образом организовав рабочее пространство. Важнейшим элементом самостоятельной работы является проработка материалов прошедших занятий (анализ конспектов, чтение рекомендованной литературы) и подготовка к следующим

лекциям/семинарам. Литературу, рекомендованную в программе курса, следует, по возможности, читать в течение всего семестра, концентрируясь на обусловленных программой курса темах.

Существенную часть самостоятельной работы магистранта представляет самостоятельное изучение учебно-методических изданий, лекционных конспектов, интернет-ресурсов и пр. Подготовка к семинарским занятиям, опросам также является важной формой работы магистранта. Самостоятельная работа может вестись как индивидуально, так и при содействии преподавателя.

6.2 Рекомендации по распределению учебного времени по видам самостоятельной работы и разделам дисциплины

Тема 1. Введение:

1.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 2 часа.

1.2. Подготовка к лабораторным занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций – 2 часа. Итого: 4 часа.

Тема 2. Морфологический анализ:

2.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 2 часа.

2.2. Подготовка к лабораторным занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций – 2 часа. Итого: 4 часа.

Тема 3. Языковые модели:

3.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 2 часа.

3.2. Подготовка к лабораторным занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций – 2 часа. Итого: 4 часа.

Тема 4. Анализ тональности:

4.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 2 часа.

4.2. Подготовка к лабораторным занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций – 2 часа. Итого: 4 часа.

Тема 5. Синтаксический анализ:

5.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 2 часа.

5.2. Подготовка к лабораторным занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций – 2 часа. Итого: 4 часа.

Тема 6. Извлечение информации:

6.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 2 часа.

6.2. Подготовка к лабораторным занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций – 2 часа. Итого: 4 часа.

Тема 7. Автоматическое реферирование:

7.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 2 часа.

7.2. Подготовка к лабораторным занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций – 2 часа. Итого: 4 часа.

Тема 8. Вычислительная семантика:

8.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 2 часа.

8.2. Подготовка к лабораторным занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций – 2 часа. Итого: 4 часа.

Тема 9. Ответы на вопросы:

9.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 2 часа.

9.2. Подготовка к лабораторным занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций – 2 часа. Итого: 4 часа.

Тема 10. Машинный перевод (МП):

10.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 2 часа.

10.2. Подготовка к лабораторным занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций – 2 часа. Итого: 4 часа.

6.3 Перечень основных вопросов по изучаемым темам для самостоятельной работы обучающихся по дисциплине

Вопросы для самостоятельной подготовки по темам дисциплины:

1. Исправление опечаток.
2. Распознавание звучащей речи.
3. Морфологический анализ: основные подходы.
4. Нейронные языковые модели.
5. Анализ тональности: постановка задачи, приложения, данные, методы решения.
6. Синтаксический анализ: формализмы для представления синтаксической структуры, данные, подходы к решению, оценка качества.

7. Методы извлечения информации на основе машинного обучения.
8. Нейронные сети для автоматического реферирования.
9. Дистрибутивные семантические модели.
10. Машинный перевод: методы оценки качества.

6.4 Перечень литературы для самостоятельной работы

1. Ганегедара, Т. Обработка естественного языка с TensorFlow : монография / Т. Ганегедара ; пер. с анг. В. С. Яценкова. - Москва : ДМК Пресс, 2020. - 382 с. - ISBN 978-5-97060-756-5. - Текст : электронный. - URL: https://znanium.com/catalog/product/1094940 .
2. Дуглас А., Л. Люк, Д.А. Анализ сетей (графов) в среде R. Руководство пользователя / Д.А. Люк ; пер. с анг. А.В. Груздева. - Москва : ДМК Пресс, 2017. - 250 с. - ISBN 978-5-97060-428-1. - Текст : электронный. - URL: https://znanium.com/catalog/product/1028090
3. Иванова, Н.Ю. Системное и прикладное программное обеспечение : учебное пособие / Н.Ю. Иванова, В.Г. Маняхина ; Московский педагогический государственный университет. – Москва : Прометей, 2011. – 202 с. : ил.,табл., схем. – Режим доступа: по подписке. – URL: http://biblioclub.ru/index.php?page=book&id=105792
4. Кравцова, Е. Д. Логика и методология научных исследований [Электронный ресурс] : учеб. пособие / Е. Д. Кравцова, А. Н. Городищева. – Красноярск : Сиб. федер. ун-т, 2014. – 168 с. - ISBN 978-5-7638-2946-4 - Режим доступа: http://znanium.com/catalog.php?bookinfo=507377

6.5 Перечень учебно-методического обеспечения для самостоятельной работы

Для обеспечения самостоятельной работы магистрантов по дисциплине «Инструменты обработки естественного языка» разработано учебно-методическое обеспечение в составе:

1. Контрольные задания для подготовки к процедурам текущего контроля (п. 7.2 Рабочей программы).
2. Типовые задания для подготовки к промежуточной аттестации (п. 7.4 Рабочей программы).
3. Рекомендуемые основная, дополнительная литература, Интернет-ресурсы и справочные системы (п. 8, 9 Рабочей программы).
4. Рабочая программа дисциплины размещена в электронной информационно-образовательной среде Университета на электронном учебно-методическом ресурсе АНООВО «ЕУСПб» — образовательном портале LMS Sakai — Sakai@EU.

7. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕЙ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ

7.1 Показатели, критерии и оценивание компетенций и индикаторов их достижения в процессе текущей аттестации

Информация о содержании и процедуре текущего контроля успеваемости, методике оценивания знаний, умений и навыков обучающегося в ходе текущего контроля доводятся научно-педагогическими работниками Университета до сведения обучающегося на первом занятии по данной дисциплине.

Текущий контроль предусматривает подготовку магистрантов к каждому лабораторному занятию, участие в опросах, диспутах, подготовку практических заданий, выполнение контрольных заданий, активное слушание на лекциях. Магистрант должен присутствовать на семинарских занятиях, отвечать на поставленные вопросы, показывая, что прочитал разбираемую литературу, представлять содержательные реплики по обсуждаемым вопросам.

Текущий контроль проводится в форме устных опросов и оценивания участия магистрантов в проходящих диспутах, оценивания выполненных практических заданий, контрольных работ, демонстрирующих степень знакомства с дополнительной литературой.

Таблица 5

Показатели, критерии и оценивание компетенций и индикаторов их достижения в процессе текущей аттестации

Наименование темы (раздела)	Код компетенции	Индикаторы компетенций	Коды ЗУВ (в соотв. с табл. 1)	Формы текущего контроля	Результаты текущего контроля
Введение	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Опрос 1 Диспут 1	зачтено/ не зачтено зачтено/ не зачтено
Морфологический анализ	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Контрольное задание 1 Контрольное задание 2 Опрос 2 Контрольное задание 3	зачтено/ не зачтено зачтено/ не зачтено зачтено/ не зачтено зачтено/ не зачтено
Языковые модели	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Контрольное задание 4 Контрольное задание 5 Контрольное задание 6	зачтено/ не зачтено зачтено/ не зачтено зачтено/ не зачтено
Анализ тональности	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Практическое задание 1	зачтено/ не зачтено
Синтаксический анализ	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Контрольное задание 7 Контрольное задание 8 Практическое задание 2	зачтено/ не зачтено зачтено/ не зачтено зачтено/ не зачтено
Извлечение информации	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3.	З (УК-1) У (УК-1) В (УК-1)	Практическое задание 3	зачтено/ не зачтено

Наименование темы (раздела)	Код компетенции	Индикаторы компетенций	Коды ЗУВ (в соотв. с табл. 1)	Формы текущего контроля	Результаты текущего контроля
		ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (ПК-3) У (ПК-3) В (ПК-3)		
Автоматическое реферирование	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Контрольное задание 9	зачтено/ не зачтено
Вычислительная семантика	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Диспут 2 Практическое задание 4	зачтено/ не зачтено зачтено/ не зачтено
Ответы на вопросы	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Практическое задание 5	зачтено/ не зачтено зачтено/ не зачтено
Машинный перевод (МП)	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Контрольное задание 10	зачтено/ не зачтено

Таблица 6

Критерии оценивания

Формы текущего контроля успеваемости	Критерии оценивания
Опрос	Ответ отсутствует или является односложным, или содержит существенные ошибки – не зачтено Магистрант в ответах демонстрирует знание всех теоретических положений, (развернуто) отвечает на все поставленные вопросы, предлагает обоснования при ответе на все или большинство поставленных вопросов; несущественные ошибки не снижают качество ответа — зачтено

Формы текущего контроля успеваемости	Критерии оценивания
Диспут	Пассивность, участие без представления аргументов и обоснования точки зрения, несформированность навыков профессиональной коммуникации в группе — не зачтено Представление аргументированной научной позиции, обоснование точки зрения в диспуте, демонстрация навыков профессиональной коммуникации в группе — зачтено
Практическое задание	магистрант выполняет задание частично или с существенными недочетами (некорректно сформулирован исследовательский вопрос, не определены основные агенты, некорректно выбраны методы исследования, требования к содержанию, структуре, логике, аргументации, оформлению не выполнены) – не зачтено, полное и правильное выполнение задания в соответствии с требованиями к содержанию, структуре, логике, аргументации, оформлению с возможным небольшим количеством погрешностей (например, плохо выдержанная структура текста, недостаточная аргументация отдельных тезисов) – зачтено
Контрольное задание	магистрант выполняет задание частично или с существенными недочетами (некорректно сформулирован исследовательский вопрос, не определены основные агенты, некорректно выбраны методы исследования, требования к содержанию, структуре, логике, аргументации, оформлению не выполнены) – не зачтено, полное и правильное выполнение задания в соответствии с требованиями к содержанию, структуре, логике, аргументации, оформлению с возможным небольшим количеством погрешностей (например, плохо выдержанная структура текста, недостаточная аргументация отдельных тезисов) – зачтено

7.2. Контрольные задания для текущей аттестации

Примерный материал опросов, диспутов, практических заданий:

Тема 1. Введение.

Опрос 1.

1. Какие основные инструменты и приложения используются при обработке естественного языка?
2. История развития методов обработки естественного языка?

Диспут 1.

Как технологии обработки естественного языка меняют повседневную жизнь людей.

Тема 2. Морфологический анализ.

Контрольное задание 1.

Три слова в списке имеют одинаковый тип словоизменения, одно – отличается. Отметьте это слово.

1) чело
2) табло
3) сверло
4) кайло.

Контрольное задание 2.

Какие термины относятся к морфологической обработке? Выберите номера правильных ответов:

1) лемматизация;
2) лемминг;
3) стемминг;
4) прокрастинация;
5) парсинг.

Опрос 2.

1. Каковы основные подходы к морфологическому анализу?
2. Какова основная методика отбора данных для морфологического анализа?
3. Каковы основные инструменты морфологического анализа?

Контрольное задание 3.

Отметьте правильный набор граммем (в нотации НКРЯ/mystem) для выделенного слова в предложении:

Лично я ловлю покемонов не одобряю.

1) S,жен,неод=вин,ед
2) V,несов,пе=непрош,ед,изъяв,1-л
3) V,несов,пе=ед,пов,2-л
4) S,муж,неод=им,ед.

Тема 3. Языковые модели.

Контрольное задание 4.

Постройте биграммную языковую модель на основе корпуса:

<s> Вася любит мороженое </s>
<s> Лена любит малину </s>
<s> Вася любит Лену </s>
<s> Георгий ест мороженое </s>
<s> Лена рисует яблоко </s>
<s> Георгий любит Катю </s>
<s> Георгий любит смотреть, как Лена ест мороженое </s>

Упорядочите предложения по убыванию оценок вероятностей на основе построенной языковой модели.

1. <s> Лена любит мороженое </s>
2. <s> Лена рисует малину </s>
3. <s> Вася любит Катю </s>.

Контрольное задание 5.

Постройте биграммную языковую модель на основе корпуса:

<s> Вася любит мороженое </s>
<s> Лена любит малину </s>
<s> Вася любит Лену </s>
<s> Георгий ест мороженое </s>
<s> Лена рисует яблоко </s>
<s> Георгий любит Катю </s>
<s> Георгий любит смотреть, как Лена ест мороженое </s>

Вычислите перплексию модели на предложении:

<s> Георгий любит малину </s>

Обратите внимание, что для вычисления вероятности предложения (и, соответственно, перплексии) используются вероятности 4-х биграмм (N=4 в формуле вычисления перплексии).

Контрольное задание 6.

Пусть у нас есть корпус, содержащий 10000 предложений, размер словаря – 1500 уникальных слов (включая специальные «слова» – маркеры начала и конца предложений). Некоторые частоты униграмм:

ем 100
дуриан 1
и 5000
не 3000
морщусь 50

и биграмм:
<s> ем 20
ем дуриан 0
дуриан и 0
и не 300
не морщусь 15
морщусь </s> 5.

Примените сглаживание Лапласа (α=1, сглаживание «+1») для оценки вероятностей биграмм и оцените на их основе вероятность предложения <s> ем дуриан и не морщусь </s>

В качестве ответа введите натуральный логарифм оценки вероятности предложения.

Тема 4. Анализ тональности.

Практическое задание 1

Реализуйте анализатор тональности одним из способов. Данные – английские предложений из отзывов о фильмах из Stanford Sentiment Treebank (из датасета взяты только целые предложения; 5-уровневая разметка приведена к трехуровневой – негативный, нейтральный, позитивный). Вы можете использовать подход на основе словаря тонально окрашенных слов (например, SentiWords) или обучить классификатор на тренировочных данных. Вы можете использовать тренировочные данные и в первом случае – чтобы подобрать пороги для классификации предложений на основе весов словаря. Постройте матрицу ошибок (confusion matrix) на тестовом наборе. Оцените правильность (accuracy, доля правильно классифицированных предложений) классификатора на тестовом наборе. Проанализируйте неверно классифицированные предложения, сделайте предположения о причинах неверной классификации, предложите улучшения.

Тема 5. Синтаксический анализ.

Контрольное задание 7.

Выберите правильный разбор на составляющие предложения:

Советник губернатора Чукотского автономного округа Романа Абрамовича Роман Копин победил на повторных выборах главы администрации Чаунского района

1. [[[Советник [[губернатора [Чукотского [автономного округа]]] [Романа Абрамовича]]] [Роман Копин]] [победил [на [повторных [выборах [главы [администрации [Чаунского района]]]]]]]
2. [[[Советник [[губернатора [Чукотского [автономного округа]]] [Романа Абрамовича]]] [Роман Копин]] [[[победил на] [повторных [выборах [главы [администрации [Чаунского района]]]]]]]
3. [[[[[Советник губернатора] [Чукотского [автономного округа]]] [Романа Абрамовича]] [Роман Копин]] [победил [на [[повторных выборах] [главы [администрации [Чаунского района]]]]]]
4. [[[Советник [[губернатора [Чукотского [автономного округа]]] [Романа Абрамовича]]] [[Роман Копин] победил]] [на [[повторных выборах] [главы [администрации [Чаунского района]]]]]

[[[Советник [[губернатора Чукотского] [[автономного округа] [Романа Абрамовича]]]] [Роман Копин]] [победил [на [повторных [выборах [главы [администрации [Чаунского района]]]]]]]

Контрольное задание 8.

Выберите правильный разбор предложения в терминах зависимостей:

0 root 1 Активно 2 обсуждается 3 роль 4 нашей 5 страны 6 в 7 современном 8 быстро 9 меняющемся 10 мире, 11 проходящем 12 через 13 переломный 14 этап.

В приведенных ниже вариантах разбора каждая пара – зависимость (хозяин, слуга), знаки препинания не учитываются, используется традиционный подход при установлении зависимостей с участием предлогов (не как в universal dependencies).

1. (0, 3) (3, 2) (2, 1) (3, 5) (5, 4) (3, 6) (6, 10) (10, 7) (10, 9) (9, 8) (10, 11) (11, 12) (12, 14) (14, 13)
2. (0, 2) (2, 1) (2, 3) (3, 5) (5, 4) (3, 6) (6, 10) (10, 7) (10, 9) (9, 8) (10, 11) (11, 12) (12, 14) (14, 13)
3. (0, 2) (2, 1) (2, 3) (3, 5) (5, 4) (5, 6) (6, 10) (10, 7) (10, 9) (9, 8) (10, 11) (11, 12) (12, 14) (14, 13)
4. (0, 2) (2, 1) (2, 3) (3, 5) (5, 4) (3, 6) (6, 7) (7, 8) (8, 9) (9, 10) (10, 11) (11, 12) (12, 14) (14, 13)
5. (0, 2) (2, 3) (3, 5) (5, 4) (3, 6) (6, 10) (10, 7) (10, 9) (2, 8) (10, 11) (11, 1) (11, 12) (12, 14) (14, 13)

Практическое задание 2

Постройте деревья зависимостей для 2000 предложений «Войны и Мира» и «Братьев Карамазовых» с помощью библиотеки Stanza.

Оцените производительность библиотеки (предложения/с).

Проведите ручную оценку качества разбора на случайных 20 предложениях (по 10 из каждой книги).

Посчитайте среднюю глубину дерева разбора для каждого из романов. Посчитайте корреляцию между длиной предложения в словах и глубиной дерева разбора. Оцените на нескольких примерах, насколько глубина дерева соответствует субъективной сложности понимания предложения.

Тема 6. Извлечение информации.

Практическое задание 3

Примените и оцените модуль извлечения именованных сущностей Natasha (https://github.com/natasha/natasha). Для тестирования используйте текст и соответствующую разметку. Тестовые данные содержат только разметку для людей (PER) и организаций (ORG). Рассчитайте F1 для каждого типа сущностей и общее значение F1.

Тема 7. Автоматическое реферирование.

Контрольное задание 9.

Пусть у нас есть два «реферата-образца»:

1: карп лещ лещ щука сазан

2: лещ карп лещ сазан плотва

Упорядочите «рефераты» ниже по убыванию значения ROUGE-2.

1) лещ карп лещ сазан плотва
2) плотва сазан лещ карп сазан
3) щука сазан карп лещ сазан
4) карп карась лещ окунь лещ
5) щука сазан лещ карп сазан

Тема 8. Вычислительная семантика.

Диспут 2.

Возможные негативные последствия использования больших предобученных моделей для генерации текста.

Практическое задание 4

Реализуйте генератор юмора по мотивам работы Alessandro Valitutti et al. “Let Everything Turn Well in Your Wife”: Generation of Adult Humor Using Lexical Constraints, 2013. На входе генератора – новостной заголовок, в котором надо заменить одно слово. Предлагаемый алгоритм:

1. Проведите разбор предложения с помощью библиотеки Stanza. На основе результатов разбора выберете слово-кандидат на замену.
2. Найдите антоним для слова в WordNet (используйте интерфейс библиотеки NLTK).
3. Если антоним не нашелся, то найдите несколько слов, близких по звучанию или рифму с помощью datamuse api.
4. Получите вектора fasttext для начального слова и вариантов замены. Среди этих слов найдите самое далекое по косинусному расстоянию.

Оцените 20 модификаций по шкале от 0 (совсем не смешно) до 3 (очень смешно), приведите среднюю оценку.

Тема 9. Ответы на вопросы.

Практическое задание 5

Обучите модель для выделения ответа на вопрос из параграфа на тренировочной части русскоязычных данных TyDi QA. Оцените качество модели на данных для настройки (development set, EM/F1). Проанализируйте случаи, с которыми модель справилась хуже всего. Сделайте предположение, в чем сложность этих случаев.

Тема 10. Машинный перевод (МП).

Контрольное задание 10

Для фразы

Call me what instrument you will, though you can fret me, you cannot play upon me.

Есть два образцовых перевода (для простоты знаки препинания в образцах и вариантах перевода удалены):

- *назовите меня каким угодно инструментом вы хоть и можете меня терзать но играть на мне не можете*
- *объявите меня каким угодно инструментом вы можете расстроить меня но играть на мне нельзя*

Упорядочите варианты перевода ниже по убыванию BLEU-2 (метрика на основе униграмм и биграмм). Помните, что стандартная метрика BLEU не предполагает лемматизацию текстов.

- *позвони мне на каком инструменте вы будете хотя вы можете беспокоиться меня но вы не можете играть на мне*
- *назовите мне какой инструмент вы хотите хотя можете меня беспокоить но вы не можете играть на меня*
- *позвони мне какой инструмент ты будешь хотя ты можешь меня волновать но ты не можешь играть на меня*
- *назовите меня какой инструмент вы будете хотя вы можете раздражать меня все же вы не можете играть на меня*
- *считай меня чем тебе угодно ты можешь мучить меня но не играть мною*

● *назови меня каким угодно инструментом ты можешь меня расстроить но не играть на мне.*

7.3 Показатели, критерии и оценивание компетенций и индикаторов их достижения в процессе промежуточной аттестации

Форма промежуточной аттестации — зачет с оценкой, выставляемый на основе 2-х письменных работ (эссе).

Перед зачетом с оценкой проводится консультация, на которой преподаватель отвечает на вопросы магистрантов.

В результате промежуточного контроля знаний студенты получают оценку по дисциплине.

Таблица 7

Показатели, критерии и оценивание компетенций и индикаторов их достижения в процессе промежуточной аттестации

Форма промежуточной аттестации/вид промежуточной аттестации	**Коды компетенций**	**Индикаторы компетенций** *(в соотв. с Таблицей 1)*	**Коды ЗУВ** (в соответствии с Таблицей 1)	**Критерии оценивания**	**Оценка**
Зачет с оценкой/ Письменная работа (эссе)	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Эссе соответствует следующим требованиям: сформулирован исследовательский вопрос, корректно выбраны методы и собраны данные, тема раскрыта, соблюдены структура и научный стиль, сформулированы выводы, аргументация убедительна, правильно оформлен библиографический аппарат и т.д., магистрант демонстрирует: глубокое усвоение программного материала, изложение его исчерпывающе, последовательно, четко, умение делать обоснованные выводы, соблюдение норм устной и письменной литературной речи/ Эссе успешно представлено на защите. Магистрант дает правильный ответ на теоретический вопрос, при условии, что отдельные неточности, допускаемые в ходе ответа, никак не снижают общего качества ответа. Для ответа характерно: • глубокое усвоение программного материала, • изложение его исчерпывающе, последовательно, четко,	Зачтено, отлично

Форма промежуточной аттестации/вид промежуточной аттестации	Коды компетенций	Индикаторы компетенций *(в соотв. с Таблицей 1)*	Коды ЗУВ (в соответствии с Таблицей 1)	Критерии оценивания	Оценка
				• умение делать обоснованные выводы, • соблюдение норм устной и письменной литературной речи.	
				В эссе не соблюдены некоторые требования к работе: возможно несоблюдении одного-двух требований и допущении некоторых неточностей, магистрант демонстрирует: твердое знание материала курса, последовательное изложение материала, знание теоретических положений без обоснованной их аргументации, соблюдение норм устной и письменной литературной речи; Эссе успешно представлено на защите. Магистрант верно отвечает на вопрос, указанный в билете, при условии, что ответ на вопрос характеризуется отсутствием серьезных, значимых неточностей, при следующих характеристиках ответа: • твердое знание материала курса, • последовательное изложение материала, • знание теоретических положений без обоснованной их аргументации, • соблюдение норм устной и письменной литературной речи.	Зачтено, хорошо
				Эссе содержит существенные оплошности: нарушено сразу несколько требований, например, выводы плохо обоснованы, есть фактические ошибки, магистрант при защите демонстрирует: знание основного материала, но владение им не в полном объеме,	Зачтено, удовлетворительно

Форма промежуточной аттестации/вид промежуточной аттестации	Коды компетенций	Индикаторы компетенций *(в соотв. с Таблицей 1)*	Коды ЗУВ (в соответствии с Таблицей 1)	Критерии оценивания	Оценка
				допущение существенных неточностей, недостаточно правильных формулировок, допущение нарушения логической последовательности в изложении материала, наличие нарушений норм литературной устной и письменной речи. Эссе представлено на защите. Магистрант представляет правильный ответ на теоретический вопрос, указанный в билете, при условии, что ответ на вопрос характеризуется значительными неточностями, при следующих параметрах ответа: • знание основного материала, но владение им не в полном объеме, • допущение существенных неточностей, недостаточно правильных формулировок, • допущение нарушения логической последовательности в изложении материала, • наличие нарушений норм литературной устной и письменной речи.	
				Представленное эссе не отвечает предъявляемым требованиям (либо не предоставление эссе); имеет место: незнание значительной части программного материала, наличие существенных ошибок в определениях, формулировках, понимании теоретических положений; бессистемность при ответе на поставленный вопрос, отсутствие в ответе логически корректного анализа, аргументации, классификации, наличие нарушений норм устной и письменной литературной речи.	Не зачтено, не удовлетворительно

Форма промежуточной аттестации/вид промежуточной аттестации	Коды компетенций	Индикаторы компетенций *(в соотв. с Таблицей 1)*	Коды ЗУВ (в соответствии с Таблицей 1)	Критерии оценивания	Оценка
				Магистрант представляет ответ на вопрос, характеризующийся наличием существенных ошибок в определениях, формулировках, понимании теоретических положений, свидетельствующий о некомпетентности магистранта, при следующих параметрах ответа: • незнание значительной части программного материала, • наличие существенных ошибок в определениях, формулировках, понимании теоретических положений; • бессистемность при ответе на поставленный вопрос, • отсутствие в ответе логически корректного анализа, аргументации, классификации, наличие нарушений норм устной и письменной литературной речи.	

Результаты сдачи промежуточной аттестации по направлениям подготовки уровня магистратуры оцениваются по стобалльной системе оценки в соответствии с Положением о формах, периодичности и порядке организации и проведения текущего контроля успеваемости и промежуточной аттестации обучающихся в АНООВО «ЕУСПб» следующим образом согласно таблице 7а.

Таблица 7а

Система оценки знаний обучающихся

Пятибалльная (стандартная) система	Стобалльная система оценки	Бинарная система оценки
5 (отлично)	100-81	зачтено
4 (хорошо)	80-61	
3 (удовлетворительно)	60-41	
2 (неудовлетворительно)	40 и менее	не зачтено

Результаты промежуточного контроля по дисциплине, выраженные в оценках «зачтено, удовлетворительно», «зачтено, хорошо», «зачтено, отлично», показывают уровень сформированности у обучающегося компетенций по дисциплине в соответствии с картами компетенций основной профессиональной образовательной программы высшего образования — программы магистратуры «Музейные исследования и кураторские стратегии» по направлению подготовки 51.04.04 Музеология и охрана объектов культурного и природного наследия.

Результаты промежуточного контроля по дисциплине, выраженные в оценках «не зачтено, неудовлетворительно», показывают несформированность у обучающегося

компетенций по дисциплине в соответствии с картами компетенций основной профессиональной образовательной программы высшего образования — программы магистратуры «Музейные исследования и кураторские стратегии» по направлению подготовки 51.04.04 Музеология и охрана объектов культурного и природного наследия.

7.4 Типовые задания к промежуточной аттестации

УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий

В рамках выбранной темы магистранту необходимо показать наличие знаний и умения применять различные методологии и методики системного и критического анализа проблемных ситуаций, решение которых возможно посредством применения инструментов обработки естественного языка, для чего необходимо смоделировать ситуацию, возможную в профессиональной деятельности, сформулировать, обосновать и решить задачи, которые характерны для данной ситуации, посредством соответствующих инструментов обработки естественного языка.

Примерный перечень тем 1-й письменной работы (эссе):

1. Сравнение качества и производительности морфологических библиотек для русского языка.
2. Сравнение качества и производительности библиотек анализа тональности для русского языка.
3. Сравнение качества и производительности библиотек для выделения именованных сущностей для русского языка.
4. Сравнение качества онлайн сервисов машинного перевода для пары русский-английский (одно направление).
5. Сравнительный анализ произведений Льва Толстого и Федора Достоевского с помощью стилометрических методов.
6. Гендерное смещение (gender bias) в дистрибутивных моделях русского языка (проанализировать 2-3 статические модели отсюда: https://rusvectores.org/ru/models/).
7. Сравнительный анализ двух моделей вопросно-ответного поиска для русского языка с помощью инструмента CheckList (https://github.com/marcotcr/checklist).
8. Анализ качества кросс-языкового переноса моделей вопросно-ответного поиска на данных без дообучения. Исходный английский набор данных – SquAD, тестирование – на данных TyDi QA.
9. Исследование переносимости моделей распознавания юмора: исследовать 2-3 метода классификации на 2-3 англоязычных наборах данных.
10. Анализ существующих систем вопросно-ответного поиска по базам знаний (deeppavlov, Qanswer) с помощью тестового набора данных RuBQ.
11. Систематический анализ качества генерации текстов с помощью модели ruGPT-3 для различных сценариев (https://github.com/sberbank-ai/ru-gpts).

ПК-3 Способен использовать современные методы обработки и интерпретации информации в профессиональной сфере

В рамках выбранной темы магистранту необходимо показать наличие знаний современных методов обработки изображений в области исследований культурного и природного наследия, последовательно описать процедуру применения различных методов и инструментов обработки изображений на примере конкретного научно-исследовательского проекта или возможной ситуации в работе музея по выбору магистранта.

Примерный перечень тем 2-й письменной работы:

1. Методы обработки естественного языка.
2. Машинный перевод.
3. Голосовые помощники.

4. Анализ текстов.
5. Распознавание и синтез речи.
6. Обработка естественного языка на Java.
7. Обработка естественного языка в контексте Data Science фреймворков.
8. Обработка естественного языка с TensorFlow.
9. Глубокое обучение при обработке естественного языка.
10. NLP на Python.
11. Интеграция Spark и библиотек машинного обучения.
12. Нейросетевые методы в обработке естественного языка.
13. Анализ моделей, основанных на механизме внимания и архитектуре Transformer.
14. Анализ задач музейной работы, которые может решить NLP.
15. Предобработка текста.
16. Стемминг.
17. Лемматизация.
18. Векторизация.
19. Дедубликация.
20. Семантический анализ.
21. Распознавание именованных сущностей и извлечение отношений.
22. Использование N-грамм.
23. Частеречная разметка.
24. Библиотеки для NLP.
25. Анализ примера кода на языке Scala.
26. Обзор подходов и методов к задаче автоматического извлечения именованных сущностей.

7.5 Средства оценки индикаторов достижения компетенций

Таблица 8

Средства оценки индикаторов достижения компетенций

Коды компетенций	**Индикаторы компетенций** *(в соотв. с Таблицей 1)*	**Средства оценки** *(в соотв. с Таблицами 5, 7)*
УК-1	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4.	Опрос, диспут, практическое задание, контрольное задание, письменная работа (эссе)
ПК-3	ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	Опрос, диспут, практическое задание, контрольное задание, письменная работа (эссе)

Таблица 9

Описание средств оценки индикаторов достижения компетенций

Средства оценки *(в соотв. с Таблицами 5, 7)*	**Рекомендованный план выполнения работы**
Опрос	Магистрант в ходе подготовки и участия в опросе показывает наличие практической базы знаний в рамках дисциплины, необходимой для выполнения следующих действий в области профессиональной деятельности: – анализирует проблемную ситуацию, определяет пробелы в информации, оценивать надёжность источников информации, разрабатывает стратегию решения проблемной ситуации на основе системного и междисциплинарного подходов, строит сценарии реализации стратегии, определяя возможные риски и предлагая пути их устранения; – применяет современные методы, поиска, обработки, анализа и использования информации в рамках проведения научно-исследовательских и организационных работ в области музейных исследований и кураторских стратегий.

Средства оценки *(в соотв. с Таблицами 5, 7)*	**Рекомендованный план выполнения работы**
Диспут	Магистрант в ходе подготовки и участия в диспуте показывает наличие практической базы знаний в рамках дисциплины, необходимой для выполнения следующих действий в области профессиональной деятельности: – анализирует проблемную ситуацию, определяет пробелы в информации, оценивать надёжность источников информации, разрабатывает стратегию решения проблемной ситуации на основе системного и междисциплинарного подходов, строит сценарии реализации стратегии, определяя возможные риски и предлагая пути их устранения; – применяет современные методы, поиска, обработки, анализа и использования информации в рамках проведения научно-исследовательских и организационных работ в области музейных исследований и кураторских стратегий.
Практическое задание	Магистрант в ходе подготовки и выполнения практического задания показывает наличие практической базы знаний в рамках дисциплины, необходимой для выполнения следующих действий в области профессиональной деятельности: – анализирует проблемную ситуацию, определяет пробелы в информации, оценивать надёжность источников информации, разрабатывает стратегию решения проблемной ситуации на основе системного и междисциплинарного подходов, строит сценарии реализации стратегии, определяя возможные риски и предлагая пути их устранения; – применяет современные методы, поиска, обработки, анализа и использования информации в рамках проведения научно-исследовательских и организационных работ в области музейных исследований и кураторских стратегий.
Контрольное задание	Магистрант в ходе подготовки и выполнения контрольного задания показывает наличие практической базы знаний в рамках дисциплины, необходимой для выполнения следующих действий в области профессиональной деятельности: – анализирует проблемную ситуацию, определяет пробелы в информации, оценивать надёжность источников информации, разрабатывает стратегию решения проблемной ситуации на основе системного и междисциплинарного подходов, строит сценарии реализации стратегии, определяя возможные риски и предлагая пути их устранения; – применяет современные методы, поиска, обработки, анализа и использования информации в рамках проведения научно-исследовательских и организационных работ в области музейных исследований и кураторских стратегий.
Письменная работа (эссе)	Магистрант в ходе подготовки и написания письменной работы (эссе), показывает наличие практической базы знаний в рамках дисциплины, необходимой для выполнения следующих действий в области профессиональной деятельности: – анализирует проблемную ситуацию, определяет пробелы в информации, оценивать надёжность источников информации, разрабатывает стратегию решения проблемной ситуации на основе системного и междисциплинарного подходов, строит сценарии реализации стратегии, определяя возможные риски и предлагая пути их устранения; – применяет современные методы, поиска, обработки, анализа и использования информации в рамках проведения научно-исследовательских и организационных работ в области музейных исследований и кураторских стратегий.

8. ОСНОВНАЯ И ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА

8.1 Основная литература

1. Ганегедара, Т. Обработка естественного языка с TensorFlow : монография / Т. Ганегедара ; пер. с анг. В. С. Яценкова. - Москва : ДМК Пресс, 2020. - 382 с. - ISBN 978-5-97060-756-5. - Текст : электронный. - URL: https://znanium.com/catalog/product/1094940 .

2. Тапскотт, Д. Викиномика: как массовое сотрудничество изменяет все / Д. Тапскотт, Э. Д. Уильямс. - Москва : Интеллектуальная Литература, 2020. - 456 с. - ISBN 978-5-6042878-7-3. - Текст : электронный. - URL: https://znanium.com/catalog/product/1220225 . – Режим доступа: по подписке.

8.2 Дополнительная литература

1 Богданов, Е. П. Интеллектуальный анализ данных : практикум для магистрантов направления 09.04.03 «Прикладная информатика» профиль подготовки «Информационные системы и технологии корпоративного управления» / Е. П. Богданов. -

Волгоград : ФГБОУ ВО Волгоградский ГАУ, 2019. - 112 с. - Текст : электронный. - URL: https://znanium.com/catalog/product/1087885 . – Режим доступа: по подписке

9. ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ, ИСПОЛЬЗУЕМЫЕ ПРИ ОСУЩЕСТВЛЕНИИ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА

9.1 Программное обеспечение

При осуществлении образовательного процесса магистрантами и профессорско-преподавательским составом используется следующее лицензионное программное обеспечение:

1. OS Microsoft Windows (OVS OS Platform)
2. MS Office (OVS Office Platform)
3. Adobe Acrobat Professional 11.0 MLP AOO License RU
4. Adobe CS5.5 Design Standart Win IE EDU CLP
5. ABBYY FineReader 11 Corporate Edition
6. ABBYY Lingvo x5
7. Adobe Photoshop Extended CS6 13.0 MLP AOO License RU
8. Adobe Acrobat Reader DC /Pro – бесплатно
9. Google Chrome – бесплатно
10. Opera – бесплатно
11. Mozilla – бесплатно
12. VLC – бесплатно
13. Яндекс Браузер

9.2 Перечень информационно-справочных систем и профессиональных баз данных информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины:

Информационно-справочные системы

1. Гарант.Ру. Информационно-правовой портал: http://www.garant.ru
2. Информационная система «Единое окно доступа к образовательным ресурсам»: http://window.edu.ru/
3. Открытое образование. Ассоциация «Национальная платформа открытого образования»: http://npoed.ru
4. Официальная Россия. Сервер органов государственной власти Российской Федерации: http://www.gov.ru
5. Официальный интернет-портал правовой информации. Государственная система правовой информации: http://pravo.gov.ru
6. Правовой сайт КонсультантПлюс: http://www.consultant.ru/sys
7. Российское образование. Федеральный портал: http://www.edu.ru

Профессиональные базы данных информационно-телекоммуникационной сети «Интернет»:

1. Национальная электронная библиотека НЭБ: http://www.rusneb.ru
2. Неприкосновенный запас: http://magazines.russ.ru/nz/
3. Президентская библиотека: http://www.prlib.ru
4. Российская государственная библиотека: http://www.rsl.ru/
5. Российская национальная библиотека: http://www.nlr.ru/poisk/

9.3 Лицензионные электронные ресурсы библиотеки Университета

Профессиональные базы данных:

Полный перечень доступных обучающимся профессиональных баз данных представлен на официальном сайте Университета https://eusp.org/library/electronic-resources, включая следующие базы данных:

1. **East View** – 100 ведущих российских журналов по гуманитарным наукам (архив и текущая подписка): https://dlib.eastview.com/browse;
2. **eLIBRARY.RU** — Российский информационно-аналитический портал в области науки, технологии, медицины и образования, содержащий рефераты и полные тексты научных статей и публикаций, наукометрическая база данных: http://elibrary.ru;
3. **Университетская информационная система РОССИЯ** — база электронных ресурсов для учебных программ и исследовательских проектов в области социально-гуманитарных наук: http://www.uisrussia.msu.ru/;
4. Электронные журналы по подписке (текущие номера научных зарубежных журналов).

Электронные библиотечные системы:

1. **Znanium.com** – Электронная библиотечная система **(ЭБС)** – http://znanium.com/;
2. Университетская библиотека онлайн – Электронная библиотечная система (ЭБС) – http://biblioclub.ru/

9.4 Электронная информационно-образовательная среда Университета

Образовательный процесс по дисциплине поддерживается средствами электронной информационно-образовательной среды Университета, которая включает в себя электронный учебно-методический ресурс АНООВО «ЕУСПб» — образовательный портал LMS Sakai — Sakai@EU, лицензионные электронные ресурсы библиотеки Университета, официальный сайт Университета (Европейский университет в Санкт-Петербурге [https://eusp.org/]), локальную сеть Университета и корпоративную электронную почту и обеспечивает:

- доступ к учебным планам, рабочим программам дисциплин (модулей), практик и к изданиям электронных библиотечных систем и электронным образовательным ресурсам, указанным в рабочих программах;
- фиксацию хода образовательного процесса, результатов промежуточной аттестации и результатов освоения основной образовательной программы;
- формирование электронного портфолио обучающегося, в том числе сохранение работ обучающегося, рецензий и оценок за эти работы со стороны любых участников образовательного процесса;
- взаимодействие между участниками образовательного процесса, в том числе синхронное и (или) асинхронное взаимодействие посредством сети «Интернет» (электронной почты и т.д.).

Каждый обучающийся в течение всего периода обучения обеспечен индивидуальным неограниченным доступом к электронным ресурсам библиотеки Университета, содержащей издания учебной, учебно-методической и иной литературы по изучаемой дисциплине.

10. МАТЕРИАЛЬНО-ТЕХНИЧЕСКАЯ БАЗА, НЕОБХОДИМАЯ ДЛЯ ОСУЩЕСТВЛЕНИЯ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА

В ходе реализации образовательного процесса используются специализированные многофункциональные аудитории для проведения занятий лекционного типа, занятий семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, укомплектованные специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Проведение занятий лекционного типа обеспечивается демонстрационным оборудованием.

Помещения для самостоятельной работы оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду организации.

Для лиц с ограниченными возможностями здоровья и инвалидов предоставляется возможность присутствия в аудитории вместе с ними ассистента (помощника). Для слабовидящих предоставляется возможность увеличения текста на экране (ПК). Для самостоятельной работы лиц с ограниченными возможностями здоровья в помещении для самостоятельной работы организовано одно место (ПК) с возможностями бесконтактного ввода информации и управления компьютером (специализированное лицензионное программное обеспечение – Camera Mouse, веб камера). Библиотека университета предоставляет удаленный доступ к ЭБ с возможностями для слабовидящих увеличения текста на экране ПК. Лица с ограниченными возможностями здоровья могут при необходимости воспользоваться имеющимся в университете креслом-коляской. В учебном корпусе имеется адаптированный лифт. На первом этаже оборудован специализированный туалет. У входа в здание университета для инвалидов оборудована специальная кнопка, входная среда обеспечена информационной доской о режиме работы университета, выполненной рельефно-точечным тактильным шрифтом (азбука Брайля).

ПРИЛОЖЕНИЕ 1

ДИСЦИПЛИНА ПО ВЫБОРУ

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ПО ДИСЦИПЛИНЕ

«Инструменты обработки естественного языка»

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕЙ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ

1 Показатели, критерии и оценивание компетенций и индикаторов их достижения в процессе текущей аттестации

Информация о содержании и процедуре текущего контроля успеваемости, методике оценивания знаний, умений и навыков обучающегося в ходе текущего контроля доводятся научно-педагогическими работниками Университета до сведения обучающегося на первом занятии по данной дисциплине.

Текущий контроль предусматривает подготовку магистрантов к каждому лабораторному занятию, участие в опросах, диспутах, подготовку практических заданий, выполнение контрольных заданий, активное слушание на лекциях. Магистрант должен присутствовать на семинарских занятиях, отвечать на поставленные вопросы, показывая, что прочитал разбираемую литературу, представлять содержательные реплики по обсуждаемым вопросам.

Текущий контроль проводится в форме устных опросов и оценивания участия магистрантов в проходящих диспутах, оценивания выполненных практических заданий, контрольных работ, демонстрирующих степень знакомства с дополнительной литературой.

Таблица 1

Показатели, критерии и оценивание компетенций и индикаторов их достижения в процессе текущей аттестации

Наименование темы (раздела)	Код компетенции	Индикаторы компетенций	Коды ЗУВ (в соотв. с табл. 1)	Формы текущего контроля	Результаты текущего контроля
Введение	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Опрос 1 Диспут 1	зачтено/ не зачтено зачтено/ не зачтено
Морфологический анализ	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Контрольное задание 1 Контрольное задание 2 Опрос 2 Контрольное задание 3	зачтено/ не зачтено зачтено/ не зачтено зачтено/ не зачтено зачтено/ не зачтено
Языковые модели	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Контрольное задание 4 Контрольное задание 5 Контрольное задание 6	зачтено/ не зачтено зачтено/ не зачтено зачтено/ не зачтено
Анализ тональности	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2.	З (УК-1) У (УК-1)	Практическое задание 1	зачтено/ не зачтено

Наименование темы (раздела)	Код компетенции	Индикаторы компетенций	Коды ЗУВ (в соотв. с табл. 1)	Формы текущего контроля	Результаты текущего контроля
		ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)		
Синтаксический анализ	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Контрольное задание 7 Контрольное задание 8 Практическое задание 2	зачтено/ не зачтено зачтено/ не зачтено зачтено/ не зачтено
Извлечение информации	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Практическое задание 3	зачтено/ не зачтено
Автоматическое реферирование	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Контрольное задание 9	зачтено/ не зачтено
Вычислительная семантика	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Диспут 2 Практическое задание 4	зачтено/ не зачтено зачтено/ не зачтено
Ответы на вопросы	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Практическое задание 5	зачтено/ не зачтено зачтено/ не зачтено
Машинный перевод (МП)	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2.	З (УК-1) У (УК-1)	Контрольное задание 10	зачтено/ не зачтено

Наименование темы (раздела)	Код компетенции	Индикаторы компетенций	Коды ЗУВ (в соотв. с табл. 1)	Формы текущего контроля	Результаты текущего контроля
		ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)		

Таблица 2

Критерии оценивания

Формы текущего контроля успеваемости	Критерии оценивания
Опрос	Ответ отсутствует или является односложным, или содержит существенные ошибки – не зачтено Магистрант в ответах демонстрирует знание всех теоретических положений, (развернуто) отвечает на все поставленные вопросы, предлагает обоснования при ответе на все или большинство поставленных вопросов; несущественные ошибки не снижают качество ответа — зачтено
Диспут	Пассивность, участие без представления аргументов и обоснования точки зрения, несформированность навыков профессиональной коммуникации в группе — не зачтено Представление аргументированной научной позиции, обоснование точки зрения в диспуте, демонстрация навыков профессиональной коммуникации в группе — зачтено
Практическое задание	магистрант выполняет задание частично или с существенными недочетами (некорректно сформулирован исследовательский вопрос, не определены основные агенты, некорректно выбраны методы исследования, требования к содержанию, структуре, логике, аргументации, оформлению не выполнены) – не зачтено, полное и правильное выполнение задания в соответствии с требованиями к содержанию, структуре, логике, аргументации, оформлению с возможным небольшим количеством погрешностей (например, плохо выдержанная структура текста, недостаточная аргументация отдельных тезисов) – зачтено
Контрольное задание	магистрант выполняет задание частично или с существенными недочетами (некорректно сформулирован исследовательский вопрос, не определены основные агенты, некорректно выбраны методы исследования, требования к содержанию, структуре, логике, аргументации, оформлению не выполнены) – не зачтено, полное и правильное выполнение задания в соответствии с требованиями к содержанию, структуре, логике, аргументации, оформлению с возможным небольшим количеством погрешностей (например, плохо выдержанная структура текста, недостаточная аргументация отдельных тезисов) – зачтено

2. Контрольные задания для текущей аттестации

Материалы опросов, диспутов, практических заданий:

Тема 1. Введение.

Опрос 1.

1. Какие основные инструменты и приложения используются при обработке естественного языка?
2. История развития методов обработки естественного языка?

Диспут 1.

Как технологии обработки естественного языка меняют повседневную жизнь людей.

Тема 2. Морфологический анализ.

Контрольное задание 1.

Три слова в списке имеют одинаковый тип словоизменения, одно – отличается. Отметьте это слово.

1) чело
2) табло
3) сверло
4) кайло.

Контрольное задание 2.

Какие термины относятся к морфологической обработке? Выберите номера правильных ответов:

1) лемматизация;
2) лемминг;
3) стемминг;
4) прокрастинация;
5) парсинг.

Опрос 2.

1. Каковы основные подходы к морфологическому анализу?
2. Какова основная методика отбора данных для морфологического анализа?
3. Каковы основные инструменты морфологического анализа?

Контрольное задание 3.

Отметьте правильный набор граммем (в нотации НКРЯ/mystem) для выделенного слова в предложении:

Лично я ловлю покемонов не одобряю.

1) S,жен,неод=вин,ед
2) V,несов,пе=непрош,ед,изъяв,1-л
3) V,несов,пе=ед,пов,2-л
4) S,муж,неод=им,ед.

Тема 3. Языковые модели.

Контрольное задание 4.

Постройте биграммную языковую модель на основе корпуса:

<s> Вася любит мороженое </s>
<s> Лена любит малину </s>
<s> Вася любит Лену </s>
<s> Георгий ест мороженое </s>
<s> Лена рисует яблоко </s>
<s> Георгий любит Катю </s>
<s> Георгий любит смотреть, как Лена ест мороженое </s>

Упорядочите предложения по убыванию оценок вероятностей на основе построенной языковой модели.

1. <s> Лена любит мороженое </s>
2. <s> Лена рисует малину </s>
3. <s> Вася любит Катю </s>.

Контрольное задание 5.

Постройте биграммную языковую модель на основе корпуса:

<s> Вася любит мороженое </s>

<s> Лена любит малину </s>
<s> Вася любит Лену </s>
<s> Георгий ест мороженое </s>
<s> Лена рисует яблоко </s>
<s> Георгий любит Катю </s>
<s> Георгий любит смотреть, как Лена ест мороженое </s>

Вычислите перплексию модели на предложении:
<s> Георгий любит малину </s>
Обратите внимание, что для вычисления вероятности предложения (и, соответственно, перплексии) используются вероятности 4-х биграмм (N=4 в формуле вычисления перплексии).

Контрольное задание 6.

Пусть у нас есть корпус, содержащий 10000 предложений, размер словаря – 1500 уникальных слов (включая специальные «слова» – маркеры начала и конца предложений). Некоторые частоты униграмм:
ем 100
дуриан 1
и 5000
не 3000
морщусь 50

и биграмм:
<s> ем 20
ем дуриан 0
дуриан и 0
и не 300
не морщусь 15
морщусь </s> 5.
Примените сглаживание Лапласа (α=1, сглаживание «+1») для оценки вероятностей биграмм и оцените на их основе вероятность предложения <s> ем дуриан и не морщусь </s>
В качестве ответа введите натуральный логарифм оценки вероятности предложения.

Тема 4. Анализ тональности.

Практическое задание 1

Реализуйте анализатор тональности одним из способов. Данные – английские предложений из отзывов о фильмах из Stanford Sentiment Treebank (из датасета взяты только целые предложения; 5-уровневая разметка приведена к трехуровневой – негативный, нейтральный, позитивный). Вы можете использовать подход на основе словаря тонально окрашенных слов (например, SentiWords) или обучить классификатор на тренировочных данных. Вы можете использовать тренировочные данные и в первом случае – чтобы подобрать пороги для классификации предложений на основе весов словаря. Постройте матрицу ошибок (confusion matrix) на тестовом наборе. Оцените правильность (accuracy, доля правильно классифицированных предложений) классификатора на тестовом наборе. Проанализируйте неверно классифицированные предложения, сделайте предположения о причинах неверной классификации, предложите улучшения.

Тема 5. Синтаксический анализ.

Контрольное задание 7.

Выберите правильный разбор на составляющие предложения:
Советник губернатора Чукотского автономного округа Романа Абрамовича Роман

Копин победил на повторных выборах главы администрации Чаунского района

1. [[[Советник [[губернатора [Чукотского [автономного округа]]] [Романа Абрамовича]]] [Роман Копин]] [победил [на [повторных [выборах [главы [администрации [Чаунского района]]]]]]]
2. [[[Советник [[губернатора [Чукотского [автономного округа]]] [Романа Абрамовича]]] [Роман Копин]] [[[победил на] [повторных [выборах [главы [администрации [Чаунского района]]]]]]]
3. [[[[[Советник губернатора] [Чукотского [автономного округа]]] [Романа Абрамовича]] [Роман Копин]] [победил [на [[повторных выборах] [главы [администрации [Чаунского района]]]]]]
4. [[[Советник [[губернатора [Чукотского [автономного округа]]] [Романа Абрамовича]]] [[Роман Копин] победил]] [на [[повторных выборах] [главы [администрации [Чаунского района]]]]]
5. [[[Советник [[губернатора Чукотского] [[автономного округа] [Романа Абрамовича]]]] [Роман Копин]] [победил [на [повторных [выборах [главы [администрации [Чаунского района]]]]]]]

Контрольное задание 8.

Выберите правильный разбор предложения в терминах зависимостей:

0 root 1 Активно 2 обсуждается 3 роль 4 нашей 5 страны 6 в 7 современном 8 быстро 9 меняющемся 10 мире, 11 проходящем 12 через 13 переломный 14 этап.

В приведенных ниже вариантах разбора каждая пара – зависимость (хозяин, слуга), знаки препинания не учитываются, используется традиционный подход при установлении зависимостей с участием предлогов (не как в universal dependencies).

1. (0, 3) (3, 2) (2, 1) (3, 5) (5, 4) (3, 6) (6, 10) (10, 7) (10, 9) (9, 8) (10, 11) (11, 12) (12, 14) (14, 13)
2. (0, 2) (2, 1) (2, 3) (3, 5) (5, 4) (3, 6) (6, 10) (10, 7) (10, 9) (9, 8) (10, 11) (11, 12) (12, 14) (14, 13)
3. (0, 2) (2, 1) (2, 3) (3, 5) (5, 4) (5, 6) (6, 10) (10, 7) (10, 9) (9, 8) (10, 11) (11, 12) (12, 14) (14, 13)
4. (0, 2) (2, 1) (2, 3) (3, 5) (5, 4) (3, 6) (6, 7) (7, 8) (8, 9) (9, 10) (10, 11) (11, 12) (12, 14) (14, 13)
5. (0, 2) (2, 3) (3, 5) (5, 4) (3, 6) (6, 10) (10, 7) (10, 9) (2, 8) (10, 11) (11, 1) (11, 12) (12, 14) (14, 13)

Практическое задание 2

Постройте деревья зависимостей для 2000 предложений «Войны и Мира» и «Братьев Карамазовых» с помощью библиотеки Stanza.

Оцените производительность библиотеки (предложения/с).

Проведите ручную оценку качества разбора на случайных 20 предложениях (по 10 из каждой книги).

Посчитайте среднюю глубину дерева разбора для каждого из романов. Посчитайте корреляцию между длиной предложения в словах и глубиной дерева разбора. Оцените на нескольких примерах, насколько глубина дерева соответствует субъективной сложности понимания предложения.

Тема 6. Извлечение информации.

Практическое задание 3

Примените и оцените модуль извлечения именованных сущностей Natasha (https://github.com/natasha/natasha). Для тестирования используйте текст и соответствующую разметку. Тестовые данные содержат только разметку для людей (PER) и организаций (ORG). Рассчитайте F1 для каждого типа сущностей и общее значение F1.

Тема 7. Автоматическое реферирование.

Контрольное задание 9.

Пусть у нас есть два «реферата-образца»:
1: карп лещ лещ щука сазан
2: лещ карп лещ сазан плотва

Упорядочите «рефераты» ниже по убыванию значения ROUGE-2.
1) лещ карп лещ сазан плотва
2) плотва сазан лещ карп сазан
3) щука сазан карп лещ сазан
4) карп карась лещ окунь лещ
5) щука сазан лещ карп сазан

Тема 8. Вычислительная семантика.

Диспут 2.

Возможные негативные последствия использования больших предобученных моделей для генерации текста.

Практическое задание 4

Реализуйте генератор юмора по мотивам работы Alessandro Valitutti et al. “Let Everything Turn Well in Your Wife”: Generation of Adult Humor Using Lexical Constraints, 2013. На входе генератора – новостной заголовок, в котором надо заменить одно слово. Предлагаемый алгоритм:

1. Проведите разбор предложения с помощью библиотеки Stanza. На основе результатов разбора выберете слово-кандидат на замену.
2. Найдите антоним для слова в WordNet (используйте интерфейс библиотеки NLTK).
3. Если антоним не нашелся, то найдите несколько слов, близких по звучанию или рифму с помощью datamuse api.
4. Получите вектора fasttext для начального слова и вариантов замены. Среди этих слов найдите самое далекое по косинусному расстоянию.

Оцените 20 модификаций по шкале от 0 (совсем не смешно) до 3 (очень смешно), приведите среднюю оценку.

Тема 9. Ответы на вопросы.

Практическое задание 5

Обучите модель для выделения ответа на вопрос из параграфа на тренировочной части русскоязычных данных TyDi QA. Оцените качество модели на данных для настройки (development set, EM/F1). Проанализируйте случаи, с которыми модель справилась хуже всего. Сделайте предположение, в чем сложность этих случаев.

Тема 10. Машинный перевод (МП).

Контрольное задание 10

Для фразы
Call me what instrument you will, though you can fret me, you cannot play upon me.

Есть два образцовых перевода (для простоты знаки препинания в образцах и вариантах перевода удалены):

- *назовите меня каким угодно инструментом вы хоть и можете меня терзать но играть на мне не можете*

- *объявите меня каким угодно инструментом вы можете расстроить меня но играть на мне нельзя*

Упорядочите варианты перевода ниже по убыванию BLEU-2 (метрика на основе униграмм и биграмм). Помните, что стандартная метрика BLEU не предполагает лемматизацию текстов.

- *позвони мне на каком инструменте вы будете хотя вы можете беспокоиться меня но вы не можете играть на мне*
- *назовите мне какой инструмент вы хотите хотя можете меня беспокоить но вы не можете играть на меня*
- *позвони мне какой инструмент ты будешь хотя ты можешь меня волновать но ты не можешь играть на меня*
- *назовите меня какой инструмент вы будете хотя вы можете раздражать меня все же вы не можете играть на меня*
- *считай меня чем тебе угодно ты можешь мучить меня но не играть мною*
- *назови меня каким угодно инструментом ты можешь меня расстроить но не играть на мне.*

3 Показатели, критерии и оценивание компетенций и индикаторов их достижения в процессе промежуточной аттестации

Форма промежуточной аттестации — зачет с оценкой, выставляемый на основе 2-х письменных работ (эссе).

Перед зачетом с оценкой проводится консультация, на которой преподаватель отвечает на вопросы магистрантов.

В результате промежуточного контроля знаний студенты получают оценку по дисциплине.

Таблица 3

Показатели, критерии и оценивание компетенций и индикаторов их достижения в процессе промежуточной аттестации

Форма промежуточной аттестации/вид промежуточной аттестации	**Коды компетенций**	**Индикаторы компетенций** *(в соотв. с Таблицей 1)*	**Коды ЗУВ** (в соответствии с Таблицей 1)	**Критерии оценивания**	**Оценка**
Зачет с оценкой/ Письменная работа (эссе)	УК-1 ПК-3	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4. ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	З (УК-1) У (УК-1) В (УК-1) З (ПК-3) У (ПК-3) В (ПК-3)	Эссе соответствует следующим требованиям: сформулирован исследовательский вопрос, корректно выбраны методы и собраны данные, тема раскрыта, соблюдены структура и научный стиль, сформулированы выводы, аргументация убедительна, правильно оформлен библиографический аппарат и т.д., магистрант демонстрирует: глубокое усвоение программного материала, изложение его исчерпывающе, последовательно, четко, умение делать обоснованные выводы,	Зачтено, отлично

Форма промежуточной аттестации/вид промежуточной аттестации	Коды компетенций	Индикаторы компетенций *(в соотв. с Таблицей 1)*	Коды ЗУВ (в соответствии с Таблицей 1)	Критерии оценивания	Оценка
				соблюдение норм устной и письменной литературной речи/ Эссе успешно представлено на защите. Магистрант дает правильный ответ на теоретический вопрос, при условии, что отдельные неточности, допускаемые в ходе ответа, никак не снижают общего качества ответа. Для ответа характерно: • глубокое усвоение программного материала, • изложение его исчерпывающе, последовательно, четко, • умение делать обоснованные выводы, • соблюдение норм устной и письменной литературной речи.	
				В эссе не соблюдены некоторые требования к работе: возможно несоблюдении одного-двух требований и допущении некоторых неточностей, магистрант демонстрирует: твердое знание материала курса, последовательное изложение материала, знание теоретических положений без обоснованной их аргументации, соблюдение норм устной и письменной литературной речи; Эссе успешно представлено на защите. Магистрант верно отвечает на вопрос, указанный в билете, при условии, что ответ на вопрос характеризуется отсутствием серьезных, значимых неточностей, при следующих характеристиках ответа: • твердое знание материала курса, • последовательное изложение материала,	Зачтено, хорошо

Форма промежуточной аттестации/вид промежуточной аттестации	**Коды компетенций**	**Индикаторы компетенций** *(в соотв. с Таблицей 1)*	**Коды ЗУВ** (в соответствии с Таблицей 1)	**Критерии оценивания**	**Оценка**
				• знание теоретических положений без обоснованной их аргументации, • соблюдение норм устной и письменной литературной речи.	
				Эссе содержит существенные оплошности: нарушено сразу несколько требований, например, выводы плохо обоснованы, есть фактические ошибки, магистрант при защите демонстрирует: знание основного материала, но владение им не в полном объеме, допущение существенных неточностей, недостаточно правильных формулировок, допущение нарушения логической последовательности в изложении материала, наличие нарушений норм литературной устной и письменной речи. Эссе представлено на защите. Магистрант представляет правильный ответ на теоретический вопрос, указанный в билете, при условии, что ответ на вопрос характеризуется значительными неточностями, при следующих параметрах ответа: • знание основного материала, но владение им не в полном объеме, • допущение существенных неточностей, недостаточно правильных формулировок, • допущение нарушения логической последовательности в изложении материала, • наличие нарушений норм литературной устной и письменной речи.	Зачтено, удовлетворительно

Форма промежуточной аттестации/вид промежуточной аттестации	**Коды компетенций**	**Индикаторы компетенций** *(в соотв. с Таблицей 1)*	**Коды ЗУВ** (в соответствии с Таблицей 1)	**Критерии оценивания**	**Оценка**
				Представленное эссе не отвечает предъявляемым требованиям (либо не предоставление эссе); имеет место: незнание значительной части программного материала, наличие существенных ошибок в определениях, формулировках, понимании теоретических положений; бессистемность при ответе на поставленный вопрос, отсутствие в ответе логически корректного анализа, аргументации, классификации, наличие нарушений норм устной и письменной литературной речи. Магистрант представляет ответ на вопрос, характеризующийся наличием существенных ошибок в определениях, формулировках, понимании теоретических положений, свидетельствующий о некомпетентности магистранта, при следующих параметрах ответа: • незнание значительной части программного материала, • наличие существенных ошибок в определениях, формулировках, понимании теоретических положений; • бессистемность при ответе на поставленный вопрос, • отсутствие в ответе логически корректного анализа, аргументации, классификации, наличие нарушений норм устной и письменной литературной речи.	Не зачтено, не удовлетворительно

Результаты сдачи промежуточной аттестации по направлениям подготовки уровня магистратуры оцениваются по стобалльной системе оценки в соответствии с Положением о формах, периодичности и порядке организации и проведения текущего контроля

успеваемости и промежуточной аттестации обучающихся в АНООВО «ЕУСПб» следующим образом согласно таблице 3а.

Таблица 3а

Система оценки знаний обучающихся

Пятибалльная (стандартная) система	Стобалльная система оценки	Бинарная система оценки
5 (отлично)	100-81	зачтено
4 (хорошо)	80-61	
3 (удовлетворительно)	60-41	
2 (неудовлетворительно)	40 и менее	не зачтено

Результаты промежуточного контроля по дисциплине, выраженные в оценках «зачтено, удовлетворительно», «зачтено, хорошо», «зачтено, отлично», показывают уровень сформированности у обучающегося компетенций по дисциплине в соответствии с картами компетенций основной профессиональной образовательной программы высшего образования — программы магистратуры «Музейные исследования и кураторские стратегии» по направлению подготовки 51.04.04 Музеология и охрана объектов культурного и природного наследия.

Результаты промежуточного контроля по дисциплине, выраженные в оценках «не зачтено, неудовлетворительно», показывают несформированность у обучающегося компетенций по дисциплине в соответствии с картами компетенций основной профессиональной образовательной программы высшего образования — программы магистратуры «Музейные исследования и кураторские стратегии» по направлению подготовки 51.04.04 Музеология и охрана объектов культурного и природного наследия.

4 Задания к промежуточной аттестации

УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий

В рамках выбранной темы магистранту необходимо показать наличие знаний и умения применять различные методологии и методики системного и критического анализа проблемных ситуаций, решение которых возможно посредством применения инструментов обработки естественного языка, для чего необходимо смоделировать ситуацию, возможную в профессиональной деятельности, сформулировать, обосновать и решить задачи, которые характерны для данной ситуации, посредством соответствующих инструментов обработки естественного языка.

Перечень тем 1-й письменной работы (эссе) с эталонными ответами по дисциплине по выбору:

1. Сравнение качества и производительности морфологических библиотек для русского языка.

Эталонный ответ

Морфологическая обработка естественного языка — это процесс анализа слов с целью определения их грамматических характеристик, таких как часть речи, род, число, падеж и т. д. Морфологические библиотеки представляют собой инструменты, которые помогают автоматизировать этот процесс.

Для сравнения качества и производительности различных морфологических библиотек необходимо учитывать следующие факторы:

* Точность: насколько точно библиотека может определить грамматические характеристики слов.
* Полнота: насколько полно библиотека охватывает все возможные формы слов.
* Скорость работы: насколько быстро библиотека обрабатывает текст.
* Поддержка различных форматов данных: может ли библиотека обрабатывать тексты в разных форматах, таких как HTML, XML, JSON и др.
* Простота использования: насколько легко библиотека интегрируется с другими

инструментами обработки текста.
* Документация и поддержка: наличие подробной документации и поддержки от разработчиков библиотеки.
На рынке существует множество морфологических библиотек для русского языка. Вот некоторые из них:
1. Pymorphy2: это библиотека Python, которая использует словарь для определения грамматических характеристик слов. Она обладает высокой точностью и полнотой, но может быть медленной при обработке больших объёмов данных.
2. Mystem: это коммерческая библиотека, разработанная компанией Яндекс. Она также использует словарь для определения грамматических характеристик слов, но обладает более высокой скоростью работы.
3. TextBlob: это библиотека Python, которая предоставляет простой интерфейс для обработки текстов. Она включает в себя модуль для морфологического анализа, который может быть менее точным, чем Pymorphy2 или Mystem, но более быстрым.
4. Spacy: это ещё одна библиотека Python для обработки естественного языка, которая включает в себя морфологический анализ. Spacy обладает высокой скоростью работы и хорошей точностью, а также поддерживает различные форматы данных.
5. Stanza: это библиотека для обработки естественного языка на основе машинного обучения, которая может использоваться для морфологического анализа. Stanza обладает высокой точностью, но может потребовать больше времени для обучения модели.
Выбор конкретной библиотеки зависит от конкретных потребностей проекта. Если требуется высокая точность и полнота, то можно выбрать Pymorphy2 или Mystem. Если важна скорость работы, то можно рассмотреть TextBlob или Spacy. Если же требуется гибкость и возможность настройки параметров, то Stanza может быть хорошим выбором.

2. Сравнение качества и производительности библиотек анализа тональности для русского языка.

Эталонный ответ
Анализ тональности — это метод обработки естественного языка, который позволяет определить отношение автора к объекту или событию. Он используется для анализа отзывов клиентов, новостных статей, социальных медиа и других текстовых данных.
Для анализа тональности используются различные библиотеки и инструменты, которые предоставляют функции для обработки текста, определения тональности и визуализации результатов. В этой статье мы рассмотрим некоторые из наиболее популярных библиотек анализа тональности для русского языка и сравним их по качеству и производительности.
1. TextBlob — это библиотека Python, которая предоставляет простые и удобные функции для работы с текстом. Она включает в себя инструменты для токенизации, лемматизации, стемминга, определения частей речи, синтаксического разбора и анализа тональности. TextBlob использует машинное обучение для определения тональности на основе словаря эмоций.
Качество: TextBlob может быть полезен для быстрого и простого анализа тональности небольших объёмов данных. Однако его результаты могут быть неточными и субъективными, особенно для сложных и неоднозначных текстов.
Производительность: TextBlob работает быстро и эффективно, но его скорость может зависеть от размера данных и сложности задачи.
2. VADER (Valence Aware Dictionary and sEntiment Reasoner) — это ещё одна библиотека Python, разработанная специально для анализа тональности текстов на английском и русском языках. Она основана на словаре эмоций и правилах грамматики, которые позволяют определять тональность на уровне слов, фраз и предложений. VADER также учитывает контекст и интенсивность эмоций, что делает его более точным и надёжным.
Качество: VADER обеспечивает более точные и объективные результаты, чем TextBlob,

благодаря использованию словаря эмоций и правил грамматики. Это делает его подходящим для анализа больших объёмов данных и сложных текстов.
Производительность: VADER работает медленнее, чем TextBlob, но это компенсируется его точностью и надёжностью.
3. SentiRuEval — это набор инструментов и данных для оценки качества анализа тональности русскоязычных текстов. Он включает в себя наборы данных, аннотации и метрики, а также инструменты для сравнения и оценки различных методов анализа тональности. SentiRuEval позволяет оценить качество и точность различных библиотек и алгоритмов.
Сравнение качества и производительности библиотек анализа тональности зависит от конкретных требований и условий задачи. Если вам нужен быстрый и простой анализ небольших объёмов данных, то TextBlob может быть хорошим выбором. Если же вам нужны более точные и надёжные результаты для больших объёмов данных и сложных текстов, то VADER может быть предпочтительнее. SentiRuEval предоставляет инструменты для оценки и сравнения различных методов и библиотек анализа тональности, что позволяет выбрать оптимальный вариант для вашей задачи.
В целом, выбор библиотеки анализа тональности зависит от ваших целей, ресурсов и ограничений. Важно провести тестирование и сравнение различных методов, чтобы найти наилучший вариант для вашего проекта.

3. Сравнение качества и производительности библиотек для выделения именованных сущностей для русского языка.

Эталонный ответ
Сравнение качества и производительности библиотек для выделения именованных сущностей для русского языка
Выделение именованных сущностей (Named Entity Recognition, NER) — это задача обработки естественного языка, которая заключается в распознавании и классификации имён собственных, таких как имена людей, названия организаций, географические объекты и т. д. В русском языке эта задача может быть особенно сложной из-за морфологии, многозначности и других особенностей языка.
Для решения задачи NER существует множество библиотек и инструментов. Рассмотрим некоторые из них:
1. Spacy — библиотека машинного обучения с открытым исходным кодом для обработки естественного языка. Она предоставляет инструменты для токенизации, лемматизации, синтаксического анализа и других задач. Spacy поддерживает множество языков, включая русский. Библиотека использует алгоритмы машинного обучения для обучения моделей на основе данных.
2. Stanza — ещё одна библиотека с открытым исходным кодом, предназначенная для обработки текста. Stanza также поддерживает русский язык и предоставляет инструменты для NER, токенизации, тегирования частей речи и других задач.
3. TextBlob — простая библиотека для обработки текстов на Python. TextBlob предоставляет базовые инструменты для работы с текстом, такие как токенизация, стемминг, определение частей речи и другие. Однако она не специализируется на NER.
4. DeepPavlov — фреймворк для создания диалоговых систем на основе машинного обучения. DeepPavlov предоставляет инструменты для различных задач обработки естественного языка, включая NER. Он использует глубокие нейронные сети для обучения моделей.
5. Flair — библиотека для обработки естественного языка на основе PyTorch. Flair предоставляет инструменты для токенизации, стемминга, определения частей речи, NER и других задач. Библиотека поддерживает множество языков, в том числе русский.

Качество и производительность библиотек для NER зависят от нескольких факторов, таких как:
* Обучающие данные: библиотеки, использующие машинное обучение, требуют обучающих данных для обучения моделей. Качество этих данных может влиять на точность результатов.
* Алгоритмы: разные библиотеки могут использовать различные алгоритмы для обучения и предсказания. Алгоритмы, основанные на глубоких нейронных сетях, обычно обеспечивают более высокую точность, но требуют больше вычислительных ресурсов.
* Оптимизация: некоторые библиотеки оптимизированы для определённых задач или платформ. Например, Stanza оптимизирована для использования на мобильных устройствах.
* Поддержка языка: некоторые библиотеки лучше поддерживают определённые языки. Например, Spacy и Stanza предоставляют хорошую поддержку для русского языка.
тобы сравнить качество и производительность библиотек, можно провести следующие эксперименты:
* Использовать каждую библиотеку для выполнения задачи NER на одном и том же наборе данных. Сравнить результаты и определить, какая библиотека обеспечивает лучшую точность.
* Замерить время выполнения каждой библиотеки на одном и том же наборе данных. Определить, какая библиотека работает быстрее.
* Протестировать библиотеки на разных наборах данных и платформах. Оценить их гибкость и масштабируемость.
В целом, выбор библиотеки для NER зависит от конкретных требований проекта. Если требуется высокая точность и поддержка русского языка, то Spacy или Stanza могут быть хорошим выбором. Если важна скорость выполнения, то можно рассмотреть TextBlob или DeepPavlov.

4. Сравнение качества онлайн сервисов машинного перевода для пары русский-английский (одно направление).

Эталонный ответ
Качество онлайн-сервисов машинного перевода для пары русский-английский можно сравнить по нескольким критериям:
* Точность: насколько точно перевод отражает смысл исходного текста.
* Понятность: насколько легко понять переведенный текст.
* Грамматика: правильность грамматики и синтаксиса в переведенном тексте.
* Лексика: богатство словаря и выбор слов в переводе.
* Стиль: соответствие стиля переведенного текста стилю исходного текста.

Для сравнения качества онлайн-сервисов можно использовать следующие методы:
1. Тестирование на конкретных примерах: выбрать несколько текстов на русском языке и перевести их с помощью разных сервисов. Затем оценить качество перевода по указанным выше критериям.
2. Анализ отзывов пользователей: изучить отзывы пользователей о различных сервисах машинного перевода. Это позволит получить представление о том, какие сервисы считаются наиболее качественными.
3. Сравнение результатов с профессиональным переводом: сравнить переводы, выполненные сервисами машинного перевода, с переводами, выполненными профессиональными переводчиками. Это даст возможность оценить точность и понятность переводов.
4. Использование метрик оценки качества: существуют специальные метрики, которые

позволяют количественно оценить качество машинного перевода. Например, BLEU (Bilingual Evaluation Understudy) — это метрика, которая оценивает сходство между машинным переводом и эталонным переводом.
5. Проведение экспериментов: провести эксперименты, чтобы сравнить качество переводов, выполненных разными сервисами. В экспериментах могут участвовать люди, которые будут оценивать переводы по определенным критериям.

На основе этих методов можно составить объективное представление о качестве онлайн-сервисов машинного перевода для пары русский-английский. Однако следует учитывать, что результаты могут различаться в зависимости от конкретных условий и требований к переводу.

Вот некоторые популярные онлайн-сервисы машинного перевода для этой языковой пары:
* Яндекс Переводчик;
* PROMT;
* DeepL;
* Google Translate.

Каждый из этих сервисов имеет свои особенности и преимущества. Например, Яндекс Переводчик предлагает широкий спектр функций, таких как перевод в реальном времени, перевод видео и аудио, а также перевод с использованием искусственного интеллекта. PROMT известен своими точными переводами технических текстов. DeepL отличается высоким качеством переводов и способностью сохранять стиль исходного текста. Google Translate является одним из самых популярных сервисов машинного перевода благодаря своей доступности и простоте использования.

Выбор конкретного сервиса зависит от ваших потребностей и предпочтений. Если вам нужен точный перевод технических или научных текстов, то PROMT может быть хорошим выбором. Если вы ищете сервис, который сохраняет стиль оригинального текста, то DeepL может быть более подходящим. А если вам нужен простой и доступный сервис для перевода общих текстов, то Google Translate может быть лучшим вариантом.

5. Сравнительный анализ произведений Льва Толстого и Федора Достоевского с помощью стилометрических методов.

Эталонный ответ
Стилометрический анализ — это метод исследования текстов, основанный на количественных характеристиках. Он позволяет выявить особенности стиля автора или жанра произведения. Для анализа используются различные параметры: длина предложений, частота использования определённых слов и конструкций, синтаксические особенности и т. д.

Для сравнительного анализа произведений Льва Толстого и Фёдора Достоевского можно использовать следующие стилометрические методы:
1. Длина предложений. Можно сравнить среднюю длину предложений в произведениях обоих авторов. Это может дать представление о стиле каждого из них. Например, у Толстого предложения обычно более длинные и сложные, чем у Достоевского.
2. Частота использования определённых слов. Можно проанализировать частоту использования различных слов в произведениях Толстого и Достоевского. Это позволит выявить характерные для каждого автора слова и выражения. Например, Толстой часто использует длинные описательные конструкции, а Достоевский — короткие и лаконичные фразы.

3. Синтаксические особенности. Можно изучить синтаксическую структуру предложений в произведениях Толстого и Достоевского. Это поможет понять, как каждый автор строит свои фразы и выражает мысли. Например, Толстой использует более сложные и разветвлённые синтаксические конструкции, чем Достоевский.
4. Лексический состав. Можно провести анализ лексического состава произведений Толстого и Достоевского, чтобы определить, какие слова и выражения они используют чаще всего. Это даст представление об их стиле и тематике произведений. Например, в произведениях Достоевского часто встречаются слова, связанные с психологией и внутренним миром человека.
5. Семантический анализ. Можно применить методы семантического анализа к произведениям Толстого и Достоевского, чтобы выявить основные темы и идеи, которые они затрагивают. Это поможет лучше понять их творчество и сравнить его между собой.
6. Анализ структуры текста. Можно исследовать структуру произведений Толстого и Достоевского, включая их композицию, сюжетные линии и персонажей. Это также позволит выявить различия в их подходе к написанию.
7. Статистический анализ. Можно использовать статистические методы для сравнения произведений Толстого и Достоевского по различным параметрам. Это даст более объективное представление о различиях между ними.

Конечно, стилометрический анализ не может полностью заменить качественный анализ произведений, но он может быть полезным инструментом для изучения творчества писателей и выявления их индивидуальных особенностей.

Важно отметить, что результаты стилометрического анализа могут различаться в зависимости от выбранных параметров и методов. Поэтому важно проводить исследование с учётом всех возможных факторов и интерпретировать результаты с осторожностью.

6. Гендерное смещение (gender bias) в дистрибутивных моделях русского языка (проанализировать 2-3 статические модели отсюда: https://rusvectores.org/ru/models/).

Эталонный ответ
Гендерное смещение (gender bias) в дистрибутивных моделях русского языка — это явление, при котором модели обработки естественного языка могут демонстрировать предвзятость по отношению к определённому полу или гендеру. Это может происходить из-за того, что модели обучаются на данных, которые содержат гендерные стереотипы или предубеждения.

Для анализа гендерного смещения в дистрибутивных моделях можно рассмотреть несколько примеров моделей, доступных на сайте https://rusvectores.org/ru/models/.

1. Модель «Смысл ↔ Текст» — модель машинного перевода, которая обучена на большом объёме текстовых данных и способна генерировать тексты на основе заданных смыслов. Модель может быть подвержена гендерному смещению, если данные, на которых она обучалась, содержат стереотипные представления о мужчинах и женщинах. Например, модель может ассоциировать определённые профессии с определённым полом.

2. Модель GloVe — модель векторного представления слов, которая обучается на больших объёмах текстовых данных. Модель может использовать гендерно-предвзятые слова для

представления определённых понятий, что может привести к гендерному смещению в результатах. Например, модель может связывать определённые характеристики с мужским или женским полом.

3. Модель FastText — ещё одна модель векторного представления слов. Она также может быть подвержена гендерному смещению из-за использования гендерно-предвзятых слов для представления понятий.

Чтобы оценить степень гендерного смещения, необходимо провести анализ данных, на которых обучались модели, и результатов, которые они генерируют. Для этого можно использовать различные методы, такие как:
* Анализ ассоциаций — метод, который позволяет выявить ассоциации между словами и понятиями. Если модель связывает определённые слова с определёнными гендерными характеристиками, это может свидетельствовать о гендерном смещении.
* Тестирование на гендерную предвзятость — метод, который заключается в том, чтобы проверить, насколько точно модель классифицирует гендерные характеристики. Если модель допускает ошибки в классификации, это может указывать на гендерное смещение.

В целом, гендерное смещение в моделях обработки естественного языка является актуальной проблемой, которую необходимо учитывать при разработке и использовании таких моделей. Чтобы уменьшить гендерное смещение, можно использовать следующие подходы:
* Разнообразие данных — использование разнообразных данных, содержащих информацию о различных гендерах. Это позволит моделям обучаться на более широком спектре представлений о гендере.
* Контроль качества данных — проверка данных на наличие гендерных стереотипов и предубеждений. Это поможет предотвратить обучение моделей на предвзятых данных.
* Методы дегендеризации — применение методов, которые позволяют устранить гендерное смещение из данных или результатов. К таким методам относятся, например, методы балансировки данных или методы переобучения моделей.

7. Сравнительный анализ двух моделей вопросно-ответного поиска для русского языка с помощью инструмента CheckList (https://github.com/marcotcr/checklist).

Эталонный ответ
Вопросно-ответные системы — это информационные системы, которые позволяют пользователям задавать вопросы на естественном языке и получать ответы в виде фактов или мнений.

Для анализа двух моделей вопросно-ответного поиска для русского языка можно использовать инструмент CheckList. Это набор метрик для оценки качества работы систем обработки естественного языка. Он позволяет провести сравнительный анализ двух моделей и выявить их сильные и слабые стороны.

Основные метрики, которые можно оценить с помощью инструмента CheckList:
* Точность (Precision) — доля правильных ответов среди всех ответов, выданных системой.
* Полнота (Recall) — доля вопросов, на которые система дала правильный ответ, среди всех вопросов, заданных пользователями.
* F1-мера (F1 score) — среднее гармоническое точности и полноты. Она учитывает как точность, так и полноту ответов системы.
* Время ответа (Response time) — время, которое требуется системе для выдачи ответа на вопрос пользователя.

* Количество ошибок (Error rate) — количество ошибок, допущенных системой при ответе на вопросы пользователей.

Также можно рассмотреть дополнительные метрики, такие как покрытие (Coverage) — процент вопросов, на которые система может дать ответ, и удовлетворённость пользователей (User satisfaction) — оценка пользователями качества ответов системы.

Чтобы провести сравнительный анализ двух моделей, необходимо задать им одинаковые вопросы и сравнить результаты. Для этого можно использовать генератор вопросов, который позволит создать большое количество разнообразных вопросов по определённой теме.

После того как обе модели ответят на заданные вопросы, можно проанализировать полученные результаты и сделать выводы о том, какая модель работает лучше.

Например, одна модель может быть более точной, но менее полной, а другая — более полной, но менее точной. Также можно сравнить время ответа и количество ошибок обеих моделей.

На основе полученных результатов можно сделать вывод о том, какая из моделей является более эффективной для решения задач вопросно-ответного поиска.

Инструмент CheckList позволяет провести объективный анализ и получить достоверные результаты. Однако он не учитывает некоторые аспекты, такие как качество вопросов и контекст, в котором они задаются. Поэтому для более полного анализа необходимо также учитывать эти факторы.

Таким образом, сравнительный анализ двух моделей вопросно-ответного поиска с помощью инструмента CheckList позволит выявить их преимущества и недостатки, а также определить, какая модель лучше подходит для решения конкретных задач.

8. Анализ качества кросс-языкового переноса моделей вопросно-ответного поиска на данных без дообучения. Исходный английский набор данных – SquAD, тестирование – на данных TyDi QA.

Эталонный ответ
Кросс-языковой перенос (cross-lingual transfer) — это метод, который позволяет использовать модели, обученные на одном языке, для решения задач на другом языке. Это может быть полезно, когда данные на целевом языке ограничены или недоступны.

В контексте обработки естественного языка (NLP) кросс-языковой перенос применяется к различным задачам, таким как машинный перевод, распознавание речи и вопросно-ответный поиск. В данном случае рассматривается применение кросс-языкового переноса в задаче вопросно-ответного поиска.

Для анализа качества кросс-языкового переноса необходимо провести сравнение результатов работы модели на исходном английском наборе данных (например, SquAD) и на целевом наборе данных на другом языке (например, TyDi QA).

SquAD — это набор данных для задачи вопросно-ответного поиска, который содержит вопросы и соответствующие им ответы из статей Википедии. Он широко используется для обучения и тестирования моделей вопросно-ответного поиска.

TyDi QA — это ещё один набор данных для вопросно-ответного поиска, но уже на других языках. Он содержит вопросы и ответы на разных языках, включая английский, французский, испанский, немецкий и другие.

Чтобы провести анализ качества кросс-языкового переноса, необходимо выполнить следующие шаги:
1. Обучить модель на исходном наборе данных SquAD.
2. Перенести модель на целевой язык без дополнительного обучения (дообучения).
3. Протестировать модель на целевом наборе данных TyDi QA.
4. Оценить качество ответов модели с помощью метрик, таких как точность, полнота и F1-мера.
5. Сравнить результаты с результатами моделей, специально обученных на целевом наборе данных.

Результаты анализа могут показать, насколько хорошо модель, обученная на английском языке, может быть перенесена на другой язык без дообучения. Если результаты хорошие, то это означает, что модель может эффективно использовать общие знания о структуре вопросов и ответов, которые не зависят от конкретного языка. Однако если результаты плохие, то модель может столкнуться с трудностями при переносе на новый язык из-за различий в грамматике, лексике и семантике.

Анализ качества кросс-языкового переноса позволяет оценить потенциал моделей для использования в многоязычных системах обработки естественного языка и определить направления для дальнейшего улучшения.

Важно отметить, что результаты анализа могут зависеть от конкретных методов и параметров модели. Поэтому для получения более точных результатов рекомендуется проводить эксперименты с различными моделями и параметрами.

9. Исследование переносимости моделей распознавания юмора: исследовать 2-3 метода классификации на 2-3 англоязычных наборах данных.

Эталонный ответ
Распознавание юмора — это задача обработки естественного языка, которая заключается в определении того, является ли текст юмористическим или нет. Это может быть полезно для различных приложений, таких как анализ тональности, анализ социальных медиа и т. д.

Переносимость модели — это способность модели работать на новых данных, которые не использовались при её обучении. Это важно для того, чтобы модель была применима в реальных условиях, где данные могут быть разнообразными и непредсказуемыми.

Существует несколько методов классификации, которые можно использовать для распознавания юмора. Вот некоторые из них:
* Методы на основе правил: эти методы используют заранее определённые правила для определения юмора. Например, они могут искать определённые слова или фразы, которые обычно ассоциируются с юмором. Эти методы просты в реализации, но могут быть ограничены в своей способности распознавать новые формы юмора.
* Статистические методы: эти методы обучают модель на большом наборе данных, содержащем юмористические и не юмористические тексты. Модель использует статистические закономерности в данных для определения юмора. Эти методы могут быть более точными, чем методы на основе правил, но они требуют большого количества данных

для обучения.
* Нейронные сети: эти методы используют нейронные сети для моделирования сложных взаимосвязей между словами и контекстом. Нейронные сети могут быть обучены на больших объёмах данных и способны распознавать сложные формы юмора. Однако они также требуют больших вычислительных ресурсов для обучения и могут быть менее интерпретируемыми, чем другие методы.

Для исследования переносимости этих методов необходимо провести эксперименты на нескольких англоязычных наборах данных. Вот несколько шагов, которые можно предпринять:
1. Выбор наборов данных: выберите два или три англоязычных набора данных, содержащих юмористические и не юмористические тексты. Наборы данных должны быть достаточно разнообразными, чтобы отразить различные формы юмора.
2. Предварительная обработка данных: предварительно обработайте данные, удалив ненужные символы и разделив текст на токены (слова или символы).
3. Обучение моделей: обучите модели на каждом наборе данных отдельно. Для методов на основе правил и статистических методов используйте соответствующие алгоритмы. Для нейронных сетей используйте архитектуру, подходящую для задачи распознавания юмора.
4. Тестирование моделей: протестируйте модели на тестовых данных, не использованных при обучении. Сравните результаты с эталонными метками.
5. Анализ результатов: проанализируйте результаты тестирования, чтобы определить, насколько хорошо модели распознают юмор на новых данных. Обратите внимание на точность, полноту и F-меру.
6. Выводы: сделайте выводы о переносимости каждого метода на основе результатов тестирования. Определите, какие методы лучше всего подходят для распознавания юмора на новых данных.

Это лишь общий план исследования. В зависимости от конкретных условий и целей эксперимента, он может потребовать дополнительных шагов или изменений.

Важно отметить, что исследование переносимости требует тщательного планирования и анализа. Необходимо выбрать подходящие наборы данных, методы и метрики для оценки результатов. Также необходимо учитывать ограничения каждого метода и возможные проблемы, такие как переобучение и недообучение.

10. Анализ существующих систем вопросно-ответного поиска по базам знаний (deeppavlov, Qanswer) с помощью тестового набора данных RuBQ.

Эталонный ответ
DeepPavlov и QAnswer — это две системы вопросно-ответного поиска, которые используют обработку естественного языка (NLP) для поиска ответов на вопросы пользователей. Обе системы основаны на глубоком обучении и могут быть использованы для решения различных задач, связанных с обработкой текста.

DeepPavlov — это библиотека машинного обучения с открытым исходным кодом, которая предоставляет инструменты для создания и обучения моделей обработки естественного языка. DeepPavlov включает в себя множество готовых моделей и компонентов, которые можно легко комбинировать для создания сложных систем обработки текста. В рамках библиотеки также есть готовые модели вопросно-ответных систем.

QAnswer — это система вопросно-ответного поиска по базам знаний, разработанная компанией Яндекс. Система основана на использовании больших объёмов данных и

машинного обучения для поиска релевантных ответов на запросы пользователей. QAnswer использует различные методы обработки текста, такие как выделение ключевых слов, анализ контекста и семантический поиск, чтобы найти наиболее подходящие ответы на заданные вопросы.

Обе системы имеют свои преимущества и недостатки. DeepPavlov предоставляет широкий спектр инструментов и моделей для обработки текста, что позволяет создавать сложные системы обработки естественного языка с учётом специфики конкретной задачи. Однако для использования DeepPavlov требуется знание основ машинного обучения и обработки текста. QAnswer, в свою очередь, является готовой системой вопросно-ответного поиска, которая может быть легко интегрирована в существующие приложения и сервисы. Однако QAnswer имеет ограниченные возможности настройки и не всегда может обеспечить высокую точность ответов.

Для анализа существующих систем вопросно-ответного поиска можно использовать тестовый набор данных RuBQ. RuBQ — это набор вопросов и ответов, который был разработан специально для оценки качества работы систем вопросно-ответного поиска. Набор данных содержит вопросы на русском языке и соответствующие им ответы, что делает его подходящим для тестирования систем, работающих с русскоязычными данными.

Анализ существующих систем с помощью тестового набора данных RuBQ позволяет оценить их точность, полноту и релевантность ответов. Точность определяется как доля правильных ответов среди всех ответов, предоставленных системой. Полнота определяется как доля вопросов, на которые система смогла дать ответ. Релевантность определяется как соответствие ответа запросу пользователя.

На основе анализа результатов тестирования можно сделать выводы о сильных и слабых сторонах каждой из систем и выбрать ту, которая лучше всего подходит для конкретной задачи. Например, если требуется высокая точность ответов, то предпочтение следует отдать системе с более высокой точностью. Если же требуется широкий охват тем, то можно выбрать систему с большим количеством вопросов в базе знаний.

В целом, обе системы, DeepPavlov и QAnswer, являются эффективными инструментами для вопросно-ответного поиска и могут использоваться для различных целей. Выбор между ними зависит от конкретных требований и ограничений проекта.

11. Систематический анализ качества генерации текстов с помощью модели ruGPT-3 для различных сценариев (https://github.com/sberbank-ai/ru-gpts).

Эталонный ответ
Для анализа качества генерации текстов моделью ruGPT-3 необходимо провести ряд исследований и экспериментов, чтобы оценить её способность генерировать тексты, соответствующие определённым критериям.

1. Оценка точности: точность генерации текста можно оценить, сравнив его с исходными данными или с ожидаемым результатом. Это может быть сделано вручную или с использованием автоматических инструментов.

2. Анализ структуры и содержания: важно проверить, насколько хорошо модель понимает структуру и содержание текста. Для этого можно использовать различные метрики, такие как BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation) и другие.

3. Проверка на наличие ошибок: необходимо убедиться, что в тексте нет грамматических, орфографических и пунктуационных ошибок. Это можно сделать с помощью специальных программ или вручную.

4. Тестирование на разных данных: модель должна быть способна генерировать текст на разные темы и в разных стилях. Необходимо провести тестирование на различных наборах данных, чтобы убедиться в её универсальности.

5. Сравнение с другими моделями: полезно сравнить результаты работы ruGPT-3 с результатами других моделей, чтобы определить её преимущества и недостатки.

6. Исследование влияния параметров: можно исследовать влияние различных параметров модели на качество генерации текста, таких как размер модели, количество эпох обучения и т. д.

7. Анализ контекста: важно учитывать контекст, в котором будет использоваться модель. Например, если модель используется для генерации ответов на вопросы, то необходимо проверить её способность понимать вопросы и давать адекватные ответы.

8. Изучение влияния обучения: можно изучить, как влияет обучение модели на её способность генерировать качественный текст. Для этого можно провести эксперименты с различными методами обучения и параметрами.

9. Анализ результатов: после проведения всех исследований необходимо проанализировать полученные результаты и сделать выводы о качестве генерации текстов моделью ruGPT-3.

Эти шаги помогут провести систематический анализ качества генерации текстов моделью ruGPT-3 и получить более полное представление о её возможностях и ограничениях.

Пример использования модели ruGPT-3

Модель ruGPT-3 может быть использована для создания текстов различных жанров и стилей. Она может генерировать описания товаров, новости, статьи, диалоги и многое другое. Модель также может быть обучена на конкретных данных, например, на медицинских записях или юридических документах, чтобы улучшить её способность генерировать специализированные тексты.

Однако стоит отметить, что модель не всегда может точно передать смысл исходного текста или создать уникальный контент. Поэтому важно проводить тщательный анализ и проверку результатов перед их использованием.

В целом, модель ruGPT-3 является мощным инструментом для генерации текстов, который может быть использован в различных областях. Однако для достижения наилучших результатов необходимо проводить систематический анализ её работы и учитывать особенности каждого конкретного сценария.

ПК-3 Способен использовать современные методы обработки и интерпретации информации в профессиональной сфере

В рамках выбранной темы магистранту необходимо показать наличие знаний современных методов обработки изображений в области исследований культурного и

природного наследия, последовательно описать процедуру применения различных методов и инструментов обработки изображений на примере конкретного научно-исследовательского проекта или возможной ситуации в работе музея по выбору магистранта.

Перечень тем 2-й письменной работы с эталонными ответами:

1. Методы обработки естественного языка.

Эталонный ответ

Обработка естественного языка (Natural Language Processing, NLP) — это область искусственного интеллекта, которая занимается анализом и пониманием человеческого языка. Она включает в себя множество методов и подходов для работы с текстами на естественном языке.

Методы обработки естественного языка можно разделить на несколько категорий:

1. Морфологический анализ. Этот метод включает в себя анализ структуры слов и их форм. Он позволяет определить части речи, род, число, падеж и другие морфологические характеристики слов. Морфологический анализ является основой для многих других методов обработки естественного языка.
2. Синтаксический анализ. Синтаксический анализ позволяет определить структуру предложения и его составляющие элементы. Это включает в себя определение подлежащего, сказуемого, дополнения и других синтаксических элементов. Синтаксический анализ помогает понять смысл предложения и его связь с другими предложениями.
3. Семантический анализ. Семантический анализ направлен на понимание смысла слов и предложений. Он учитывает контекст, в котором используются слова, и определяет их значения. Семантический анализ может быть использован для определения синонимов, антонимов, гипернимов и гипонимов.
4. Распознавание речи. Распознавание речи — это метод, который позволяет преобразовывать речь в текст. Он используется для создания систем голосового управления, распознавания команд и других приложений, где требуется взаимодействие с пользователем через речь.
5. Генерация текста. Генерация текста — это процесс создания новых текстов на основе входных данных. Генерация может быть использована для создания описаний товаров, генерации ответов на вопросы и других задач, требующих создания нового контента.
6. Машинный перевод. Машинный перевод — это автоматический перевод текста с одного языка на другой. Машинный перевод использует алгоритмы для анализа исходного текста и создания перевода на целевой язык.
7. Анализ тональности. Анализ тональности позволяет определить эмоциональную окраску текста. Он может использоваться для определения настроения пользователей, анализа отзывов и других задач, связанных с эмоциями.
8. Извлечение информации. Извлечение информации — это процесс извлечения фактов и событий из текста. Извлечение информации используется для создания баз данных, поиска информации и других задач, где необходимо извлечь конкретные данные из текста.
9. Автоматическое суммирование. Автоматическое суммирование — это создание краткого изложения текста. Оно может быть использовано для создания резюме статей, новостей и других текстов.
10. Диалоговые системы. Диалоговые системы позволяют взаимодействовать с пользователями через естественный язык. Они могут использоваться для создания чат-ботов, виртуальных помощников и других диалоговых систем.

Это лишь некоторые из методов обработки естественного языка, которые используются в современных системах искусственного интеллекта. Каждый из этих методов имеет свои особенности и области применения.

2. Машинный перевод.

Эталонный ответ
Машинный перевод — это процесс автоматического перевода текста с одного языка на другой с помощью компьютерных программ.

Машинный перевод используется для упрощения и ускорения процесса перевода, особенно в тех случаях, когда требуется перевести большое количество текстов или когда необходимо быстро получить перевод. Однако машинный перевод не всегда обеспечивает высокое качество перевода, так как он может допускать ошибки и неточности.

Существует несколько подходов к машинному переводу:
* Статистический машинный перевод (SMT) использует статистические модели для определения наиболее вероятного перевода каждого слова или фразы. SMT требует большого количества обучающих данных для создания моделей.
* Нейронный машинный перевод (NMT) использует нейронные сети для обучения моделей перевода. NMT может обеспечить более высокое качество перевода по сравнению со SMT, но требует больших вычислительных ресурсов.
* Гибридный машинный перевод сочетает в себе элементы SMT и NMT для достижения наилучшего качества перевода.

Для улучшения качества машинного перевода используются различные методы, такие как:
* Предварительная обработка текста, которая включает в себя токенизацию, лемматизацию и стемминг.
* Пост-обработка текста, которая исправляет ошибки машинного перевода и улучшает читаемость перевода.
* Использование параллельных корпусов, которые содержат тексты на исходном языке и их переводы на целевой язык. Параллельные корпуса используются для обучения моделей машинного перевода.

В настоящее время машинный перевод широко используется в различных областях, таких как бизнес, наука и образование. Машинный перевод также является предметом исследований в области искусственного интеллекта и обработки естественного языка.

Преимущества машинного перевода:
* Скорость и эффективность: машинный перевод позволяет быстро и эффективно переводить большие объёмы текста.
* Доступность: машинные переводчики доступны онлайн и могут быть использованы бесплатно или за небольшую плату.
* Автоматизация: машинный перевод автоматизирует процесс перевода, что снижает вероятность ошибок, связанных с человеческим фактором.

Недостатки машинного перевода:
* Качество перевода: машинный перевод может допускать ошибки и неточности, особенно при переводе сложных текстов.
* Необходимость предварительной обработки и пост-обработки: для улучшения качества перевода требуется предварительная обработка исходного текста и пост-обработка машинного перевода.
* Ограниченность языковых пар: некоторые машинные переводчики поддерживают только определённые языковые пары.

Несмотря на недостатки, машинный перевод продолжает развиваться и совершенствоваться, и его использование становится всё более распространённым. В будущем машинный перевод будет играть ещё более важную роль в обработке естественного языка и коммуникации между людьми, говорящими на разных языках.

3. Голосовые помощники.

Эталонный ответ
Голосовые помощники — это компьютерные программы, которые способны распознавать и понимать голосовые команды человека. Они используются для выполнения различных задач, таких как поиск информации в интернете, управление устройствами умного дома, планирование дел и т. д.

Голосовые помощники работают на основе алгоритмов машинного обучения и обработки естественного языка (NLP). Они используют методы распознавания речи для преобразования аудиосигнала в текст, а затем анализируют этот текст с помощью NLP-моделей для понимания его смысла. После этого они выполняют соответствующие действия или предоставляют пользователю информацию.

Существует множество голосовых помощников, каждый из которых имеет свои особенности и функции. Некоторые из них являются самостоятельными приложениями, такими как Siri от Apple, Google Assistant от Google и Cortana от Microsoft. Другие интегрированы в операционные системы, такие как Amazon Alexa для устройств Amazon Echo и Google Home. Третьи являются частью других сервисов, таких как Яндекс Алиса, которая работает на платформе Яндекс.Браузер.

Для работы голосового помощника необходимо установить соответствующее приложение на устройство или использовать совместимое устройство с поддержкой голосового управления. Пользователь может взаимодействовать с голосовым помощником через микрофон или динамик устройства, произнося команды или задавая вопросы. Голосовой помощник будет обрабатывать эти команды и предоставлять пользователю необходимую информацию или выполнять требуемые действия.

Преимущества использования голосовых помощников:
* Удобство: голосовые помощники позволяют пользователям выполнять задачи без необходимости использования клавиатуры или мыши. Это особенно полезно для людей с ограниченными возможностями или тех, кто хочет упростить свою работу.
* Эффективность: голосовые помощники могут быстро и точно выполнять сложные задачи, такие как поиск информации или перевод текста.
* Доступность: голосовые помощники доступны на различных устройствах и платформах, что делает их доступными для широкого круга пользователей.

Однако у голосовых помощников есть и некоторые недостатки:
* Ограниченные возможности: голосовые помощники всё ещё имеют ограниченные способности к пониманию сложных запросов и контекстуальной информации.
* Проблемы с конфиденциальностью: использование голосовых помощников может привести к сбору и обработке личных данных пользователя.
* Зависимость от технологий: работа голосовых помощников зависит от качества алгоритмов и доступности интернета, что может вызвать проблемы при сбоях или отсутствии подключения.

В целом, голосовые помощники представляют собой перспективное направление развития технологий искусственного интеллекта и обработки естественного языка. Они могут стать более совершенными и функциональными в будущем, предоставляя пользователям новые возможности для взаимодействия с компьютерами и другими устройствами.

4. Анализ текстов.

Эталонный ответ
Анализ текстов — это процесс извлечения информации из текстовых данных с использованием методов обработки естественного языка (NLP). Анализ текстов включает в себя несколько этапов:

1. Предварительная обработка: на этом этапе текст преобразуется в формат, удобный для анализа. Это может включать удаление знаков препинания, приведение текста к нижнему регистру, удаление стоп-слов и т. д.

2. Токенизация: текст разбивается на отдельные слова или токены.

3. Лемматизация: слова приводятся к их словарной форме, что упрощает дальнейший анализ.

4. Стемминг: удаление окончаний слов для получения основы слова.

5. Морфологический анализ: определение морфологических характеристик слов, таких как часть речи, род, число и т. п.

6. Синтаксический анализ: построение синтаксического дерева предложения для определения его структуры.

7. Семантический анализ: выявление смысла текста на основе его лексических и синтаксических особенностей.

8. Извлечение информации: выделение ключевых фактов, событий, объектов и отношений из текста.

9. Классификация и кластеризация: группировка текстов по определённым категориям или кластерам на основе их содержания.

10. Анализ тональности: определение эмоциональной окраски текста (позитивный, негативный, нейтральный).

Для проведения анализа текстов используются различные методы и алгоритмы машинного обучения, такие как нейронные сети, деревья решений, методы кластеризации и др. В зависимости от задачи, могут применяться разные подходы к анализу текстов. Например, для классификации новостей можно использовать методы машинного обучения на основе признаков, а для выявления тем в тексте — методы тематического моделирования.

Анализ текстов находит применение в различных областях, таких как маркетинг, медицина, юриспруденция и другие. Он позволяет автоматизировать процессы обработки больших объёмов текстовых данных, выявлять закономерности и тенденции, а также принимать обоснованные решения на основе полученной информации.

5. Распознавание и синтез речи.

Эталонный ответ
Распознавание речи — это процесс преобразования аудиосигнала в текст. Системы распознавания речи используются для голосового управления устройствами, автоматического создания субтитров, перевода речи в текст и других целей.

Синтез речи — это технология, которая преобразует текст в речь. Синтезаторы речи используются в различных приложениях и устройствах, таких как голосовые помощники, системы оповещения, навигационные системы и другие.

Процесс распознавания и синтеза речи включает несколько этапов:
1. Сбор данных: на этом этапе происходит запись аудиосигнала или текста, который будет обрабатываться системой.
2. Предварительная обработка: на данном этапе аудиосигнал очищается от шумов и помех, а текст нормализуется и обрабатывается для улучшения качества распознавания.
3. Обработка сигнала: здесь происходит извлечение признаков из аудиосигнала или анализ текста для определения его структуры и содержания.
4. Классификация: на основе извлечённых признаков система определяет, какой текст или звук был произнесён.
5. Интерпретация: после классификации система интерпретирует полученный результат и выводит его в виде текста или звука.
6. Оценка: на последнем этапе система оценивает качество распознавания или синтеза и при необходимости корректирует свои параметры.

Для распознавания речи используются различные методы, такие как скрытые марковские модели (HMM), глубокие нейронные сети (DNN) и рекуррентные нейронные сети (RNN). Для синтеза речи применяются алгоритмы, основанные на моделях генерации текста и акустических моделях.

Современные системы распознавания и синтеза речи достигли высокого уровня точности и могут использоваться в различных областях, таких как медицина, образование, бизнес и другие. Однако они всё ещё имеют некоторые ограничения, связанные с акцентом, шумом и другими факторами, которые могут повлиять на качество распознавания и синтеза.

В будущем ожидается дальнейшее развитие систем распознавания и синтеза речи, включая улучшение алгоритмов обработки сигналов, использование более мощных вычислительных ресурсов и разработку новых методов обучения моделей. Это позволит повысить точность и надёжность этих систем и расширить их применение в различных сферах деятельности.

6. Обработка естественного языка на Java.

Эталонный ответ
Обработка естественного языка (Natural Language Processing, NLP) — это область искусственного интеллекта, которая занимается взаимодействием между компьютерами и человеческим языком.

Обработка естественного языка на Java может включать в себя несколько этапов:
1. Токенизация: разделение текста на отдельные слова или токены. Это может быть сделано с помощью регулярных выражений или других методов.

2. Стемминг и лемматизация: приведение слов к их основной форме. Стемминг удаляет окончания слов, а лемматизация использует словари для определения базовой формы слова.
3. Морфологический анализ: определение морфологических характеристик слов, таких как часть речи, род, число и т. д.
4. Синтаксический анализ: построение синтаксического дерева предложения. Это позволяет понять структуру предложения и его смысл.
5. Семантический анализ: определение значения слов и предложений на основе контекста.
6. Распознавание именованных сущностей: выделение имён людей, мест, организаций и других важных объектов из текста.
7. Извлечение информации: извлечение фактов и отношений из текста. Например, извлечение даты, времени, цены и т. п.
8. Генерация текста: создание нового текста на основе входных данных. Это может быть перевод, перефразирование или генерация новых предложений.
9. Анализ тональности: определение эмоциональной окраски текста. Это может использоваться для анализа отзывов клиентов или других текстовых данных.
10. Машинный перевод: перевод текста с одного языка на другой.

Для обработки естественного языка на Java можно использовать различные библиотеки и фреймворки, такие как Stanford CoreNLP, OpenNLP, GATE и другие. Эти инструменты предоставляют готовые функции для выполнения различных задач обработки естественного языка.

Также важно учитывать, что обработка естественного языка является сложной задачей, требующей глубоких знаний в области лингвистики, математики и информатики. Поэтому для успешной реализации проектов по обработке естественного языка необходимо иметь хорошее понимание этих областей.

7. Обработка естественного языка в контексте Data Science фреймворков.

Эталонный ответ
Обработка естественного языка (Natural Language Processing, NLP) — это область искусственного интеллекта и машинного обучения, которая занимается анализом и пониманием человеческого языка. Она включает в себя множество задач, таких как распознавание речи, машинный перевод, анализ тональности текста, генерация текста и другие.

В контексте Data Science фреймворков обработка естественного языка может быть использована для решения различных задач анализа данных, связанных с текстом. Вот некоторые из них:

1. Анализ тональности текста: определение эмоциональной окраски текста (позитивной, негативной или нейтральной). Это может быть полезно для анализа отзывов клиентов, новостных статей и других текстовых данных.
2. Распознавание именованных сущностей: выделение в тексте имён людей, мест, организаций и других объектов. Это может использоваться для создания каталогов товаров, поиска информации о событиях и т. д.
3. Машинный перевод: автоматический перевод текста с одного языка на другой. Это может быть полезным для международных компаний, научных исследований и других областей, где требуется перевод большого объёма текста.
4. Генерация текста: создание нового текста на основе заданного шаблона или контекста. Это может применяться для генерации описаний товаров, новостей, рекламных текстов и других видов контента.

5. Извлечение ключевых фраз: выделение наиболее важных слов и фраз из текста. Это может помочь в анализе содержания текста, поиске информации и других задачах.
6. Сентимент-анализ: определение общего настроения текста (например, положительное, отрицательное или нейтральное).
7. Классификация текста: отнесение текста к определённому классу или категории на основе его содержания.
8. Автоматическое реферирование: создание краткого изложения текста.
9. Понимание запросов на естественном языке: интерпретация запросов пользователей и предоставление соответствующих ответов.

Для обработки естественного языка в Data Science используются различные инструменты и библиотеки, такие как NLTK, SpaCy, Gensim, TensorFlow и другие. Они предоставляют широкий спектр функций для работы с текстом, включая токенизацию, лемматизацию, стемминг, векторизацию и другие операции.

Обработка естественного языка является важным инструментом для Data Scientist, который позволяет анализировать и понимать текстовые данные, а также использовать их для принятия решений и прогнозирования.

8. Обработка естественного языка с TensorFlow.

Эталонный ответ
Обработка естественного языка (Natural Language Processing, NLP) — это область искусственного интеллекта, которая занимается анализом и пониманием человеческого языка. Она включает в себя множество задач, таких как распознавание речи, машинный перевод, анализ тональности текста и т. д.

TensorFlow — это библиотека машинного обучения с открытым исходным кодом, разработанная компанией Google. Она предоставляет инструменты для создания и обучения моделей машинного обучения, а также для их развёртывания в различных средах. TensorFlow широко используется в области обработки естественного языка для решения таких задач, как:
* Распознавание именованных сущностей (Named Entity Recognition, NER) — определение имён людей, мест, организаций и других объектов в тексте.
* Анализ тональности (Sentiment Analysis) — определение эмоциональной окраски текста.
* Машинный перевод (Machine Translation) — автоматический перевод текста с одного языка на другой.
* Генерация текста (Text Generation) — создание нового текста на основе заданного контекста.
* Извлечение информации (Information Extraction) — извлечение структурированной информации из неструктурированных данных.

Для обработки естественного языка с помощью TensorFlow можно использовать различные модели машинного обучения, такие как рекуррентные нейронные сети (RNN), свёрточные нейронные сети (CNN) и трансформеры. Эти модели обучаются на больших наборах данных, содержащих тексты на естественном языке. После обучения модели могут быть использованы для выполнения различных задач обработки естественного языка, таких как классификация, кластеризация и предсказание.

Вот несколько примеров того, как можно использовать TensorFlow для обработки естественного языка:

1. Распознавание именованных сущностей: можно создать модель, которая будет определять имена людей, места, организации и другие объекты в тексте. Для этого можно использовать RNN или CNN.
2. Анализ тональности: можно создать модель, которая будет определять эмоциональную окраску текста. Для этого можно использовать регрессию или классификацию.
3. Машинный перевод: можно создать модель, которая будет переводить текст с одного языка на другой. Для этого можно использовать трансформеры, такие как BERT или GPT.
4. Генерация текста: можно создать модель, которая будет создавать новый текст на основе заданного контекста. Для этого можно использовать генеративно-состязательные сети (GAN).
5. Извлечение информации: можно создать модель, которая будет извлекать структурированную информацию из неструктурированных данных, таких как электронные письма, веб-страницы и социальные сети. Для этого можно использовать методы извлечения информации, такие как извлечение отношений (Relation Extraction) или извлечение фактов (Fact Extraction).

В целом, обработка естественного языка с TensorFlow позволяет создавать мощные и гибкие модели, которые могут решать сложные задачи анализа и понимания человеческого языка. Однако для этого требуется глубокое понимание как обработки естественного языка, так и машинного обучения.

9. Глубокое обучение при обработке естественного языка.

Эталонный ответ

Глубокое обучение — это область машинного обучения, которая использует искусственные нейронные сети для решения сложных задач. Оно позволяет компьютерам учиться на больших объёмах данных и находить сложные закономерности.

В контексте обработки естественного языка глубокое обучение используется для анализа и понимания текста, речи и других форм естественного языка. Оно может быть использовано для таких задач, как:
* Распознавание речи: глубокое обучение может использоваться для распознавания речи и преобразования её в текст. Это полезно для создания голосовых помощников, систем автоматического перевода и других приложений.
* Машинный перевод: глубокое обучение можно использовать для улучшения качества машинного перевода. Нейронные сети могут учитывать контекст и семантику слов, что позволяет им создавать более точные переводы.
* Анализ тональности: глубокое обучение также может использоваться для анализа тональности текста. Нейронные сети способны определять эмоциональное состояние автора и классифицировать текст как позитивный, негативный или нейтральный.
* Генерация текста: с помощью глубокого обучения можно создавать модели, которые генерируют текст на основе входных данных. Эти модели могут быть использованы для создания новостных статей, описаний товаров и других типов контента.

Для реализации глубокого обучения при обработке естественного языка используются различные архитектуры нейронных сетей, такие как рекуррентные нейронные сети (RNN), свёрточные нейронные сети (CNN) и трансформеры. Каждая из этих архитектур имеет свои преимущества и недостатки, и выбор зависит от конкретной задачи и доступных данных.

Рекуррентные нейронные сети (RNN) хорошо подходят для обработки последовательностей данных, таких как текст или речь. Они используют обратные связи для учёта контекста предыдущих элементов последовательности.

Свёрточные нейронные сети (CNN) часто используются для обработки изображений, но они также могут быть адаптированы для работы с текстом. CNN могут извлекать признаки из текста и использовать их для классификации или генерации.

Трансформеры — это относительно новая архитектура, разработанная для обработки естественного языка. Трансформеры основаны на механизме внимания, который позволяет им учитывать контекст при принятии решений. Они широко используются в задачах машинного перевода, анализа тональности и генерации текста.

Глубокое обучение продолжает развиваться, и новые архитектуры и методы постоянно разрабатываются. Это делает его мощным инструментом для обработки естественного языка и решения сложных задач в этой области.

10. NLP на Python.

Эталонный ответ
Обработка естественного языка (NLP) — это область искусственного интеллекта, которая занимается анализом и пониманием человеческого языка. Она включает в себя множество задач, таких как распознавание речи, машинный перевод, анализ тональности текста и многое другое.

Python — это популярный язык программирования, который широко используется для разработки различных приложений, включая приложения для обработки естественного языка. В Python существует множество библиотек и фреймворков, которые упрощают разработку NLP-приложений.

Вот некоторые из наиболее популярных библиотек и фреймворков для NLP на Python:
* NLTK (Natural Language Toolkit) — это библиотека для обработки естественного языка, которая предоставляет инструменты для токенизации, лемматизации, стемминга, синтаксического анализа и других задач.
* SpaCy — это ещё одна популярная библиотека для NLP, которая предлагает более быстрые и эффективные алгоритмы для многих задач, таких как токенизация, лемматизация, распознавание именованных сущностей и т. д.
* TensorFlow — это фреймворк машинного обучения от Google, который также можно использовать для создания моделей NLP. TensorFlow предоставляет множество инструментов для предварительной обработки данных, обучения моделей и их оценки.
* Keras — это высокоуровневый API для TensorFlow, который упрощает создание и обучение моделей глубокого обучения, включая модели для NLP.
* Hugging Face — это платформа для машинного обучения, которая предоставляет доступ к предварительно обученным моделям NLP, таким как BERT, GPT-3 и другим. Эти модели можно легко интегрировать в свои приложения с помощью API или кода.

Для начала работы с NLP на Python необходимо установить Python и необходимые библиотеки и фреймворки. Затем можно приступить к изучению основ NLP и созданию своих первых приложений. Вот несколько шагов, которые можно предпринять:
1. Изучить основы NLP, такие как токенизация, лемматизация, стемминг, синтаксический анализ и другие.
2. Познакомиться с библиотеками и фреймворками для NLP, такими как NLTK, SpaCy, TensorFlow и Keras.
3. Создать своё первое приложение для NLP, например, для распознавания именованных сущностей или анализа тональности текста.

4. Использовать предварительно обученные модели NLP для решения конкретных задач, таких как машинный перевод или генерация текста.
5. Экспериментировать с различными алгоритмами и методами для улучшения производительности и точности своих моделей.
6. Публиковать свои результаты и делиться опытом с другими разработчиками.

В целом, NLP на Python предоставляет широкие возможности для исследования и разработки приложений, которые могут понимать и анализировать человеческий язык. Это может быть полезно для различных областей, таких как медицина, финансы, маркетинг и многие другие.

11. Интеграция Spark и библиотек машинного обучения.

Эталонный ответ
Apache Spark — это фреймворк с открытым исходным кодом для распределённой обработки неструктурированных и слабоструктурированных данных. Он предоставляет набор инструментов для работы с большими данными, включая обработку в реальном времени (streaming), машинное обучение, обработку графов и SQL-запросы.

Spark поддерживает интеграцию с различными библиотеками машинного обучения, что позволяет использовать их вместе для анализа данных и решения задач машинного обучения. Вот некоторые из них:
1. MLlib (Machine Learning Library) — библиотека машинного обучения в составе Spark, которая предоставляет инструменты для различных алгоритмов машинного обучения, таких как классификация, регрессия, кластеризация и другие. MLlib интегрируется со Spark для обеспечения масштабируемости и эффективности при работе с большими объёмами данных.
2. XGBoost — популярный алгоритм градиентного бустинга, который может быть интегрирован со Spark через Spark MLlib. Это позволяет использовать XGBoost для задач классификации, регрессии и ранжирования на больших объёмах данных.
3. LightGBM — ещё один алгоритм градиентного бустинга, который также может быть интегрирован с Spark. LightGBM предлагает высокую скорость обучения и хорошую точность, что делает его полезным инструментом для задач машинного обучения на больших данных.
4. TensorFlow — популярная библиотека машинного обучения от Google, которая может быть интегрирована с Apache Spark. TensorFlow предоставляет мощные инструменты для глубокого обучения, такие как нейронные сети и свёрточные нейронные сети. Интеграция с Spark позволяет масштабировать эти алгоритмы на большие объёмы данных.
5. Scikit-learn — библиотека машинного обучения для Python, которая также может быть использована с Apache Spark через PySpark. Scikit-learn предоставляет широкий спектр алгоритмов машинного обучения, которые могут быть применены к данным, обработанным с помощью Spark.
6. H2O — платформа машинного обучения с открытым исходным кодом, которая предлагает широкий спектр алгоритмов и инструментов для машинного обучения и обработки данных. H2O также может быть интегрирована со Spark для выполнения задач машинного обучения на больших наборах данных.
7. Keras — высокоуровневый API для создания нейронных сетей, который можно интегрировать с Apache Spark для глубокого обучения на больших объёмах данных. Keras предоставляет простой и понятный интерфейс для разработки моделей нейронных сетей.
8. Caffe — фреймворк для глубокого обучения с открытым исходным кодом, который также можно интегрировать со Spark. Caffe предоставляет эффективные инструменты для обучения и использования глубоких нейронных сетей на больших данных.

9. MXNet — масштабируемый фреймворк машинного обучения, который поддерживает интеграцию со Spark. MXNet предлагает гибкость и эффективность при разработке и обучении моделей машинного обучения.

Интеграция этих библиотек с Apache Spark позволяет разработчикам использовать мощные инструменты машинного обучения для анализа больших объёмов данных и решения сложных задач. Это делает Apache Spark мощным инструментом для специалистов по обработке естественного языка, позволяя им применять алгоритмы машинного обучения к текстовым данным и извлекать из них полезную информацию.

12. Нейросетевые методы в обработке естественного языка.

Эталонный ответ

Нейросетевые методы в обработке естественного языка — это применение искусственных нейронных сетей для анализа, понимания и обработки естественного языка. Нейронные сети представляют собой математические модели, которые имитируют работу человеческого мозга и способны обучаться на больших объёмах данных.

В обработке естественного языка нейросетевые методы используются для решения различных задач, таких как:
* Распознавание речи: преобразование речи в текст или определение ключевых слов и фраз из аудиозаписей.
* Машинный перевод: перевод текста с одного языка на другой.
* Генерация текста: создание текстов на основе заданных условий или шаблонов.
* Анализ тональности: определение эмоциональной окраски текста (например, позитивный, негативный, нейтральный).
* Классификация документов: определение тематики документа или его принадлежности к определённому классу.
* Извлечение информации: выделение ключевой информации из текста.
* Ответы на вопросы: поиск ответов на заданные вопросы в тексте.

Для реализации этих задач используются различные типы нейросетей, такие как рекуррентные нейронные сети (RNN), свёрточные нейронные сети (CNN) и трансформеры. Рекуррентные нейросети хорошо подходят для работы с последовательностями данных, такими как текст или речь. Свёрточные нейросети эффективны для обработки изображений, но также могут быть адаптированы для работы с текстом. Трансформеры, основанные на архитектуре внимания, позволяют учитывать контекст при анализе данных и широко применяются в машинном переводе и генерации текста.

Преимущества нейросетевых методов в обработке естественного языка включают высокую точность и способность к обучению на больших объёмах данных, что позволяет им адаптироваться к различным задачам и условиям. Однако эти методы также требуют значительных вычислительных ресурсов и времени для обучения моделей. Кроме того, они могут быть подвержены переобучению и не всегда способны объяснить свои результаты.

Несмотря на эти ограничения, нейросетевые методы продолжают активно развиваться и находить новые применения в области обработки естественного языка, обеспечивая более точные и эффективные решения для различных задач.

13. Анализ моделей, основанных на механизме внимания и архитектуре Transformer.

Эталонный ответ

Механизм внимания (attention mechanism) — это метод, который позволяет модели машинного обучения сосредоточиться на определённых частях входных данных при выполнении задачи. Это особенно полезно в задачах обработки естественного языка (NLP), где модель должна понимать контекст и отношения между словами и фразами.

Одним из самых известных примеров использования механизма внимания является архитектура Transformer, разработанная компанией Google. Она стала основой для многих современных моделей NLP, таких как BERT, GPT-3 и другие.

Основные принципы работы архитектуры Transformer:
1. Использование механизма самовнимания (self-attention) для определения важности каждого слова или токена во входном предложении.
2. Применение нескольких слоёв (блоков) для последовательного анализа входных данных.
3. Обучение модели на больших объёмах текстовых данных с использованием методов глубокого обучения.
4. Возможность генерации текста на основе входных данных или ответов на вопросы.
5. Способность к обучению на различных языках и задачах NLP.

Архитектура Transformer имеет несколько преимуществ перед другими моделями NLP:
* Масштабируемость: благодаря использованию механизма внимания, модель может обрабатывать длинные последовательности данных без потери информации.
* Универсальность: архитектура Transformer может быть адаптирована для различных задач NLP, включая перевод, генерацию текста и ответы на вопросы.
* Эффективность: благодаря параллельной обработке данных, модель работает быстро и эффективно.

Однако у архитектуры Transformer есть и некоторые недостатки:
* Сложность: модель имеет большое количество параметров, что затрудняет её обучение и интерпретацию результатов.
* Зависимость от данных: для достижения хороших результатов модель требует большого количества обучающих данных.

В целом, архитектура Transformer является мощным инструментом для обработки естественного языка. Она позволяет создавать модели, которые могут понимать и генерировать текст на уровне, близком к человеческому. Однако для успешного применения этой архитектуры необходимо иметь достаточно данных для обучения и понимания принципов работы механизма внимания.

Для анализа моделей, основанных на механизме внимания и архитектуре Transformer, можно использовать следующие методы:

* Оценка качества перевода или генерации текста с помощью метрик, таких как точность, полнота и F-мера.
* Анализ внимания модели для понимания того, на какие части входных данных она обращает внимание при принятии решений.
* Визуализация результатов работы модели для выявления возможных проблем и ошибок.

Эти методы позволяют оценить эффективность и качество моделей, а также выявить их слабые места и области для улучшения.

Таким образом, анализ моделей, основанных на механизме внимания и архитектуре Transformer, является важным шагом для понимания их работы и возможностей. Он

позволяет разработчикам и исследователям определить сильные и слабые стороны этих моделей и использовать их для решения различных задач обработки естественного языка.

14. Анализ задач музейной работы, которые может решить NLP.

Эталонный ответ
Обработка естественного языка (NLP) — это область искусственного интеллекта, которая занимается анализом и пониманием человеческого языка. Она может быть полезна в различных областях, включая музейную работу.

Вот несколько задач музейной работы, которые могут быть решены с помощью NLP:
1. Автоматическое аннотирование экспонатов. NLP может использоваться для автоматического создания описаний экспонатов на основе их названий, описаний и других атрибутов. Это может ускорить процесс создания описаний для новых экспонатов и сделать его более точным.
2. Анализ отзывов посетителей. NLP можно использовать для анализа отзывов посетителей о музее и его экспонатах. Это позволит выявить тенденции в отзывах, определить наиболее популярные экспонаты и улучшить качество обслуживания посетителей.
3. Классификация экспонатов по темам. NLP может помочь классифицировать экспонаты по различным темам, таким как искусство, история, наука и т. д. Это облегчит поиск экспонатов для посетителей и сделает экспозицию более организованной.
4. Создание персонализированных рекомендаций. На основе анализа данных о посетителях и экспонатах, NLP может создавать персонализированные рекомендации для посетителей. Например, он может рекомендовать экспонаты, которые соответствуют интересам посетителя, или предлагать маршруты по музею, которые будут наиболее интересными.
5. Распознавание объектов на изображениях. С помощью методов компьютерного зрения и NLP можно распознавать объекты на фотографиях экспонатов. Это поможет автоматизировать процесс каталогизации экспонатов и упростить поиск информации о них.
6. Генерация описаний изображений. Используя методы обработки естественного языка, можно автоматически генерировать описания изображений экспонатов. Это будет полезно для людей с нарушениями зрения или тех, кто не владеет языком, на котором написаны описания экспонатов.
7. Поиск информации об экспонатах. NLP может быть использован для создания системы поиска информации об экспонатах музея. Посетители смогут искать экспонаты по названию, описанию или другим атрибутам.
8. Перевод описаний экспонатов. Если музей работает с посетителями из разных стран, то NLP может быть полезен для перевода описаний экспонатов на разные языки. Это сделает музей более доступным для международной аудитории.
9. Анализ посещаемости музея. NLP может анализировать данные о посещаемости музея, чтобы выявить закономерности и тенденции. Это поможет оптимизировать работу музея и повысить его эффективность.
10. Создание чат-бота для ответов на вопросы посетителей. Чат-бот, основанный на методах NLP, может отвечать на часто задаваемые вопросы посетителей о музее, экспонатах и часах работы. Это упростит процесс получения информации для посетителей.
Это лишь некоторые из возможных применений NLP в музейной работе. В зависимости от конкретных потребностей музея, могут быть разработаны и другие решения на основе NLP.

15. Предобработка текста.

Эталонный ответ

Предобработка текста — это этап обработки естественного языка (Natural Language Processing, NLP), который включает в себя ряд задач по подготовке текстовых данных для дальнейшего анализа и обработки.

Предобработка может включать в себя следующие шаги:
1. Токенизация — разделение текста на отдельные слова или токены. Это может быть сделано с помощью пробелов, знаков препинания или других разделителей.
2. Стемминг и лемматизация — приведение слов к их основной форме. Стемминг — это процесс отсечения окончаний слов, а лемматизация — более сложный процесс, который учитывает морфологию слова и приводит его к словарной форме.
3. Удаление стоп-слов — удаление часто встречающихся слов, которые не несут смысловой нагрузки, таких как предлоги, союзы и артикли.
4. Нормализация текста — преобразование текста в единый формат, например, перевод всех слов в нижний регистр или удаление специальных символов.
5. Обработка опечаток и ошибок — исправление орфографических и пунктуационных ошибок, чтобы улучшить качество текста.
6. Сегментация текста — разбиение текста на более мелкие фрагменты, такие как предложения или абзацы.
7. Фильтрация — удаление нерелевантных или нежелательных фрагментов текста, таких как спам или реклама.
8. Извлечение признаков — выделение ключевых характеристик текста, которые будут использоваться для дальнейшей обработки. Например, можно извлечь список ключевых слов или определить тональность текста.
9. Векторизация — представление текста в виде числового вектора, который будет использоваться в алгоритмах машинного обучения.

Эти шаги помогают подготовить текстовые данные для анализа, делая их более структурированными и удобными для работы алгоритмов. Предобработка является важным этапом в обработке естественного языка, поскольку она позволяет повысить точность и эффективность последующих этапов анализа.

16. Стемминг.

Эталонный ответ
Стемминг — это процесс нахождения основы слова, которая называется «стемма». Стемма — это часть слова без окончаний и суффиксов.

Стеммеры используются для приведения слов к нормальной форме, чтобы упростить их сравнение и обработку. Это особенно полезно в задачах обработки естественного языка (NLP), таких как информационный поиск, классификация документов и извлечение информации.

Существует несколько алгоритмов стемминга, но наиболее распространённым является алгоритм Портера. Он был разработан Мартином Портером в 1980 году и представляет собой набор правил для удаления окончаний и суффиксов из слов. Алгоритм Портера широко используется в различных языках и платформах.

Процесс стемминга включает следующие шаги:
1. Предварительная обработка: удаление знаков препинания, цифр и стоп-слов.
2. Разделение на слоги: слово разбивается на части, которые могут быть основой или суффиксом.
3. Применение правил: применяются правила Портера для определения основы слова.

4. Удаление окончаний: удаляются окончания, если они не являются частью основы.
5. Нормализация: основа слова приводится к стандартному виду.

Преимущества стемминга включают простоту реализации и высокую скорость работы. Однако стемминг может привести к потере семантической информации, так как разные слова могут иметь одну и ту же основу. Также стеммер может неправильно определить основу слова в некоторых случаях, что может повлиять на результаты обработки текста.

Несмотря на эти недостатки, стемминг остаётся полезным инструментом для предварительной обработки текстов в NLP. Он позволяет упростить сравнение слов и ускорить работу алгоритмов машинного обучения.

17. Лемматизация.

Эталонный ответ
Лемматизация — это процесс приведения слова к его словарной форме, то есть к лемме.

В русском языке лемма — это начальная форма слова, которая используется в словарях. Например, для глагола «бежит» лемма будет «бежать», а для прилагательного «умный» — «умный».

Лемматизация используется в обработке естественного языка (Natural Language Processing, NLP) для упрощения и стандартизации текста. Это помогает компьютерам лучше понимать и обрабатывать текст, а также облегчает поиск информации и анализ данных.

Процесс лемматизации включает в себя несколько этапов:
1. Определение части речи. Сначала необходимо определить часть речи каждого слова в тексте. Для этого используются различные методы, такие как морфологические анализаторы или правила грамматики.
2. Приведение к начальной форме. Затем каждое слово приводится к своей начальной форме. Это может быть сделано с помощью правил склонения и спряжения или с использованием словаря лемм.
3. Обработка исключений. В некоторых случаях правила лемматизации могут не работать корректно. Например, слово «лучше» может быть как наречием, так и сравнительной степенью прилагательного. В таких случаях необходимо использовать дополнительные правила или словари для правильного определения леммы.

Существуют различные алгоритмы и методы лемматизации, которые могут различаться по точности и скорости работы. Некоторые из них основаны на правилах грамматики, другие используют статистические методы или машинное обучение. Выбор метода зависит от конкретных задач и требований к обработке текста.

Преимущества лемматизации:
* Упрощение текста: Лемматизация позволяет упростить текст, удалив окончания и суффиксы, что делает его более удобным для обработки и анализа.
* Стандартизация: Приведение слов к их начальной форме помогает стандартизировать текст и сделать его более однородным.
* Улучшение поиска: Лемматизированный текст легче искать, поскольку все слова приведены к одной форме.
* Анализ данных: Лемматизация помогает анализировать данные, связанные с частотами слов, темами и другими аспектами текста.

Однако стоит отметить, что лемматизация не всегда является оптимальным решением для всех задач обработки естественного языка. В некоторых случаях более эффективным может быть использование других методов, таких как стемминг или морфологический анализ.

18. Векторизация.

Эталонный ответ
Векторизация — это процесс преобразования текстовых данных в числовые векторы, которые могут быть использованы для машинного обучения и анализа.

Векторизация текста позволяет компьютерам обрабатывать и понимать текстовые данные, что является важным шагом в обработке естественного языка (Natural Language Processing, NLP).

Существует несколько методов векторизации:
* TF-IDF (Term Frequency — Inverse Document Frequency) — метод, который учитывает частоту появления каждого слова в документе и его важность относительно всего корпуса документов. Этот метод используется для определения ключевых слов в тексте.
* Word2Vec — метод, основанный на нейронных сетях, который позволяет создавать векторные представления слов. Эти векторы отражают семантические отношения между словами, что позволяет использовать их для различных задач, таких как классификация, кластеризация и поиск похожих текстов.
* GloVe (Global Vectors for Word Representation) — ещё один метод векторизации, который также использует статистические данные о словах для создания векторных представлений. GloVe учитывает глобальные отношения между словами во всём корпусе текстов, что делает его более точным и эффективным, чем некоторые другие методы.

Процесс векторизации включает следующие шаги:
1. Предварительная обработка текста: удаление стоп-слов, приведение к нижнему регистру, токенизация (разбиение на отдельные слова или фразы) и т. д. Это помогает упростить текст и сделать его более удобным для обработки.
2. Выбор метода векторизации: выбор подходящего метода зависит от конкретной задачи и доступных данных. Например, если необходимо определить ключевые слова, то TF-IDF может быть хорошим выбором. Если же требуется создать векторные представления для использования в нейронной сети, то Word2Vec или GloVe могут быть более подходящими.
3. Применение выбранного метода: использование выбранного метода для преобразования текста в числовые векторы. В результате получается матрица, где каждая строка представляет собой векторное представление одного слова или фразы.
4. Использование векторов: полученные векторы можно использовать для различных задач машинного обучения, таких как классификация, регрессия, кластеризация и т. п.

Таким образом, векторизация является важным этапом в обработке естественного языка, позволяющим компьютерам понимать и анализировать текстовые данные.

19. Дедубликация.

Эталонный ответ
Дедубликация — это процесс удаления дублирующихся данных из массива информации. В контексте обработки естественного языка (Natural Language Processing, NLP) дедубликация может быть применена к текстам, чтобы исключить повторяющиеся фрагменты и повысить качество анализа данных.

Методы дедубликации в NLP:
1. Удаление стоп-слов. Это наиболее распространённый метод дедубликации, который заключается в удалении из текстов слов, не несущих смысловой нагрузки, таких как предлоги, союзы, артикли и т. д. Этот метод позволяет сократить размер текстов и упростить их обработку.
2. Стемминг и лемматизация. Эти методы позволяют привести слова к их основной форме, что также помогает уменьшить размер текстов и улучшить их сопоставимость.
3. Использование векторных представлений слов. Векторные представления слов позволяют представить каждое слово в виде вектора в многомерном пространстве. Это позволяет сравнивать слова по их семантической близости и удалять дубликаты.
4. Синтаксический анализ. Синтаксический анализ позволяет анализировать структуру предложений и выявлять повторяющиеся фразы или конструкции.
5. Семантический анализ. Семантический анализ позволяет учитывать контекст и смысл слов при дедубликации. Например, можно удалить дубликаты, если они имеют одинаковое значение в разных формах.
6. Применение алгоритмов машинного обучения. Алгоритмы машинного обучения, такие как кластеризация или классификация, могут быть использованы для выявления дубликатов на основе сходства между текстами.
7. Ручная проверка. В некоторых случаях может потребоваться ручная проверка результатов дедубликации для обеспечения качества.

Дедубликация является важным этапом в обработке естественного языка, поскольку она позволяет повысить точность и эффективность анализа данных, а также сократить объём обрабатываемой информации.

20. Семантический анализ.

Эталонный ответ
Семантический анализ — это процесс извлечения смысла из текста или речи. Он включает в себя понимание значения слов, фраз и предложений, а также их взаимосвязи друг с другом.

Семантический анализ является важной частью обработки естественного языка (Natural Language Processing, NLP) и используется во многих приложениях, таких как машинный перевод, поиск информации, автоматическое реферирование и т. д.

Основные задачи семантического анализа:
* Определение значения отдельных слов и словосочетаний. Это может быть сделано с помощью словарей, тезаурусов и других ресурсов.
* Понимание отношений между словами и фразами. Например, «Иван любит Машу» — это отношение любви между Иваном и Машей.
* Выявление структуры предложения и его смысла. Например, предложение «Иван купил книгу» имеет структуру подлежащее-сказуемое-дополнение и означает, что Иван совершил действие покупки книги.

Для выполнения семантического анализа используются различные методы и алгоритмы, такие как:
* Синтаксический анализ. Это процесс разбора структуры предложения с целью выявления его частей речи, грамматических категорий и связей между ними. Синтаксический анализ позволяет получить более точное представление о смысле предложения.

* Лексический анализ. Это процесс определения значения слов на основе их контекста. Лексический анализ может включать в себя такие операции, как определение части речи слова, его значения и синонимов.
* Онтологический анализ. Это процесс сопоставления слов и фраз с понятиями онтологии, которая представляет собой формальное описание предметной области. Онтологический анализ позволяет более точно понимать смысл текста.
* Тематический анализ. Это метод классификации текстов по темам на основе ключевых слов и фраз. Тематический анализ позволяет группировать тексты по их содержанию.

В целом, семантический анализ играет ключевую роль в обработке естественного языка и позволяет компьютерам лучше понимать и интерпретировать текст. Это, в свою очередь, открывает новые возможности для создания интеллектуальных систем, способных к общению и взаимодействию с людьми на естественном языке.

21. Распознавание именованных сущностей и извлечение отношений.

Эталонный ответ

Распознавание именованных сущностей (Named Entity Recognition, NER) — это задача обработки естественного языка, которая заключается в определении и классификации именованных объектов (сущностей) в тексте. Именованные сущности могут быть различными: имена людей, названия организаций, географические объекты, даты, время и т.д.

Извлечение отношений (Relation Extraction) — это процесс определения связей между именованными сущностями в тексте. Например, из текста «Джон работает в компании Microsoft» можно извлечь отношение «работает в».

Оба процесса являются важными этапами обработки естественного языка и широко используются в различных приложениях, таких как анализ данных, информационный поиск, машинный перевод и другие.

Для распознавания именованных сущностей и извлечения отношений используются различные методы машинного обучения, такие как:
* Методы на основе правил (Rule-based methods): используют заранее определённые правила для идентификации именованных сущностей. Эти методы требуют большого количества ручной работы по созданию правил, но могут обеспечить высокую точность при наличии хорошо структурированных данных.
* Статистические методы (Statistical methods): основаны на использовании статистических моделей для предсказания именованных сущностей на основе их контекста. Эти методы могут быть более гибкими, но требуют больших объёмов обучающих данных.
* Нейронные сети (Neural networks): современные методы, основанные на глубоких нейронных сетях, которые могут автоматически извлекать признаки из данных и обучаться на них. Они могут достичь высокой точности, но также требуют больших объёмов данных для обучения.

Процесс распознавания именованных сущностей включает следующие шаги:
1. Предварительная обработка: удаление стоп-слов, приведение слов к нормальной форме, токенизация и т. д.
2. Выбор признаков: определение признаков, которые будут использоваться для классификации сущностей (например, часть речи, контекст и т. п.).
3. Обучение модели: обучение модели на наборе данных с размеченными именованными сущностями.

4. Классификация: применение обученной модели к новому тексту для определения именованных сущностей.

Извлечение отношений также включает несколько этапов:
1. Определение сущностей: идентификация именованных сущностей в тексте.
2. Извлечение признаков: выделение признаков, связанных с каждой парой сущностей, например, расстояние между ними, тип отношений и т.п.
3. Применение модели: использование модели для предсказания отношений между сущностями на основе выделенных признаков.

В целом, распознавание именованных сущностей и извлечение отношений являются важными задачами обработки естественного языка. Они позволяют автоматизировать процессы анализа данных и извлечения информации из текстов, что может быть полезно во многих областях, включая бизнес, науку и образование.

22. Использование N-грамм.

Эталонный ответ
N-граммы — это последовательности из N элементов, где N может быть любым числом. В обработке естественного языка (NLP) N-граммы используются для моделирования и анализа текстовых данных.

Применение N-грамм в NLP:
1. Распознавание речи. N-граммы помогают улучшить качество распознавания речи, учитывая контекст и последовательность слов.
2. Машинный перевод. N-граммы позволяют учитывать контекст при переводе, что делает его более точным и естественным.
3. Автоматическое реферирование. N-граммы могут использоваться для выделения ключевых фраз и предложений, что упрощает процесс создания резюме текста.
4. Анализ тональности. N-граммы можно использовать для определения эмоциональной окраски текста, например, для анализа отзывов клиентов или новостных статей.
5. Генерация текста. N-граммы могут служить основой для генерации текстов, таких как описания товаров или новости.
6. Поиск информации. N-граммы помогают в поиске информации, позволяя учитывать не только отдельные слова, но и их контекст.
7. Обработка запросов. В поисковых системах N-граммы используются для понимания запросов пользователей и предоставления более точных результатов.
8. Сентимент-анализ. С помощью N-грамм можно определить отношение автора к теме или объекту, что важно для анализа мнений и отзывов.
9. Извлечение сущностей. N-граммы применяются для извлечения именованных сущностей, таких как имена людей, названия организаций и т. д., что полезно для различных задач NLP.
10. Классификация текстов. N-граммы служат признаками для классификации текстов по темам или категориям.

В зависимости от задачи и доступных данных, можно использовать разные виды N-грамм: униграммы (последовательности из одного элемента), биграммы (из двух элементов), триграммы (из трёх элементов) и так далее. Выбор подходящего размера N-грамм зависит от контекста и целей исследования.

N-граммы являются эффективным инструментом для обработки естественного языка и могут применяться в различных областях NLP. Они позволяют учесть контекст, улучшить

понимание и анализ текстовых данных, а также повысить точность и эффективность алгоритмов.

23. Частеречная разметка.

Эталонный ответ
Частеречная разметка — это процесс определения части речи каждого слова в тексте. Это один из основных этапов обработки естественного языка, который используется для анализа и понимания текста.

Части речи — это категории слов, которые используются для классификации слов по их функции в предложении. В русском языке выделяют следующие части речи: существительное, прилагательное, глагол, наречие, местоимение, числительное, предлог, союз, частица, междометие.

Частеречную разметку можно выполнить вручную или автоматически с помощью алгоритмов машинного обучения. Автоматическая разметка может быть выполнена с использованием различных методов, таких как статистические методы, методы на основе правил и методы машинного обучения.

Статистические методы основаны на использовании статистических данных о частотности слов в разных частях речи. Эти методы могут быть эффективными для языков с хорошо изученными статистическими свойствами.

Методы на основе правил используют набор правил для определения части речи слова. Эти правила могут быть основаны на морфологических, синтаксических и семантических свойствах слов. Методы на основе правил могут быть более точными, чем статистические методы, но они требуют больших усилий для разработки и настройки.

Методы машинного обучения используют алгоритмы машинного обучения для обучения модели, которая может предсказывать часть речи слова на основе его контекста. Эти методы могут быть очень точными, но они также требуют большого количества обучающих данных.

Для выполнения частеречной разметки необходимо сначала определить список частей речи, которые будут использоваться в разметке. Затем необходимо разработать алгоритм или модель, которая будет определять часть речи каждого слова. Наконец, необходимо применить этот алгоритм или модель к тексту, чтобы получить частеречную разметку.

Частеречная разметка является важным этапом обработки естественного языка и используется во многих приложениях, таких как машинный перевод, распознавание речи и анализ текста. Она позволяет лучше понять структуру и содержание текста, а также улучшить качество обработки естественного языка.

24. Библиотеки для NLP.

Эталонный ответ
Обработка естественного языка (NLP) — это область искусственного интеллекта, которая занимается анализом и пониманием человеческого языка. Для этого используются различные алгоритмы и методы машинного обучения.

Библиотеки для NLP — это наборы инструментов и функций, которые помогают разработчикам создавать приложения для обработки естественного языка. Они предоставляют доступ к различным алгоритмам и методам, таким как токенизация, лемматизация, стемминг, синтаксический анализ, семантический анализ и другие.

Вот некоторые из наиболее популярных библиотек для NLP:
1. NLTK (Natural Language Toolkit) — это библиотека для Python, которая предоставляет широкий спектр инструментов для обработки текста на естественном языке. Она включает в себя функции для токенизации, лемматизации, стемминга, синтаксического анализа, семантического анализа и других задач. NLTK также предоставляет доступ к большим наборам данных для обучения моделей.
2. SpaCy — ещё одна популярная библиотека для Python, предназначенная для обработки текста. SpaCy предоставляет более быстрые и эффективные алгоритмы для токенизации, лемматизации, стемминга и других задач, чем NLTK. Кроме того, SpaCy имеет встроенную поддержку для многих языков, включая английский, французский, немецкий и другие.
3. Gensim — библиотека для Python, которая предназначена для создания и обучения моделей тематического моделирования. Gensim предоставляет инструменты для векторизации текста, кластеризации, классификации и других задач тематического моделирования.
4. Stanford CoreNLP — набор инструментов для Java, который предоставляет множество функций для обработки текста, таких как токенизация, лемматизация, синтаксический разбор, извлечение именованных сущностей и другие. Stanford CoreNLP также имеет возможность работать с различными языками, такими как английский, испанский, китайский и другие.
5. OpenNLP — ещё один набор инструментов для Java, предназначенный для обработки текста и речи. OpenNLP предоставляет функции для токенизации, лемматизации, синтаксического разбора, распознавания именованных сущностей и других задач.
6. TextBlob — библиотека для Python, предоставляющая простые и удобные инструменты для обработки текста. TextBlob включает функции для токенизации, лемматизации, определения частей речи, извлечения именованных сущностей и других задач.
7. Pattern — библиотека для Java, которая предоставляет функции для работы с регулярными выражениями и обработки текста. Pattern может использоваться для токенизации, лемматизации и других задач обработки текста.
8. Apache OpenNLP — проект Apache Software Foundation, который предоставляет набор инструментов для обработки текста и речи на Java. Apache OpenNLP включает функции для токенизации, лемматизации, распознавания именованных сущностей, синтаксического разбора и других задач.
9. Keras — библиотека глубокого обучения для Python, которая может быть использована для создания моделей обработки естественного языка, таких как рекуррентные нейронные сети (RNN) и свёрточные нейронные сети (CNN). Keras предоставляет простой и удобный интерфейс для построения и обучения этих моделей.
10. TensorFlow — ещё одна библиотека глубокого обучения, которая также может быть использована для создания моделей NLP. TensorFlow предоставляет более мощный и гибкий инструмент для построения и обучения моделей, но требует больше усилий для настройки и использования.
Это лишь некоторые из библиотек для NLP, доступных для разработчиков. Выбор конкретной библиотеки зависит от языка программирования, задачи и требований проекта.

25. Анализ примера кода на языке Scala.

Эталонный ответ
Анализ примера кода на языке Scala может включать в себя несколько аспектов, таких как:

* Синтаксис и структура кода: изучение основных конструкций языка, таких как классы, методы, переменные, циклы и т. д.
* Обработка данных: анализ того, как код обрабатывает данные, например, с помощью функций или библиотек.
* Логика программы: понимание того, какие задачи решает программа и как она это делает.

Для анализа примера кода на Scala необходимо иметь базовые знания о языке и его синтаксисе. Вот пример анализа кода на Scala:
```scala
object Main {
  def main(args: Array[String]): Unit = {
    val numbers = List(1, 2, 3, 4, 5)
    numbers.foreach(println)
  }
}
```
В этом примере кода создаётся объект Main, который содержит метод main. Метод main принимает массив строк (args) и возвращает Unit. Внутри метода main создаётся список чисел (numbers), а затем вызывается метод foreach для этого списка. Метод foreach перебирает все элементы списка и вызывает функцию println для каждого элемента. Таким образом, этот код выводит на экран все числа из списка numbers.

Этот пример демонстрирует использование основных конструкций Scala, таких как объекты, методы и списки. Он также показывает, как можно использовать метод foreach для перебора элементов списка и выполнения над ними определённых действий.

Анализ кода может быть более глубоким и включать в себя изучение используемых библиотек, обработку ошибок и другие аспекты. Это зависит от конкретной задачи, которую решает код.

Вот некоторые дополнительные шаги, которые можно предпринять при анализе кода на Scala:
1. Изучить документацию по используемым библиотекам и функциям.
2. Проверить, есть ли в коде обработка ошибок.
3. Проанализировать, насколько эффективно код использует ресурсы.
4. Оценить, насколько код легко читать и понимать.
5. Рассмотреть, можно ли улучшить код с точки зрения производительности или читаемости.
6. Подумать, какие изменения можно внести в код, чтобы он соответствовал определённым требованиям или стандартам.
7. Провести тестирование кода, чтобы убедиться в его корректности.
8. Сравнить код с другими примерами или решениями аналогичных задач.
9. Сделать выводы о качестве кода и его пригодности для решения поставленной задачи.

26. Обзор подходов и методов к задаче автоматического извлечения именованных сущностей.

Эталонный ответ
Извлечение именованных сущностей (Named Entity Recognition, NER) — это задача обработки естественного языка, которая заключается в распознавании и классификации имён собственных, таких как имена людей, названия организаций, географические объекты и т. д. В контексте обработки текста извлечение именованных сущностей является важным

шагом для понимания содержания текста и его дальнейшего использования в различных приложениях, таких как информационный поиск, анализ социальных сетей, автоматическое аннотирование и др.

Подходы к извлечению именованных сущностей:
1. Статистические методы. Эти методы основаны на использовании статистических моделей, которые обучаются на большом наборе данных с размеченными именованными сущностями. Статистические модели могут быть обучены с использованием различных алгоритмов машинного обучения, таких как скрытые марковские модели (HMM), условные случайные поля (CRF) и нейронные сети.
2. Методы на основе правил. Эти методы используют заранее определённые правила для идентификации именованных сущностей. Правила могут быть основаны на морфологических, синтаксических или семантических характеристиках слов.
3. Гибридные методы. Гибридные методы объединяют преимущества статистических и основанных на правилах методов. Они могут использовать статистические модели для определения вероятных кандидатов на именованные сущности, а затем применять правила для уточнения результатов.
4. Нейронные сети. Современные подходы к извлечению именованных сущностей часто используют глубокие нейронные сети, такие как рекуррентные нейронные сети (RNN) и свёрточные нейронные сети (CNN). Нейронные сети позволяют автоматически извлекать именованные сущности без необходимости явного определения правил или признаков.
5. Трансформеры. Трансформеры представляют собой архитектуру нейронных сетей, основанную на механизме внимания, который позволяет учитывать контекст при обработке последовательностей. Трансформеры показали высокую эффективность в задачах обработки естественного языка, включая извлечение именованных сущностей.
6. Графовые методы. Графовые модели представляют текст в виде графа, где узлы представляют слова, а рёбра — отношения между словами. Извлечение именованных сущностей может быть выполнено путём анализа структуры графа и поиска кластеров, соответствующих именованным сущностям.
7. Обучение с подкреплением. Методы обучения с подкреплением могут использоваться для оптимизации процесса извлечения именованных сущностей путём максимизации вознаграждения за правильные предсказания.
8. Многоязычные методы. Для многоязычных задач извлечения именованных сущностей могут применяться методы, учитывающие особенности разных языков, такие как использование языковых моделей для учёта морфологии и синтаксиса.
9. Контекстуальные методы. Контекстуальные методы учитывают контекст, в котором встречаются слова, для более точного извлечения именованных сущностей. Это может включать учёт соседних слов, частей речи и других лингвистических характеристик.

Выбор подхода зависит от конкретных требований задачи, доступных ресурсов и качества данных. Важно также учитывать сложность задачи и возможность адаптации метода к новым условиям.

27. Обзор подходов и методов к задаче морфологического анализа.

Эталонный ответ
Морфологический анализ — это процесс определения морфологических характеристик слова, таких как часть речи, род, число, падеж и т. д. Это важный этап обработки естественного языка (Natural Language Processing, NLP), который используется в различных приложениях, включая машинный перевод, распознавание речи и информационный поиск.

Существует несколько подходов и методов к задаче морфологического анализа:

1. Стемминг и лемматизация. Стемминг — это метод сокращения слов до их основы или стеммы. Лемматизация — это более сложный процесс, который включает в себя определение словарной формы слова (леммы). Эти методы используются для упрощения текста и повышения эффективности поиска.
2. Правила и шаблоны. Этот подход основан на использовании правил и шаблонов для определения морфологических характеристик слов. Правила могут быть основаны на морфологии конкретного языка или на общих принципах морфологии. Шаблоны представляют собой последовательности символов, которые соответствуют определённым морфологическим характеристикам.
3. Статистические методы. Статистические модели, такие как скрытые марковские модели (Hidden Markov Models, HMM) и условные случайные поля (Conditional Random Fields, CRF), могут использоваться для предсказания морфологических характеристик слов на основе контекста. Эти модели обучаются на больших наборах данных и могут адаптироваться к различным языкам и контекстам.
4. Нейронные сети. Нейронные сети, особенно рекуррентные нейронные сети (Recurrent Neural Networks, RNN), могут быть обучены для выполнения задачи морфологического анализа. Они могут учитывать контекст и семантику слов, что позволяет им делать более точные предсказания.
5. Онтологии и словари. Онтологии и словари могут использоваться для определения морфологических характеристик слов путём сопоставления с известными категориями и значениями. Этот подход может быть эффективным для определённых языков и областей, где существуют хорошо структурированные онтологии и словари.
6. Комбинированные подходы. В реальных системах обработки естественного языка часто используются комбинированные подходы, включающие несколько из перечисленных методов. Например, статистические модели могут использоваться для предварительного анализа, а затем результаты уточняются с помощью правил и шаблонов.
Выбор подхода и метода зависит от конкретных требований и ограничений системы обработки естественного языка. Важно также учитывать точность, скорость и сложность реализации каждого подхода.

28. Способы векторного представления слов.

Эталонный ответ
Векторное представление слов — это метод, который позволяет представить слова из естественного языка в виде числовых векторов. Это делается для того, чтобы компьютер мог понимать и обрабатывать текст так же, как это делает человек. Векторное представление слов используется в различных задачах обработки естественного языка (Natural Language Processing, NLP), таких как машинный перевод, классификация текстов, поиск информации и другие.

Существует несколько способов векторного представления слов:
1. One-hot encoding — это самый простой способ представления слов. В этом методе каждому слову присваивается уникальный вектор, состоящий из нулей и единицы. Например, слово «кот» может быть представлено как [1, 0, 0], а слово «собака» — как [0, 1, 0]. Этот метод прост в реализации, но он не учитывает семантические связи между словами.
2. TF-IDF (Term Frequency — Inverse Document Frequency) — это более сложный способ представления слов, который учитывает частоту появления слова в документе и его распространённость в корпусе документов. TF-IDF представляет каждое слово как вектор, где каждый элемент вектора соответствует весу слова в данном документе или корпусе. Веса вычисляются на основе частоты появления слова и его обратной частоты в документах корпуса. Этот метод хорошо подходит для задач классификации и поиска информации.

3. Word2Vec — это один из самых популярных методов векторного представления слов, основанный на нейронных сетях. Word2Vec использует контекст, в котором появляется слово, для определения его значения. Он обучает модель предсказывать слова, которые появляются рядом с данным словом в тексте. Модель обучается на большом корпусе текстов и создаёт векторное представление каждого слова. Эти векторы можно использовать для различных задач NLP, таких как классификация, кластеризация и поиск сходства.
4. GloVe (Global Vectors for Word Representation) — ещё один метод векторного представления слов, также основанный на статистике совместного появления слов в текстах. GloVe использует матрицу совместной встречаемости слов для создания векторных представлений слов. Эта матрица содержит информацию о том, какие слова часто встречаются вместе в одном контексте. Glove использует эту матрицу для обучения модели, которая создаёт векторные представления слов таким образом, чтобы они отражали их семантические отношения.
5. BERT (Bidirectional Encoder Representations from Transformers) — это современный метод векторного представления слов, использующий глубокие нейронные сети и внимание (attention). BERT обучается предсказывать пропущенные слова в предложениях и текстах, что позволяет ему создавать векторные представления, учитывающие контекст и семантику слов. BERT широко используется в задачах машинного перевода, классификации текстов и других задачах NLP.

Выбор метода векторного представления зависит от конкретной задачи и доступных данных.

29. Автоматическое построение онтологий.

Эталонный ответ
Автоматическое построение онтологий — это процесс создания онтологии с использованием алгоритмов и методов машинного обучения. Онтологии представляют собой формальное описание понятий и отношений между ними в определённой предметной области.

Онтология — это структура данных, которая описывает основные понятия и отношения между ними. Она используется для представления знаний о предметной области и может быть использована для решения различных задач, таких как поиск информации, классификация документов и т. д.

Построение онтологий может осуществляться вручную или автоматически. Ручное построение требует значительных усилий и времени, поэтому автоматическое построение становится всё более актуальным.

Для автоматического построения онтологий используются различные методы и подходы:
* Методы извлечения терминов и определений. Эти методы основаны на анализе текстов и извлечении из них терминов и их определений. Термины могут быть извлечены с помощью методов обработки естественного языка, таких как токенизация, лемматизация и стемминг. Определения могут быть извлечены с использованием методов машинного обучения, таких как методы опорных векторов или нейронные сети.
* Методы кластеризации. Эти методы позволяют группировать термины в кластеры на основе их семантической близости. Кластеризация может быть выполнена с использованием различных алгоритмов, таких как k-means, иерархическая кластеризация и другие.
* Методы построения иерархий. Эти методы позволяют строить иерархии понятий на основе отношений между терминами. Иерархии могут быть построены с использованием

методов анализа графов, таких как алгоритмы поиска кратчайшего пути или алгоритмы обхода графа.
* Методы машинного обучения. Эти методы используют алгоритмы машинного обучения для автоматического построения онтологий. Алгоритмы машинного обучения могут быть использованы для классификации терминов, извлечения отношений и других задач.

Автоматическое построение онтологий имеет ряд преимуществ перед ручным построением:
* Автоматическое построение позволяет создавать онтологии больших объёмов данных за короткое время.
* Автоматическое построение может быть выполнено без участия человека, что снижает вероятность ошибок.
* Автоматическое построение может использовать методы машинного обучения, которые позволяют адаптировать онтологию к новым данным.

Однако автоматическое построение также имеет некоторые недостатки:
* Автоматическое построение не всегда может обеспечить высокое качество онтологии. Это связано с тем, что методы автоматического построения могут не учитывать все аспекты предметной области или могут быть подвержены влиянию шума в данных.
* Автоматическое построение требует наличия большого объёма данных для обучения алгоритмов машинного обучения. Если данных недостаточно, то качество онтологии может быть низким.

В целом, автоматическое построение онтологий является перспективным направлением исследований. Оно может использоваться для создания онтологий различных предметных областей, таких как медицина, наука, технологии и другие. Однако для достижения высокого качества онтологий необходимо разработать более совершенные методы и алгоритмы автоматического построения.

30. Постройте дерево зависимостей для следующего предложения: If you don’t know the answer, guess.

Эталонный ответ
Дерево зависимостей — это графическое представление синтаксической структуры предложения, где каждое слово связано с другими словами через синтаксические отношения.
Дерево зависимостей для предложения «If you don’t know the answer, guess» может выглядеть следующим образом:

* Предложение (sentence)
  * Сказуемое (verb phrase): don't know
    * Подлежащее (subject): you
    * Дополнение (object): the answer
  * Глагол (verb): guess

Это один из возможных вариантов построения дерева зависимостей. В зависимости от контекста и грамматики предложения, дерево может быть построено по-разному.

31. Приведите примеры инструментов/приложений/программ обработки текстов.

Эталонный ответ

Обработка естественного языка (Natural Language Processing, NLP) — это область искусственного интеллекта и компьютерных наук, которая занимается анализом и пониманием человеческого языка.

Инструменты обработки текстов — это программы и приложения, которые позволяют автоматизировать процесс анализа текстовых данных. Они могут использоваться для различных целей, таких как классификация, кластеризация, извлечение информации, перевод и т. д. Вот несколько примеров инструментов обработки текстов:

1. Gensim — библиотека Python для тематического моделирования и анализа текстов. Она позволяет создавать модели, которые представляют семантические отношения между словами в тексте. Gensim может быть использован для задач классификации, кластеризации и поиска похожих документов.
2. SpaCy — ещё одна библиотека Python, предназначенная для обработки естественных языков. SpaCy предоставляет инструменты для токенизации, лемматизации, стемминга, синтаксического разбора и других операций с текстами. SpaCy также поддерживает множество языков и может быть легко интегрирован с другими библиотеками.
3. NLTK (Natural Language Toolkit) — пакет Python для работы с естественными языками. NLTK предоставляет широкий спектр функций для обработки текстов, включая токенизацию, стемминг, лемматизацию, тегирование частей речи, анализ зависимостей и многое другое. NLTK также включает в себя набор данных для обучения и тестирования моделей.
4. Stanford CoreNLP — набор инструментов для обработки текстов на Java. Stanford CoreNLP включает в себя токенизатор, парсер, теггер частей речи, определитель именованных сущностей, а также другие функции. Stanford CoreNLP может быть использован для создания конвейеров обработки текстов.
5. Apache OpenNLP — проект Apache Software Foundation для обработки текстов. Apache OpenNLP предоставляет инструменты для распознавания именованных сущностей, определения частей речи, выделения предложений и других задач. Apache OpenNLP написан на Java и может быть интегрирован с различными платформами.
6. TextBlob — библиотека Python для обработки текстов. TextBlob предоставляет простые методы для классификации, токенизации, стемминга и других операций. TextBlob также включает в себя функции для анализа тональности и определения частей речи.
7. VADER (Valence Aware Dictionary and sEntiment Reasoner) — инструмент для анализа тональности текстов на английском языке. VADER использует словарь эмоций и правила для определения положительного, отрицательного или нейтрального тона текста. VADER может быть использован для автоматического анализа отзывов клиентов или социальных медиа.
8. SentiWordNet — лексическая база данных для английского языка, содержащая информацию о полярности слов. SentiWordNet может быть использована для определения тональности текста путём вычисления среднего значения полярности всех слов в нём.
9. Polyglot — платформа для обработки текстов, поддерживающая более 100 языков. Polyglot предоставляет инструменты для токенизации, нормализации, лемматизации и других операций. Polyglot также включает в себя модули для машинного перевода и анализа тональности.
10. Microsoft Azure Cognitive Services — облачные сервисы Microsoft для обработки изображений, видео, аудио и текстов. Azure Cognitive Services предлагает API для распознавания речи, перевода, анализа тональности, извлечения ключевых фраз и других функций.

Это лишь некоторые примеры инструментов обработки текстов. Существует множество других программ и приложений, которые могут быть использованы для анализа и

понимания текстовых данных. Выбор инструмента зависит от конкретной задачи и требований к обработке текстов.

32. Что такое корпусная лингвистика?

Эталонный ответ
Корпусная лингвистика — это раздел языкознания, который занимается разработкой, созданием и использованием текстовых корпусов.

Текстовый корпус — это большой объём текстов на одном или нескольких языках, которые используются для изучения языка и его особенностей. Корпуса могут быть использованы для различных целей, таких как изучение грамматики, лексики, синтаксиса, семантики и других аспектов языка.

Корпусы могут быть созданы для разных языков, жанров, стилей и периодов времени. Они могут содержать тексты из книг, газет, журналов, интернет-сайтов и других источников.

Создание корпуса включает в себя несколько этапов: сбор текстов, их обработку и аннотацию. Обработка текстов включает в себя удаление лишней информации (например, метаданных), приведение текстов к единому формату и кодирование. Аннотация текстов может включать в себя разметку частей речи, синтаксических структур, семантических отношений и других элементов языка.

Для создания корпусов используются специальные программы и инструменты, такие как конкордансеры, токенизаторы, парсеры и другие.

Использование корпусов позволяет исследователям изучать язык на основе реальных данных, а не только на основе теоретических предположений. Корпусные исследования позволяют получить более точные и объективные результаты, чем традиционные методы исследования.

Основные направления корпусной лингвистики включают в себя создание корпусов, разработку методов их использования, анализ языковых явлений на основе корпусов и применение корпусов в различных областях, таких как преподавание, перевод, машинный перевод и другие.

33. Изложить основные идеи теории речевых действий.

Эталонный ответ
Теория речевых действий — это подход в лингвистике и философии языка, который рассматривает речь как целенаправленное действие. Теория была разработана Дж. Остином и Дж. Сёрлем в середине XX века.

Основные идеи теории речевых действий:
1. Речевое действие — это акт произнесения высказывания, целью которого является достижение определённого эффекта. Речевые действия могут быть успешными или неуспешными в зависимости от того, достигнута ли цель.
2. Иллокутивная сила — это намерение говорящего, которое он выражает с помощью речевого действия. Иллокутивные силы могут быть разными: утверждение, вопрос, просьба, приказ и т. д.

3. Перлокутивный эффект — это результат речевого действия, то есть то, что происходит после того, как высказывание было произнесено. Перлокутивные эффекты могут быть различными: убеждение, изменение мнения, побуждение к действию и т. п.
4. Пропозициональное содержание — это информация, которая содержится в высказывании. Пропозициональные содержания могут быть истинными или ложными.
5. Условия успешности — это условия, которые должны быть выполнены для того, чтобы речевое действие было успешным. Условия успешности зависят от иллокутивной силы речевого действия. Например, для успешного вопроса необходимо, чтобы слушающий был готов ответить на него.
6. Косвенные речевые акты — это речевые действия, в которых иллокутивная сила не совпадает с пропозициональным содержанием. Косвенные речевые акты используются для выражения вежливости, иронии, сарказма и других оттенков значения.
7. Принцип кооперации — это принцип, согласно которому участники общения должны действовать так, чтобы достичь взаимопонимания. Принцип кооперации включает в себя четыре максимы: максима количества (говори столько, сколько нужно), максима качества (говори правду), максима отношения (будь релевантным) и максима способа (говори ясно).

Теория речевых действий имеет важное значение для понимания того, как работает язык в процессе коммуникации. Она позволяет анализировать речевые действия с точки зрения их целей, намерений и эффектов.

34. Пояснить принцип работы наивного Байесовского классификатора.

Эталонный ответ
Наивный байесовский классификатор — это простой и эффективный алгоритм машинного обучения, который используется для решения задач классификации. Он основан на теореме Байеса и предполагает, что все признаки (характеристики) объекта независимы друг от друга.

Принцип работы наивного байесовского классификатора можно описать следующим образом:
1. Сбор данных. На этом этапе собираются данные, которые будут использоваться для обучения модели. Данные должны быть представлены в виде набора объектов с известными классами.
2. Выбор признаков. Из всех доступных признаков выбираются те, которые наиболее важны для задачи классификации. Это может быть сделано вручную или с помощью методов отбора признаков.
3. Обучение модели. Модель обучается на основе собранных данных. Для этого вычисляются вероятности принадлежности каждого признака к каждому классу. Эти вероятности используются для определения вероятности того, что объект принадлежит к определённому классу.
4. Классификация новых объектов. Когда модель обучена, она может использоваться для классификации новых объектов. Для этого новый объект представляется модели в виде вектора признаков, и модель определяет вероятность того, что этот объект принадлежит каждому из известных классов. Объект классифицируется как класс с наибольшей вероятностью.

Наивный байесовский классификатор имеет несколько преимуществ:
* Простота реализации.
* Эффективность при работе с большими объёмами данных.
* Возможность использования различных функций распределения вероятностей для признаков.

Однако у него есть и некоторые недостатки:
* Наивный байесовский классификатор предполагает независимость признаков, что не всегда соответствует действительности.
* Модель может быть чувствительна к выбросам — аномальным значениям признаков.

В целом, наивный байесовский классификатор является простым и эффективным алгоритмом машинного обучения, который широко используется в различных областях, таких как обработка естественного языка, распознавание речи и анализ данных.

35. В чем отличие между классификацией и кластеризацией текстов?

Эталонный ответ
Классификация и кластеризация — это два метода обработки естественного языка, которые используются для анализа текстовых данных.

Классификация текстов — это метод машинного обучения, при котором тексты распределяются по заранее определённым категориям на основе их содержания. В процессе классификации модель обучается на наборе данных с уже известными категориями и затем может автоматически классифицировать новые тексты. Например, классификация может использоваться для определения тематики текста (наука, искусство, спорт и т. д.), его эмоциональной окраски (позитивно, негативно, нейтрально) или автора (писатель, журналист, блогер).

Для классификации используются различные алгоритмы машинного обучения: наивный байесовский классификатор, метод опорных векторов, деревья решений и другие. Выбор алгоритма зависит от конкретной задачи и характеристик данных.

Кластеризация текстов — это процесс группировки схожих текстов в кластеры без использования заранее определённых категорий. Модель самостоятельно определяет, какие тексты похожи друг на друга, и объединяет их в группы. Кластеризация может быть полезна для выявления скрытых закономерностей в данных, поиска похожих документов или упрощения анализа больших объёмов информации.

В отличие от классификации, кластеризация не требует предварительного определения категорий и может быть более гибкой. Однако она также может потребовать больше времени и ресурсов для обучения модели.

Таким образом, основное отличие между классификацией и кластеризацией заключается в том, что классификация использует заранее определённые категории, а кластеризация группирует тексты на основе их сходства. Оба метода имеют свои преимущества и недостатки и могут применяться в зависимости от конкретных задач и требований к анализу данных.

36. Объяснить принципы использования регулярных выражений в поисковых системах.

Эталонный ответ
Регулярные выражения (regular expressions) — это формальный язык поиска и осуществления манипуляций с подстроками в тексте, основанный на использовании метасимволов.

Поисковые системы используют регулярные выражения для:
* Поиска по ключевым словам. Поисковые системы могут использовать регулярные выражения для более точного поиска по ключевым словам. Например, если пользователь ищет «телефонный номер», то регулярное выражение может использоваться для поиска только тех строк, которые содержат формат телефонного номера.
* Фильтрации результатов поиска. Регулярные выражения позволяют поисковым системам фильтровать результаты поиска на основе определённых критериев. Например, можно использовать регулярное выражение для исключения определённых слов или фраз из результатов поиска.
* Анализа и обработки данных. Регулярные выражения могут быть использованы для анализа и обработки текстовых данных, таких как имена файлов, адреса электронной почты и другие структурированные данные. Это позволяет поисковым системам лучше понимать содержание документов и предоставлять более точные результаты поиска.

Основные принципы использования регулярных выражений в поисковых системах включают:
1. Выбор метасимволов. Метасимволы используются для определения шаблонов поиска. Они могут включать символы, такие как * (любой символ), ? (один любой символ) и [ ] (набор символов).
2. Создание шаблона. Шаблон регулярного выражения определяет критерии поиска. Он может включать комбинации метасимволов и обычных символов.
3. Применение шаблона к тексту. Поисковая система применяет шаблон регулярного выражения к тексту, чтобы найти совпадения.
4. Обработка результатов. После нахождения совпадений поисковая система обрабатывает результаты и предоставляет их пользователю.

Использование регулярных выражений позволяет поисковым системам более точно и гибко обрабатывать запросы пользователей, что приводит к более точным и релевантным результатам поиска.

37. Привести классификацию речевых действий.

Эталонный ответ

Речевые действия можно классифицировать по различным критериям. Вот некоторые из возможных классификаций:

1. По цели высказывания:
   * Повествовательные (сообщают о каком-либо факте или событии).
   * Вопросительные (задают вопрос).
   * Побудительные (побуждают к действию).

2. По эмоциональной окраске:
   * Нейтральные (не выражают эмоций).
   * Эмоционально окрашенные (выражают эмоции).

3. По форме:
   * Устные (произносятся вслух).
   * Письменные (фиксируются на бумаге или другом носителе).

4. По структуре:
   * Простые (состоят из одного предложения).
   * Сложные (состоят из нескольких предложений).

5. По функции в коммуникации:
   * Информативные (передают информацию).
   * Регулятивные (регулируют поведение собеседника).
   * Оценочные (оценивают что-либо).
   * Этикетные (соблюдают этикет).

6. По степени воздействия на собеседника:
   * Убеждающие (стремятся убедить собеседника в чём-либо).
   * Внушающие (стремятся внушить собеседнику определённое отношение или действие).

7. По характеру взаимодействия с собеседником:
   * Диалогические (происходят в процессе диалога).
   * Монологические (происходят без непосредственного участия собеседника).

Это лишь некоторые возможные классификации речевых действий. В зависимости от контекста и целей исследования могут быть предложены и другие классификации.

Речевые действия играют важную роль в общении и взаимодействии людей. Они позволяют передавать информацию, выражать эмоции, регулировать поведение и т. д. Понимание различных типов речевых действий может помочь более эффективно общаться и достигать поставленных целей.

38. В чем отличие между тезаурусами, семантическими сетями и онтологиями?

Эталонный ответ
Тезаурус — это словарь, который содержит термины и понятия из определённой области знаний или предметной области. Тезаурусы используются для представления семантических отношений между терминами, такими как синонимия, антонимия и гиперонимия (отношение «общее-частное»).

Семантическая сеть представляет собой графическую модель, в которой узлы представляют собой понятия или объекты, а дуги — отношения между ними. Семантические сети позволяют моделировать сложные отношения между объектами и понятиями. Они широко используются в искусственном интеллекте и обработке естественного языка.

Онтология — формальная спецификация концептуализации предметной области, которая используется для моделирования понятий и отношений между ними. Онтологии позволяют описывать сложные структуры данных и знания о предметной области в виде формальной модели. Они могут быть использованы для обмена знаниями между различными системами и приложениями.

Основные различия между этими подходами:
* Цель: тезаурусы предназначены для поиска синонимов и связанных терминов, семантические сети — для моделирования сложных отношений между объектами, онтологии — для формального описания понятий и их отношений.
* Уровень абстракции: тезаурусы обычно содержат более общие термины, семантические сети могут включать более конкретные отношения, онтологии предоставляют более формальное и точное описание понятий.
* Формализм: тезаурусы используют простые структуры, такие как списки терминов и их синонимов, семантические сети используют графы для представления отношений, онтологии используют формальные языки, такие как OWL (Web Ontology Language).

В целом, эти подходы дополняют друг друга и могут использоваться совместно для создания более точных и полных моделей представления знаний. Выбор конкретного подхода зависит от целей и задач проекта.

39. Описать методы анализа социальных сетей.

Эталонный ответ

Анализ социальных сетей — это метод исследования, который позволяет изучать отношения и взаимодействия между людьми, группами и организациями. Социальные сети могут быть представлены в виде графов, где вершины представляют объекты (люди, группы, организации), а рёбра — связи между ними.

Методы анализа социальных сетей:

1. Центральность. Этот метод позволяет определить наиболее влиятельных участников социальной сети. Существует несколько видов центральности:
   * Степень центральности: показывает количество связей, которые есть у объекта. Чем больше связей у объекта, тем больше его влияние на сеть.
   * Близость центральности: измеряет, насколько быстро информация может распространяться от одного объекта к другому через кратчайший путь.
   * Междуness центральность: оценивает, сколько кратчайших путей проходит через объект. Объекты с высокой междуness центральностью играют важную роль в передаче информации между другими объектами.
2. Структурные дыры. Метод позволяет выявить пробелы в структуре социальной сети, которые могут использоваться для получения преимущества или контроля над информацией. Структурные дыры могут возникать, когда два объекта связаны друг с другом через посредника, но не напрямую.
3. Кластеризация. Метод используется для выявления групп объектов, которые тесно связаны друг с другом. Кластеризация может помочь понять структуру социальной сети и выявить скрытые закономерности.
4. Анализ сообществ. Метод направлен на выявление групп объектов, которые имеют общие характеристики или интересы. Анализ сообществ может помочь лучше понять динамику социальной сети и предсказать её развитие.
5. Анализ распространения информации. Метод изучает, как информация распространяется по социальной сети. Это может быть полезно для понимания того, какие факторы влияют на распространение информации и как можно улучшить этот процесс.
6. Визуализация. Методы визуализации позволяют представить социальную сеть в графическом виде, что упрощает её анализ и понимание. Визуализация может включать в себя использование диаграмм, графиков и других инструментов.
7. Статистический анализ. Статистические методы используются для анализа данных о социальных сетях. Они могут включать в себя анализ корреляции, регрессии и другие методы.
8. Сетевой анализ. Сетевой анализ включает в себя изучение структуры и динамики социальных сетей. Он может использоваться для определения ключевых узлов, выявления сообществ и анализа распространения информации.
9. Анализ атрибутов. Этот метод включает в себя исследование характеристик объектов в социальной сети, таких как возраст, пол, местоположение и т. д. Эти атрибуты могут влиять на поведение объектов и их взаимодействие в сети.

Эти методы могут применяться вместе или по отдельности в зависимости от целей исследования. Важно отметить, что анализ социальных сетей требует тщательного подхода к сбору и обработке данных, чтобы обеспечить точность и достоверность результатов.

40. Привести примеры использования методов кластеризации и классификации для определения авторства текстов.

Эталонный ответ
Методы кластеризации и классификации — это два основных подхода к обработке естественного языка (NLP). Они используются для решения различных задач, включая определение авторства текстов.

Кластеризация — метод машинного обучения без учителя, который позволяет группировать тексты на основе их сходства. В контексте определения авторства кластеризация может использоваться для выявления групп текстов, написанных одним автором, или для обнаружения плагиата.

Например, можно использовать кластеризацию для анализа корпуса текстов, написанных разными авторами. Алгоритм будет искать общие черты в стиле письма каждого автора и объединять тексты в соответствующие группы. Это может помочь определить, какие тексты были написаны конкретным автором.

Другой пример использования кластеризации — обнаружение плагиата. Если два текста имеют схожие характеристики, алгоритм может объединить их в один кластер. Это указывает на то, что эти тексты могут быть плагиатом друг друга.

Классификация — метод машинного обучения с учителем, который предполагает обучение модели на наборе данных с известными метками. Затем модель может классифицировать новые данные на основе этих меток. В контексте определения авторства классификация может использоваться для создания модели, которая может предсказывать автора нового текста.

Для этого необходимо обучить модель на корпусе текстов с известными авторами. Модель будет изучать особенности стиля письма каждого автора и создавать представление о том, как каждый автор пишет. Затем, когда модель получит новый текст, она сможет предсказать, кто его написал, основываясь на этом представлении.

Вот несколько примеров использования классификации для определения авторства:
* Судебные процессы: классификация может быть использована для определения автора анонимного сообщения или документа, связанного с судебным процессом.
* Литературоведение: классификация может помочь исследователям определить, какой писатель создал конкретный текст, если авторство неясно.
* Журналистика: классификация может применяться для проверки подлинности статей или репортажей, особенно в случаях, когда авторство вызывает сомнения.

В целом, методы кластеризации и классификации являются мощными инструментами для определения авторства текстов. Они могут использоваться вместе или по отдельности в зависимости от конкретной задачи и доступных данных.

41. Дать определения лингвостатистических параметров, авторского инварианта и лингвистических спектров.

Эталонный ответ
Лингвостатистические параметры — это количественные характеристики, которые используются для анализа и описания языковых явлений. Они могут быть использованы

для изучения различных аспектов языка, таких как частота использования слов, длина предложений, структура текста и т. д.

Лингвостатистические методы позволяют получить объективные данные о языке, которые могут быть полезны для лингвистических исследований, машинного перевода, автоматического распознавания речи и других областей, связанных с обработкой естественного языка.

К лингвостатистическим параметрам относятся:
* Частота слов: количество раз, которое слово встречается в тексте или корпусе текстов. Это один из наиболее важных параметров, который может использоваться для определения ключевых слов и тем в тексте.
* Длина слов: средняя или максимальная длина слов в тексте. Этот параметр может быть использован для оценки сложности текста или для создания алгоритмов сжатия текста.
* Длина предложений: средняя или максимальная длина предложений в тексте. Этот параметр также может быть использован для оценки сложности текста.
* Распределение частей речи: процентное соотношение различных частей речи (существительных, глаголов, прилагательных и т.д.) в тексте. Это может дать представление о стиле и структуре текста.
* Лексическое разнообразие: количество уникальных слов или словосочетаний в тексте. Лексическое разнообразие может указывать на богатство языка автора или на его способность к выражению мыслей.

Авторский инвариант — это набор характеристик, которые остаются неизменными для всех произведений одного автора. Авторский инвариант может включать в себя такие параметры, как стиль письма, выбор слов, синтаксические конструкции и другие особенности, которые делают произведения автора узнаваемыми.

Авторский инвариант можно использовать для идентификации автора текста, а также для анализа его творчества и стиля. Например, можно сравнить авторский инвариант двух авторов, чтобы определить, насколько они похожи или различны.

Для определения авторского инварианта можно использовать различные методы, такие как анализ частотности слов, анализ синтаксических конструкций, анализ семантических связей между словами и другие.

Лингвистические спектры — это метод представления информации о тексте в виде графика или таблицы, где каждая ось соответствует определённому параметру текста. Лингвистические спектры могут использоваться для визуализации данных о тексте, а также для сравнения текстов между собой.

Например, можно построить лингвистический спектр, где по одной оси будет отложена частота слов, а по другой — их длина. Такой спектр позволит увидеть, какие слова чаще всего встречаются в тексте и какова их средняя длина.

Также можно построить лингвистический спектр, где будут сравниваться два текста по различным параметрам, таким как распределение частей речи, лексическое разнообразие и т. п. Это позволит увидеть различия между текстами и сделать выводы об их авторстве или стиле.

42. Перечислить методы задания синтаксической структуры предложений.

Эталонный ответ
Синтаксис изучает способы соединения слов и форм слов в словосочетания и предложения, а также типы синтаксических связей. Синтаксическая структура предложения — это его строение, определённое с точки зрения синтаксиса.

Для задания синтаксической структуры предложения используют различные методы:
1. Метод непосредственных составляющих (НС). Предложение рассматривается как иерархия вложенных друг в друга элементов, каждый из которых может быть разложен на более мелкие составляющие. Этот метод позволяет анализировать структуру предложения по уровням вложенности.
2. Грамматика зависимостей. В этом методе предложение рассматривается как сеть зависимостей между словами, где каждое слово зависит от другого слова или группы слов. Зависимости могут быть разного типа: грамматические, семантические, логические и т. д.
3. Грамматика составляющих. Предложение разбивается на составляющие, которые могут быть словосочетаниями или отдельными словами. Составляющие соединяются между собой определёнными типами связей, такими как согласование, управление, примыкание и др.
4. Трансформационный метод. Метод основан на предположении о существовании ядерных структур, из которых путём трансформаций можно получить все остальные предложения. Трансформации — это преобразования, которые изменяют структуру предложения без изменения его смысла.
5. Лексико-грамматический метод. Этот метод учитывает лексическое значение слов при анализе их синтаксических функций. Например, глагол «быть» может выступать в роли связки, вспомогательного глагола или полнозначного глагола.
6. Семантико-синтаксический метод. При анализе синтаксической структуры учитывается семантика слов и отношений между ними. Например, в предложении «Он читает книгу» глагол «читает» обозначает действие, которое выполняет субъект «он» над объектом «книга».
7. Формально-семантический метод. Сочетает формальный и семантический подходы к анализу синтаксической структуры. Формальная сторона включает в себя анализ синтаксических отношений между словами, а семантическая — учёт значения этих отношений.
8. Когнитивный подход. Рассматривает синтаксическую структуру как отражение когнитивных процессов, таких как восприятие, внимание, память и мышление.
9. Корпусный подход. Основан на использовании корпусов текстов для анализа синтаксической структуры реальных высказываний. Корпуса позволяют выявить частотные модели синтаксических конструкций и их значения.

Выбор метода зависит от целей исследования и особенностей анализируемого материала.

43 Привести примеры мер семантической близости.

Эталонный ответ
Меры семантической близости — это методы и алгоритмы, которые позволяют оценить сходство или близость между двумя или более текстовыми объектами (например, словами, фразами или документами). Они используются для определения семантических отношений между этими объектами.

Существует несколько мер семантической близости, каждая из которых имеет свои особенности и может быть более подходящей для определённых задач. Вот некоторые из них:

1. Косинусное подобие (Cosine similarity) — одна из самых популярных мер семантической близости. Она основана на косинусе угла между векторами, представляющими текстовые объекты в многомерном пространстве. Чем ближе угол к нулю, тем больше похожи объекты. Косинусное подобие широко используется в задачах информационного поиска, кластеризации и классификации текстов.
2. Расстояние Левенштейна (Levenshtein distance) — мера, которая определяет количество операций вставки, удаления и замены символов, необходимых для преобразования одного слова или фразы в другое. Это расстояние часто используется для сравнения текстовых объектов, таких как имена, адреса и другие строки.
3. Сходство Жаккара (Jaccard similarity) — мера сходства, основанная на пересечении множеств. Она вычисляет отношение количества общих элементов к количеству уникальных элементов в двух наборах данных. Сходство Жаккара часто применяется в задачах анализа текста, где необходимо определить общие темы или ключевые слова.
4. Расстояния Хэмминга (Hamming distance) — метрика, используемая для измерения разницы между двумя последовательностями одинаковой длины. В контексте обработки естественного языка она может применяться для оценки схожести предложений или документов.
5. Векторное пространство (Vector space model) — метод представления текстовых данных в виде векторов в многомерном пространстве, где каждый вектор представляет собой документ или слово. Мерой семантической близости в этом случае является косинус угла между векторами.
6. Латентное размещение Дирихле (Latent Dirichlet allocation) — вероятностная модель, которая позволяет группировать документы по темам и оценивать их семантическую близость на основе этих тем.
7. Мера TF-IDF (Term frequency — inverse document frequency) — статистическая мера, которая учитывает частоту появления слова в документе и его распространённость во всей коллекции документов. Эта мера часто используется для определения ключевых слов и оценки семантической близости между документами.
8. Метод главных компонент (Principal component analysis) — математический метод, который позволяет уменьшить размерность данных и выделить основные компоненты, объясняющие большую часть дисперсии. В обработке естественного языка этот метод может использоваться для выявления основных тем или категорий в наборе документов.
Выбор конкретной меры семантической близости зависит от задачи и требований к точности и скорости вычислений.

44. Дать определения отмеченных последовательностей, контекста, дистрибутивных классов.

Эталонный ответ
Последовательность — это упорядоченный набор элементов, в котором каждый элемент имеет свой уникальный индекс. В контексте обработки естественного языка последовательности могут быть использованы для представления текстов, где элементами являются слова или символы.

Определение:

* Отмеченные последовательности — это последовательности, в которых некоторые элементы (или все элементы) помечены специальными символами или метками. Эти метки могут использоваться для обозначения различных аспектов текста, таких как части речи, семантические роли и т. д. Отмеченные последовательности используются в задачах обработки естественного языка, таких как синтаксический анализ, машинный перевод и другие.

Контекст — это окружение, в котором используется слово или фраза. Контекст может влиять на значение слова или фразы, так как он предоставляет дополнительную информацию о ситуации, в которой они используются. Например, слово «стол» может иметь разные значения в зависимости от контекста: «обеденный стол», «письменный стол» и т.д.

Дистрибутивные классы — это группы слов или фраз, которые имеют похожие дистрибутивные свойства. Дистрибуция — это способ, которым слова или фразы сочетаются с другими словами или фразами в тексте. Слова или фразы из одного дистрибутивного класса имеют схожие дистрибутивные характеристики, то есть они часто встречаются в похожих контекстах. Это позволяет использовать дистрибутивные классы для задач обработки естественного языка, таких как кластеризация, тематическое моделирование и другие.

Важно отметить, что определение дистрибутивных классов может различаться в зависимости от конкретной задачи и метода анализа. Однако, общая идея заключается в том, чтобы группировать слова или фразы на основе их сходства в использовании и контекстах, в которых они встречаются.

45. Перечислить методы классификации и кластеризации текстовой информации. Сформулировать основные принципы.

Эталонный ответ
Классификация и кластеризация текстовой информации — это методы обработки естественного языка, которые используются для анализа и структурирования текстовых данных.

Методы классификации:
* Метод опорных векторов (SVM) — метод машинного обучения, который используется для классификации текстов. Он основан на идее построения гиперплоскости, которая разделяет данные на два класса. Для этого используются опорные векторы — точки данных, которые находятся ближе всего к разделяющей гиперплоскости.
* Дерево решений — метод классификации, основанный на построении дерева решений. Дерево решений представляет собой структуру, состоящую из узлов и рёбер. В узлах дерева принимаются решения о принадлежности текста к тому или иному классу, а рёбра представляют собой условия, при которых принимается то или иное решение.
* Наивный байесовский классификатор — статистический метод классификации, который основан на теореме Байеса. Этот метод предполагает, что признаки текста независимы друг от друга. Это предположение является наивным, но оно позволяет упростить процесс классификации.
* Логистическая регрессия — метод, используемый для прогнозирования вероятности того, что текст принадлежит определённому классу. Логистическая регрессия основана на использовании логистической функции для преобразования линейного предиктора в вероятность.
* Случайный лес — ансамблевый метод классификации, который объединяет результаты нескольких деревьев решений. Случайный лес работает лучше, чем отдельное дерево решений, за счёт уменьшения переобучения и повышения обобщающей способности модели.

Основные принципы классификации:

1. Выбор признаков. Необходимо выбрать те признаки, которые наиболее важны для классификации текста. Признаки могут быть основаны на словах, словосочетаниях, частях речи, синтаксических конструкциях и т. д.
2. Обучение модели. Модель обучается на наборе данных, содержащем тексты с известными классами. Модель должна научиться предсказывать класс нового текста на основе его признаков.
3. Тестирование модели. После обучения модель тестируется на новом наборе данных, чтобы оценить её точность и надёжность.
4. Интерпретация результатов. Результаты классификации могут быть интерпретированы для понимания того, какие признаки наиболее важны для определения класса текста.

Методы кластеризации:
* K-средних — один из самых популярных методов кластеризации. Он работает путём минимизации суммарного квадратичного отклонения точек кластера от центра кластера. K-средних требует указания количества кластеров, которое должно быть известно заранее.
* Иерархическая кластеризация — метод кластеризации, который строит иерархию кластеров. Иерархия может быть представлена в виде дендрограммы, где каждый узел соответствует кластеру, а рёбра указывают на связь между кластерами.
* DBSCAN — алгоритм кластеризации на основе плотности. DBSCAN ищет плотные области данных и рассматривает их как кластеры. Алгоритм требует задания двух параметров: радиуса окрестности и минимального количества точек в окрестности.
* EM-алгоритм — вероятностный алгоритм кластеризации, основанный на методе максимального правдоподобия. EM-алгоритм итеративно обновляет параметры модели до тех пор, пока не будет достигнута сходимость.

Основные принципы кластеризации:
1. Определение меры расстояния. Необходимо определить меру расстояния между текстами, которая будет использоваться для кластеризации. Мера расстояния может быть основана на сходстве слов, семантическом сходстве и т. п.
2. Применение алгоритма кластеризации. Алгоритм кластеризации используется для группировки текстов в кластеры на основе выбранной меры расстояния.
3. Оценка качества кластеризации. Качество кластеризации можно оценить по таким критериям, как количество кластеров, размер кластеров, степень перекрытия кластеров и т. д.

46. Применение частотных методов в компьютерной лингвистике.

Эталонный ответ
Частотные методы широко применяются в компьютерной лингвистике для решения различных задач обработки естественного языка (Natural Language Processing, NLP). Вот некоторые из них:

1. Частотный анализ слов и словосочетаний. Этот метод используется для определения наиболее часто встречающихся слов или словосочетаний в тексте. Это может быть полезно для определения тематики текста, выделения ключевых слов, а также для создания автоматических аннотаций и резюме.

2. Анализ тональности текста. Частотный анализ может использоваться для определения эмоциональной окраски текста. Например, если в тексте часто встречаются слова, связанные с положительными эмоциями, то можно сделать вывод, что текст имеет позитивный тон. И наоборот, если часто встречаются слова с негативной окраской, то текст будет иметь негативный тон.

3. Распознавание именованных сущностей. Частотный анализ используется для распознавания именованных сущностей, таких как имена людей, названия организаций, даты и т. д. В этом случае частотный анализ применяется для определения наиболее вероятных кандидатов на роль именованной сущности.

4. Определение семантических отношений между словами. Частотный анализ позволяет определить семантические отношения между словами, такие как синонимия, антонимия и гипонимия. Это может быть использовано для построения семантических сетей и других моделей представления знаний.

5. Автоматическое реферирование. Частотный анализ помогает выделить ключевые фразы и предложения в тексте, которые наиболее точно отражают его содержание. Эти фразы и предложения могут быть использованы для автоматического реферирования текста.

6. Классификация текстов. Частотный анализ может быть использован для классификации текстов по их тематике или жанру. Для этого необходимо определить частотные слова и словосочетания, характерные для каждой категории текстов.

7. Извлечение информации. Частотный анализ может помочь извлечь информацию из текста, такую как факты, события, характеристики объектов и т. п. Для этого нужно определить частотные словосочетания, которые содержат нужную информацию.

8. Поиск информации. Частотные методы используются для поиска информации в больших объёмах текстовых данных. Они позволяют определить наиболее релевантные документы или фрагменты документов, соответствующие запросу пользователя.

9. Обработка запросов на естественном языке. Частотный анализ может применяться для обработки запросов на естественном языке, например, для определения смысла запроса, его темы и цели.

В целом, частотные методы являются мощным инструментом для анализа и обработки текстовых данных в компьютерной лингвистике. Они помогают выявить закономерности и тенденции в текстах, что может быть использовано для различных целей, включая машинное обучение, интеллектуальный анализ данных и другие области применения.
Обработка естественного языка (Natural Language Processing, NLP) — это область искусственного интеллекта, которая занимается взаимодействием между компьютером и человеком с помощью естественного языка.

Основные задачи обработки естественного языка:
* Распознавание речи. Преобразование речи в текст.
* Синтез речи. Генерация речи из текста.
* Машинный перевод. Перевод текста с одного языка на другой.
* Анализ тональности. Определение эмоциональной окраски текста.
* Извлечение информации. Извлечение фактов и отношений из текста.
* Генерация текста. Создание нового текста на основе входных данных.
* Диалоговые системы. Взаимодействие с пользователем через диалог.

Для решения этих задач используются различные методы и подходы, такие как машинное обучение, глубокое обучение, статистические модели и другие.

Вот несколько примеров применения обработки естественного языка:

1. Голосовые помощники. Siri от Apple, Google Assistant и Яндекс Алиса используют обработку естественного языка для понимания запросов пользователей и предоставления им ответов или выполнения действий.
2. Машинный перевод. Google Translate, Яндекс Переводчик и другие сервисы используют обработку естественного языка для перевода текстов с одного языка на другой.
3. Анализ социальных медиа. Компании и организации используют обработку естественного языка для анализа отзывов клиентов, определения трендов и выявления проблем.
4. Автоматическое создание контента. Некоторые компании используют обработку естественного языка для создания новостных статей, обзоров продуктов и других видов контента.
5. Чат-боты. Чат-боты, которые могут отвечать на вопросы пользователей, также используют обработку естественного языка.
6. Рекомендательные системы. Рекомендательные системы, которые предлагают пользователям товары, услуги или контент, основанный на их предпочтениях, также применяют обработку естественного языка для анализа поведения пользователей.
7. Диагностика заболеваний. Обработка естественного языка может использоваться для анализа медицинских записей и диагностики заболеваний.
8. Поиск информации. Поисковые системы используют обработку естественного языка, чтобы понимать запросы пользователей и предоставлять им релевантные результаты.
9. Распознавание эмоций. Обработка естественного языка позволяет определять эмоции в тексте, что может быть полезно для анализа настроения пользователей или оценки качества обслуживания.
10. Создание диалогов. Обработка естественного языка используется для создания диалоговых систем, которые могут вести беседы с пользователями на различные темы.

47. Сравнительный анализ семантических сетей и фреймовых моделей.

Эталонный ответ
Семантические сети и фреймовые модели — это два подхода к представлению знаний в искусственном интеллекте. Они используются для обработки естественного языка, машинного обучения и других задач.

Семантическая сеть представляет собой граф, узлы которого соответствуют понятиям или объектам, а рёбра — отношениям между ними. Семантические сети позволяют описывать сложные взаимосвязи между объектами и понятиями, но они могут быть громоздкими и сложными для анализа.

Основные характеристики семантических сетей:
* Гибкость: семантические сети могут представлять различные типы отношений между понятиями. Это делает их гибким инструментом для моделирования сложных систем.
* Наглядность: семантические сети представляют знания в виде графа, что делает их наглядными и понятными для человека.
* Сложность: семантические сети могут стать сложными и запутанными при моделировании больших объёмов данных.

Примеры семантических сетей включают WordNet, Cyc и другие.

Фреймовая модель представляет знания в виде иерархической структуры, состоящей из фреймов. Фреймы содержат информацию о свойствах объектов и отношениях между ними.

Фреймовые модели позволяют эффективно представлять знания о конкретных ситуациях или объектах.

Основные характеристики фреймовых моделей:
* Иерархическая структура: фреймы организованы в иерархическую структуру, что позволяет эффективно представлять сложные системы.
* Специализация: фреймы могут специализироваться на определённых типах объектов или ситуаций, что упрощает их использование.
* Ограниченность: фреймовые модели могут быть ограничены в представлении абстрактных понятий или отношений.

Примерами фреймовых моделей являются FrameNet и PropBank.

Сравнительный анализ семантических сетей и фреймовых моделей показывает, что оба подхода имеют свои преимущества и недостатки. Семантические сети более гибкие и позволяют моделировать сложные отношения, но могут быть сложными для анализа. Фреймовые модели более структурированы и специализированы, но могут ограничивать представление абстрактных понятий. Выбор подхода зависит от конкретной задачи и требований к моделированию знаний.

В целом, семантические сети подходят для задач, требующих гибкости и способности моделировать сложные связи между объектами, в то время как фреймовые модели лучше подходят для представления конкретных ситуаций и объектов. Оба подхода широко используются в обработке естественного языка и машинном обучении.

48. Направления компьютерной лингвистики.

Эталонный ответ

Компьютерная лингвистика — это область искусственного интеллекта, которая занимается разработкой алгоритмов и программ для обработки естественного языка (Natural Language Processing, NLP).

Основные направления компьютерной лингвистики:
1. Морфологический анализ. Это процесс определения грамматических характеристик слова, таких как часть речи, род, число, падеж и т. д. Морфологический анализ используется для создания электронных словарей, поисковых систем, машинного перевода и других приложений NLP.
2. Синтаксический анализ. Это процесс построения синтаксического дерева предложения, которое показывает его структуру и связи между словами. Синтаксический анализ позволяет понять смысл предложения и выполнить такие задачи, как определение подлежащего и сказуемого, выделение словосочетаний и т. п.
3. Семантический анализ. Это процесс извлечения смысла из текста на основе его структуры и контекста. Семантический анализ позволяет определить значения слов, фраз и предложений, а также их отношения друг к другу. Семантический анализ используется в таких приложениях, как вопросно-ответные системы, автоматическое реферирование, генерация текстов и т. д.
4. Прагматический анализ. Это анализ текста с точки зрения его прагматики, то есть его воздействия на читателя или слушателя. Прагматический анализ учитывает такие факторы, как цель высказывания, контекст, пресуппозиции и импликатуры. Прагматический анализ важен для таких приложений, как диалоговые системы, аргументация, убеждение и т. п.
5. Статистический анализ. Это использование статистических методов для анализа текстовых данных. Статистический анализ может применяться для решения различных задач NLP, таких как классификация, кластеризация, регрессия и т. д. Статистические

методы позволяют выявить закономерности в тексте и использовать их для предсказания или объяснения различных явлений.
6. Машинный перевод. Это автоматический перевод текста с одного языка на другой. Машинный перевод основан на использовании алгоритмов и моделей, которые учитывают грамматические, семантические и прагматические аспекты языка. Машинный перевод является одним из самых популярных и сложных приложений NLP.
7. Распознавание речи. Это автоматическое распознавание и понимание устной речи. Распознавание речи основано на использовании акустических моделей, языковых моделей и алгоритмов декодирования. Распознавание речи применяется в таких областях, как голосовые помощники, автоматизация колл-центров, переводчики и т. п.
8. Генерация речи. Это создание синтезированной речи на основе текстовой информации. Генерация речи основана на использовании фонетических, просодических и лексико-грамматических правил. Генерация речи используется в таких приложениях, как озвучивание текста, чтение новостей, обучение иностранным языкам и т. д.
9. Обработка естественного языка. Это общий термин, который охватывает все вышеперечисленные направления и другие области NLP. Обработка естественного языка включает в себя разработку алгоритмов, моделей и методов, которые позволяют компьютерам понимать и обрабатывать естественный язык так же, как это делают люди.

Это лишь некоторые из основных направлений компьютерной лингвистики. В настоящее время существует множество других направлений, которые продолжают развиваться и совершенствоваться.

49. Принципы работы автоматических систем извлечения информации.

Эталонный ответ
Автоматические системы извлечения информации (Automatic Information Extraction, AIE) — это технологии и методы, которые позволяют компьютерам автоматически извлекать структурированную информацию из неструктурированных или полуструктурированных данных.

Основные принципы работы таких систем:
1. Обработка естественного языка (Natural Language Processing, NLP) — это набор методов, алгоритмов и инструментов для анализа и обработки текстовых данных на естественном языке. NLP включает в себя такие задачи, как токенизация, лемматизация, стемминг, синтаксический анализ, семантический анализ и другие. Эти методы используются для понимания смысла текста и извлечения из него полезной информации.
2. Извлечение сущностей (Entity Extraction) — это процесс идентификации и классификации объектов, событий, фактов и других сущностей в тексте. Извлечение сущностей может быть основано на правилах, статистических моделях или машинном обучении.
3. Структурирование данных (Data Structuring) — после извлечения сущностей необходимо структурировать данные таким образом, чтобы они были пригодны для дальнейшей обработки и анализа. Структурированные данные могут быть представлены в виде таблиц, графов или других форматов.
4. Интеграция с базами данных и системами управления данными (Database Integration) — структурированные данные, извлечённые из текстов, могут быть интегрированы с существующими базами данных или системами управления данными для дальнейшего анализа и использования.
5. Машинное обучение и глубокое обучение (Machine Learning and Deep Learning) — для повышения точности и эффективности извлечения информации используются методы

машинного обучения и глубокого обучения. Они позволяют системам автоматически обучаться на больших объёмах данных и улучшать свои результаты.
6. Анализ тональности (Sentiment Analysis) — некоторые системы извлечения информации также включают анализ тональности, который позволяет определить эмоциональное содержание текста. Это может быть полезно для выявления мнений, отзывов и оценок.
7. Семантический анализ (Semantic Analysis) — семантический анализ используется для понимания значения слов и фраз в контексте. Он помогает системам лучше понимать смысл текста и извлекать более точную информацию.
8. Онтологический подход (Ontology-Based Approach) — онтологии представляют собой формальное описание понятий и отношений между ними. Использование онтологий позволяет системам более точно интерпретировать текст и извлекать информацию, соответствующую определённым концепциям.
9. Использование метаданных (Metadata Utilization) — метаданные предоставляют дополнительную информацию о тексте, такую как автор, дата создания, тема и т. д. Эта информация может быть использована для улучшения качества извлечения информации.
10. Адаптивность и обучение (Adaptability and Learning) — автоматические системы извлечения информации должны быть способны адаптироваться к новым данным и контекстам. Они также должны иметь возможность учиться на своих ошибках и улучшать свою производительность со временем.

Эти принципы обеспечивают основу для разработки и реализации автоматических систем извлечения информации, которые могут эффективно обрабатывать большие объёмы текстовых данных и предоставлять структурированную информацию для различных целей.

50. Определения центральностей разного типа.

Эталонный ответ
Центральность — это мера важности узла в графе. Она используется для определения наиболее значимых узлов в сети, которые могут играть ключевую роль в передаче информации или других процессах.

Существует несколько типов центральности:
1. Степень центральности (Degree Centrality) — это количество рёбер, инцидентных узлу. Узел с большим количеством связей считается более важным. Степень центральности рассчитывается по формуле: $C_D(v) = deg(v)$, где v — узел, а $deg(v)$ — степень узла v.

2. Близость центральность (Closeness Centrality) оценивает, насколько быстро информация может распространяться от узла к другим узлам в сети. Близость центральность рассчитывается по формуле: $$C_C(v)=\frac{n-1}{\sum_{u \in V}d(u, v)}$, где V — множество всех узлов, n — количество узлов, а $d(u,v)$ — кратчайшее расстояние между узлами u и v. Чем меньше значение близости центральности, тем важнее узел.

3. Междуness центральность (Betweenness Centrality) измеряет, сколько кратчайших путей проходит через узел. Междуness центральность вычисляется по формуле: $$C_B(v)= \sum_{s \ne v \ne t \in V}\frac{\sigma_{st}(v)}{\sigma_{st}}$, где σ_{st} — общее количество кратчайших путей между узлами s и t, а $\sigma_{st}(v)$ — количество этих кратчайших путей, проходящих через узел v. Узлы с высокой междуness центральностью играют важную роль в передаче информации между другими узлами.

4. Собственность центральность (Eigenvector Centrality) основана на идее, что важность узла определяется важностью его соседей. Собственность центральность учитывает не

только количество связей узла, но и качество этих связей. Собственность центральность можно рассчитать с помощью собственных векторов матрицы смежности графа.

5. Хаб центральность (Hub Centrality) похожа на степень центральности, но учитывает только исходящие связи узла. Хаб центральность определяется как количество исходящих рёбер из узла.

Выбор типа центральности зависит от конкретной задачи и целей анализа. Центральность помогает понять структуру сети и определить наиболее важные узлы для дальнейшего изучения.

51. Перечислить основные направления исследований социальных сетей.

Эталонный ответ
Основные направления исследований социальных сетей:

1. Анализ структуры и динамики социальных связей. В рамках этого направления изучаются различные аспекты взаимодействия пользователей в социальных сетях: формирование и эволюция социальных групп, анализ влияния одних пользователей на других, изучение распространения информации и т. д.

2. Разработка методов и алгоритмов анализа данных из социальных сетей. Это направление включает в себя создание и совершенствование методов обработки текстовых данных (например, сентимент-анализ), разработку алгоритмов для выявления сообществ и кластеризации пользователей, а также создание инструментов для визуализации данных.

3. Исследование влияния социальных сетей на общество и культуру. Здесь рассматриваются вопросы о том, как социальные сети влияют на формирование общественного мнения, распространение информации, развитие культурных трендов и т. п.

4. Изучение вопросов безопасности и конфиденциальности в социальных сетях. Это направление связано с анализом проблем, связанных с защитой персональных данных пользователей, противодействием кибербуллингу, выявлением и предотвращением распространения вредоносного контента и т.д.

5. Применение технологий социальных сетей в различных областях. Социальные сети могут быть использованы для решения разнообразных задач, таких как маркетинг, управление проектами, обучение и другие. Исследователи изучают, как можно эффективно использовать социальные сети для достижения конкретных целей.

6. Прогнозирование поведения пользователей в социальных сетях. На основе анализа данных о поведении пользователей в прошлом можно попытаться предсказать их будущие действия, например, какие посты они будут лайкать или комментировать. Это может быть полезно для персонализации рекламы, рекомендаций контента и других целей.

7. Создание новых моделей и методов для анализа социальных сетей. Исследователи постоянно разрабатывают новые подходы и алгоритмы для более глубокого понимания социальных взаимодействий и процессов, происходящих в социальных сетях.

8. Оценка качества и достоверности информации в социальных сетях. С развитием социальных сетей становится всё сложнее отличить достоверную информацию от фейков и

слухов. Исследователи работают над созданием методов и инструментов для оценки качества информации в социальных медиа.

Это лишь некоторые из основных направлений исследований в области социальных сетей. С развитием технологий и изменением потребностей общества появляются новые задачи и вызовы, которые требуют изучения и разработки новых подходов и решений.

52. Разъяснить принципы работы графематического и морфологического анализаторов.

Эталонный ответ
Графематический и морфологический анализаторы — это два важных компонента систем обработки естественного языка (NLP), которые используются для предварительной обработки текста. Они помогают преобразовать текст в более структурированный формат, который может быть использован для дальнейшего анализа и обработки.

Графематический анализатор занимается выделением основных структурных элементов текста, таких как слова, предложения, абзацы и т.д. Он также выполняет задачи по нормализации текста, такие как удаление лишних пробелов, исправление опечаток и т.п. Это позволяет упростить дальнейшую обработку текста и сделать её более эффективной.

Основные функции графематического анализатора:
* Разделение текста на токены. Токены — это отдельные элементы текста, например, слова или символы пунктуации. Графематический анализатор разбивает текст на эти токены, что упрощает его дальнейшую обработку.
* Идентификация границ слов. Анализатор определяет, где заканчивается одно слово и начинается другое. Это важно для правильного понимания смысла текста.
* Определение частей речи. Анализатор может определить, к какой части речи относится каждое слово в тексте. Это помогает в дальнейшем анализе текста и его интерпретации.

Для выполнения этих функций графематический анализатор использует различные методы и алгоритмы, включая регулярные выражения, лексический анализ и другие.

После того как текст был обработан графематическим анализатором, он передаётся на следующий этап обработки — морфологический анализ. Морфологический анализатор выполняет более глубокий анализ текста, определяя морфологические характеристики каждого слова.

Функции морфологического анализатора включают:
* Определение грамматических характеристик слов, таких как род, число, падеж и др.
* Выявление основ слов и окончаний.
* Идентификация форм глаголов.

Морфологический анализатор также использует различные алгоритмы и методы для выполнения своих функций. Например, он может использовать таблицы склонения и спряжения для определения грамматических характеристик слов.

В целом, графематический и морфологический анализаторы являются важными компонентами систем обработки естественного языка. Они позволяют преобразовать текст в структурированный формат, что облегчает его дальнейший анализ и обработку.

53. Дать формальные определения частей речи, грамматического рода и категории падежа в терминах модели языка, предложенной С. Маркусом.

Эталонный ответ
Части речи — это классы слов, которые обладают общими грамматическими и семантическими характеристиками. В модели языка, предложенной С. Маркусом, части речи определяются на основе их синтаксических функций и морфологических признаков.

С. Маркус выделяет следующие части речи:
* Существительное — часть речи, которая обозначает предмет или понятие и может выступать в роли подлежащего, дополнения или именной части сказуемого. Существительные имеют род, число и падеж.
* Прилагательное — часть речи, которая описывает признаки существительного и согласуется с ним по роду, числу и падежу. Прилагательные могут выступать в роли определения или именной части составного сказуемого.
* Глагол — часть речи, обозначающая действие или состояние и изменяющаяся по временам, лицам и числам. Глаголы могут быть переходными и непереходными, а также иметь различные виды и залоги.
* Наречие — неизменяемая часть речи, которая указывает на признак действия, качества или состояния. Наречия могут обозначать время, место, образ действия и т. д.
* Местоимение — часть речи, указывающая на предметы, признаки или количества, но не называющая их. Местоимения могут заменять существительные, прилагательные или числительные.
* Числительное — часть речи, выражающая количество или порядок предметов при счёте. Числительные делятся на количественные (обозначают количество) и порядковые (обозначают порядок).
* Предлог — служебная часть речи, связывающая слова в словосочетании и предложении. Предлоги выражают различные отношения между предметами и явлениями.
* Союз — служебная часть речи, соединяющая однородные члены предложения или простые предложения в составе сложного. Союзы бывают сочинительными (соединяют равноправные элементы) и подчинительными (выражают зависимость одного элемента от другого).
* Междометие — особая часть речи, выражающее чувства и эмоции. Междометия не изменяются и не являются членами предложения.

Грамматический род — это категория, которая определяет принадлежность существительного к одному из трёх родов: мужскому, женскому или среднему. Род существительного может влиять на согласование с прилагательными, местоимениями и глаголами.

В модели С. Маркуса род определяется на основе морфологических показателей (окончаний) и семантических характеристик (пола, одушевлённости и т. п.). Например, существительные мужского рода обычно имеют нулевое окончание в именительном падеже единственного числа (стол, дом), а существительные женского рода — окончание -а (-я) (страна, земля).

Падеж — это грамматическая категория, которая выражает отношение существительного или местоимения к другим словам в предложении. Падеж определяется на основе синтаксической функции слова (подлежащее, дополнение, определение и т. д.) и его окончания.

В русском языке существует шесть падежей: именительный, родительный, дательный, винительный, творительный и предложный. Каждый падеж имеет свои вопросы и окончания, которые помогают определить его значение. Например, именительный падеж отвечает на вопрос «кто?» или «что?» и имеет нулевое окончание (дом, стол), а родительный — на вопрос «кого?» или «чего?» и окончание -а (-я) или -у (-ю) (дома, стола).

54. Перечислить основные этапы построения автоматизированных систем обработки текстов.

Эталонный ответ
Основные этапы построения автоматизированных систем обработки текстов:
1. Сбор и подготовка данных. На этом этапе происходит сбор текстовых данных, которые будут использоваться для обучения модели. Это могут быть новостные статьи, блоги, социальные сети, научные публикации и т.д. Также необходимо провести предварительную обработку данных, чтобы удалить ненужные символы, нормализовать текст и привести его к единому формату.
2. Выбор модели. Существует множество моделей обработки естественного языка (NLP), таких как языковые модели, модели машинного перевода, модели классификации и др. Выбор модели зависит от конкретной задачи, которую необходимо решить. Например, для определения тональности текста можно использовать модель классификации, а для генерации текста – языковую модель.
3. Обучение модели. После выбора модели необходимо обучить её на собранных данных. Для этого используются методы машинного обучения, такие как обучение с учителем или без учителя. В процессе обучения модель учится предсказывать значения выходных переменных на основе входных данных.
4. Тестирование модели. После обучения модели необходимо проверить её качество на тестовых данных, которые не использовались при обучении. Если результаты тестирования удовлетворительны, то модель готова к использованию.
5. Развертывание модели. Если модель успешно прошла тестирование, её можно развернуть в производственной среде. Для этого необходимо настроить инфраструктуру, которая будет обеспечивать работу модели в режиме реального времени.
6. Мониторинг и оптимизация. После развертывания модели необходимо отслеживать её производительность и качество работы. При необходимости можно проводить оптимизацию модели, чтобы улучшить её результаты.
7. Интеграция с другими системами. Автоматизированные системы обработки текстов могут быть интегрированы с другими информационными системами, такими как системы управления контентом, поисковые системы и т.п. Это позволяет расширить функциональность и повысить эффективность работы системы.

Это лишь основные этапы построения автоматизированных систем обработки текстов. В зависимости от конкретной задачи и используемых технологий, процесс может включать дополнительные шаги.

55. Разъяснить принципы работы фрагментационного и синтаксического анализаторов.

Эталонный ответ
Обработка естественного языка (Natural Language Processing, NLP) — это область искусственного интеллекта, которая занимается анализом и пониманием человеческого языка. В процессе обработки естественного языка используются различные инструменты и методы, которые позволяют компьютерам понимать и интерпретировать текст.

Фрагментационный анализатор — это инструмент для обработки естественного языка, который разбивает текст на отдельные фрагменты или токены. Токенами могут быть слова, фразы, символы и другие элементы текста. Фрагментационный анализ является первым этапом обработки текста и позволяет упростить дальнейшую обработку данных.

Основные принципы работы фрагментационного анализатора:
1. Разделение текста на токены. Анализатор разделяет текст на отдельные слова, знаки препинания, числа и другие символы. Это позволяет упростить дальнейший анализ текста.
2. Удаление стоп-слов. Стоп-слова — это слова, которые не несут смысловой нагрузки и часто встречаются в тексте. Например, «и», «или», «а». Анализатор удаляет эти слова из текста, чтобы уменьшить размер данных и ускорить обработку.
3. Стемминг и лемматизация. Стемминг — это процесс приведения слова к его основе. Лемматизация — это более сложный процесс, который учитывает морфологию слова и приводит его к словарной форме. Анализатор может использовать эти методы для упрощения обработки текста.
4. Нормализация. Анализатор может выполнять нормализацию текста, то есть приводить все слова к одному регистру (например, к нижнему) и удалять диакритические знаки. Это упрощает сравнение слов и поиск по тексту.
5. Фильтрация. Анализатор может фильтровать токены, удаляя из них нерелевантные символы или заменяя их на более простые. Это также упрощает обработку текста.
6. Создание словаря. Анализатор создаёт словарь токенов, с которыми будет работать. Словарь может использоваться для поиска и сравнения слов в тексте.
7. Выделение признаков. Анализатор выделяет признаки из токенов, такие как часть речи, лемма, частота и т. д. Эти признаки будут использоваться на следующих этапах обработки.
8. Вывод результатов. Анализатор выводит результаты своей работы в виде списка токенов или других структур данных, которые будут использоваться другими инструментами обработки текста.

После фрагментационного анализа текст готов к дальнейшей обработке, такой как синтаксический анализ.

Синтаксический анализатор — это инструмент, который анализирует структуру предложения и определяет его синтаксические связи. Синтаксический анализ позволяет понять смысл предложения и определить его грамматическую правильность.

Основные принципы работы синтаксического анализатора:

1. Построение дерева зависимостей. Синтаксический анализатор строит дерево зависимостей, в котором каждое слово связано с другими словами в предложении. Дерево зависимостей отражает синтаксическую структуру предложения.
2. Определение частей речи. Анализатор определяет части речи каждого слова в предложении. Это помогает понять роль слова в предложении и его связь с другими словами.
3. Разрешение неоднозначности. Анализатор разрешает неоднозначность в структуре предложения, выбирая наиболее вероятный вариант. Это важно для понимания смысла предложения.
4. Проверка грамматической правильности. Анализатор проверяет грамматическую правильность предложения, выявляя ошибки и несоответствия.
5. Вывод результатов. Синтаксический анализатор выводит результаты своего анализа в виде дерева зависимостей или другой структуры данных, которая может быть использована для дальнейшего анализа текста.

Синтаксический анализ является важным этапом обработки естественного языка и позволяет компьютерам лучше понимать смысл текста.

56. Изложить основную идею теоретико-модельного подхода к исследованию семантики текстов.

Эталонный ответ
Теоретико-модельный подход — это один из методов исследования семантики текстов, который основан на использовании математических моделей и теорий для анализа содержания и смысла текста.

Основная идея теоретико-модельного подхода заключается в том, чтобы представить текст как формальную систему, которая может быть описана с помощью определённых правил и структур. Это позволяет проводить анализ текста на основе его формальных свойств и характеристик, а также выявлять закономерности и особенности его семантической структуры.

В рамках теоретико-модельного подхода используются различные методы и модели, такие как:
* Формальные грамматики, которые позволяют описывать синтаксические структуры языка и их семантические интерпретации.
* Семантические сети, представляющие собой графическое представление отношений между понятиями и объектами в тексте.
* Логические модели, основанные на формальной логике и позволяющие делать выводы о содержании текста на основе логических правил.
* Статистические модели, использующие методы машинного обучения для выявления закономерностей в данных и предсказания значений слов и фраз.

Применение теоретико-модельного подхода позволяет получить более глубокое понимание семантических аспектов текста и выявить его скрытые значения и смыслы. Этот подход широко используется в различных областях, таких как компьютерная лингвистика, искусственный интеллект и обработка естественного языка.

Однако стоит отметить, что теоретико-модельный подход имеет свои ограничения. Он не всегда способен полностью отразить все нюансы и тонкости семантики текста, особенно если речь идёт о сложных и неоднозначных контекстах. Кроме того, этот подход требует значительных усилий и ресурсов для разработки и реализации соответствующих моделей и алгоритмов.

57. Проблемы автоматизации синтаксического анализа.

Эталонный ответ
Проблемы автоматизации синтаксического анализа

Синтаксический анализ — это процесс определения структуры предложения или текста, включая отношения между словами и фразами. Автоматизация этого процесса является сложной задачей в области обработки естественного языка (NLP). Вот некоторые из проблем, с которыми сталкиваются исследователи при разработке систем автоматического синтаксического анализа:

1. Разнообразие языков: синтаксические правила разных языков могут сильно различаться, что затрудняет создание универсальных моделей.

2. Сложность синтаксических правил: даже в одном языке синтаксис может быть сложным и неоднозначным, что усложняет задачу автоматического анализа.

3. Зависимость от контекста: значение слов и фраз может меняться в зависимости от контекста, что требует учёта семантических и прагматических аспектов при анализе.

4. Омонимия и полисемия: одно и то же слово или фраза могут иметь разные значения в разных контекстах, что также усложняет процесс анализа.

5. Обработка сложных конструкций: такие конструкции, как придаточные предложения, могут быть сложными для анализа, особенно если они содержат несколько уровней вложенности.

6. Разрешение синтаксической неоднозначности: в некоторых случаях синтаксическая структура может быть неоднозначной, и необходимо использовать дополнительные методы для выбора наиболее вероятной интерпретации.

7. Обучение на больших объёмах данных: для создания точных моделей синтаксического анализа требуется большое количество размеченных данных, что может быть сложно получить для некоторых языков или областей.

8. Интеграция с другими аспектами NLP: синтаксический анализ тесно связан с другими задачами NLP, такими как семантический анализ и машинный перевод, что требует разработки комплексных моделей, учитывающих все эти аспекты.

Для решения этих проблем исследователи используют различные подходы, такие как статистические модели, основанные на машинном обучении, грамматические правила, а также гибридные методы, сочетающие оба подхода. Однако, несмотря на прогресс в этой области, автоматизация синтаксического анализа всё ещё остаётся сложной и актуальной задачей в NLP.

58. Сформулировать принципы обработки неструктурированной и плохо структурированной информации. Индексация текстов.

Эталонный ответ
Принципы обработки неструктурированной и плохо структурированной информации

Неструктурированная информация — это данные, которые не имеют заранее определённой структуры или формата. К ней относятся тексты, изображения, аудио- и видеофайлы. Плохо структурированная информация имеет некоторую структуру, но она недостаточно чёткая или сложная для автоматической обработки.

Обработка неструктурированных данных включает в себя несколько этапов:
* Сбор данных. На этом этапе происходит получение данных из различных источников, таких как веб-сайты, социальные сети, новостные порталы и т. д.
* Предварительная обработка. Этот этап включает в себя очистку данных от шума, удаление стоп-слов, приведение к единому формату и другие операции, направленные на улучшение качества данных.
* Анализ данных. Здесь применяются различные методы машинного обучения и обработки естественного языка (NLP) для извлечения полезной информации из данных. Это может

быть поиск ключевых слов, определение тем, выявление отношений между объектами и т. п.
* Интерпретация результатов. Полученные результаты анализируются и интерпретируются с целью получения выводов и принятия решений.

Для обработки неструктурированных и плохо структурированных данных используются следующие принципы:
1. Использование методов машинного обучения. Машинное обучение позволяет автоматически извлекать знания из данных без необходимости явного программирования правил. Для этого используются такие методы, как классификация, кластеризация, регрессия и др.
2. Применение методов NLP. Обработка естественного языка позволяет анализировать и понимать текст на естественном языке. Для этого применяются методы токенизации, стемминга, лемматизации, синтаксического анализа и др.
3. Создание моделей. Модели машинного обучения позволяют автоматически классифицировать, кластеризовать и предсказывать значения на основе данных. Они могут быть обучены на больших объёмах данных и адаптированы под конкретные задачи.
4. Интеграция с другими технологиями. Обработка неструктурированных данных часто интегрируется с другими технологиями, такими как базы данных, системы управления проектами, CRM-системы и т. д., чтобы обеспечить более полную и точную картину ситуации.
5. Постоянное обновление и улучшение. Методы обработки неструктурированных данных постоянно развиваются и совершенствуются. Новые алгоритмы и подходы позволяют улучшить качество и точность анализа данных.
6. Визуализация результатов. Результаты обработки данных могут быть представлены в виде графиков, диаграмм, карт и других визуальных элементов, что облегчает их понимание и интерпретацию.
7. Обеспечение безопасности данных. При обработке неструктурированных данных необходимо уделять внимание вопросам безопасности, чтобы предотвратить несанкционированный доступ к конфиденциальной информации.
8. Гибкость и масштабируемость. Системы обработки неструктурированных данных должны быть гибкими и масштабируемыми, чтобы адаптироваться к изменяющимся требованиям и объёмам данных.
9. Автоматизация процессов. Автоматизация процессов обработки данных позволяет ускорить и упростить работу, а также снизить вероятность ошибок.
10. Оценка качества. Оценка качества результатов обработки данных помогает определить эффективность используемых методов и алгоритмов.

Индексация текстов

Индексация — это процесс создания индекса, который представляет собой список терминов и их местоположений в тексте. Индексы используются для быстрого поиска информации в больших объёмах текстовых данных.

Основные этапы индексации текстов:
* Токенизация. Текст разбивается на отдельные слова или фразы (токены).
* Стемминг или лемматизация. Токены приводятся к их основной форме (основе слова или лемме).
* Удаление стоп-слов. Отбрасываются слова, не несущие смысловой нагрузки (например, «и», «а», «но» и т.п.).
* Построение инвертированного индекса. Создаётся таблица, в которой каждому слову или фразе соответствует список документов, где они встречаются.

После индексации тексты можно эффективно искать по ключевым словам, фразам или темам.

59. Изложить основные идеи подхода И. Мельчука к семантическому анализу.

Эталонный ответ

Игорь Мельчук — это один из основателей Московской семантической школы и автор теории «Смысл ⇔ Текст». В рамках этой теории он разработал модель языка, которая описывает процесс перехода от смысла к тексту и обратно.

Основные идеи подхода И. Мельчука к семантическому анализу:
1. Разделение лингвистических уровней. Мельчук предложил разделить язык на несколько уровней: семантический, синтаксический, морфологический и фонологический. Это позволило более детально изучить каждый уровень и выявить закономерности их взаимодействия.
2. Семантический анализ. Семантический уровень является центральным в теории И. Мельчука. Он включает в себя описание значений слов, фраз и предложений. Для этого используются семантические примитивы — элементарные значения, которые не могут быть разложены на более простые составляющие.
3. Синтаксис и семантика. Синтаксический уровень описывает структуру предложения, а семантический — его смысл. И. Мельчук считал, что эти два уровня тесно связаны и должны рассматриваться вместе.
4. Лексические функции. Лексические функции связывают слова с их значениями. Они позволяют описать отношения между словами и определить их роль в предложении. Например, лексическая функция Magn оценивает степень проявления признака, выраженного прилагательным.
5. Формализация семантики. И. Мельчук стремился формализовать семантику языка, чтобы сделать её более точной и предсказуемой. Для этого он использовал математические методы и логические формулы.
6. Универсальность подхода. Подход И. Мельчука может быть применён к различным языкам и позволяет проводить сравнительный анализ их семантических структур.
7. Сложность реализации. Однако подход И. Мельчука является сложным для реализации и требует глубоких знаний в области лингвистики и математики. Кроме того, некоторые аспекты теории остаются спорными и требуют дальнейшего исследования.

Подход И. Мельчука оказал значительное влияние на развитие семантического анализа и продолжает оставаться актуальным в современных исследованиях.

60. Формальные методы атрибуции текстов.

Эталонный ответ
Формальные методы атрибуции текстов — это методы, основанные на анализе статистических характеристик текста, таких как частотность слов, длина предложений, структура абзацев и т. д. Эти методы используются для определения авторства или источника текста.

Формальные методы включают в себя различные подходы и алгоритмы, которые позволяют анализировать текст и извлекать из него информацию о его авторе или источнике. Некоторые из этих методов:

* Частотный анализ. Этот метод основан на подсчёте частоты использования определённых слов или фраз в тексте. Частотные характеристики могут быть использованы для сравнения текстов и определения их сходства.
* Стемминг и лемматизация. Эти методы позволяют привести слова к их основной форме (например, «бегу», «бежит» и «бежать» будут преобразованы в «бежать»). Это позволяет более точно сравнивать тексты и определять их сходство.
* Синтаксический анализ. Метод основан на изучении структуры предложений и связей между словами. Синтаксические особенности могут использоваться для определения стиля автора или источника.
* Семантический анализ. Метод изучает смысл текста и его семантические связи. Семантические особенности могут помочь определить тематику текста и его контекст.
* Статистические методы. Методы используют статистические модели для анализа текста и определения его характеристик. Статистические модели могут включать в себя машинное обучение и нейронные сети.

Эти методы могут применяться как по отдельности, так и в комбинации друг с другом. Они широко используются в различных областях, включая криминалистику, литературную критику и научные исследования. Однако формальные методы имеют свои ограничения, и результаты могут быть не всегда точными. Поэтому важно учитывать контекст и цель исследования при использовании этих методов.

Для повышения точности атрибуции текста можно использовать комбинацию нескольких методов. Например, можно объединить частотный анализ с синтаксическим анализом или семантическим анализом. Также можно использовать машинное обучение для создания моделей, которые учитывают различные аспекты текста при определении авторства.

В целом, формальные методы атрибуции текстов являются мощным инструментом для анализа и понимания текстов. Они позволяют исследователям получить представление о стиле автора, теме текста и других характеристиках, которые могут быть полезны для различных целей.

61. Сформулировать принципы построения систем представления знаний для компьютерной обработки.

Эталонный ответ
Принципы построения систем представления знаний для компьютерной обработки

Системы представления знаний (СПЗ) — это компьютерные системы, которые позволяют хранить и обрабатывать знания в определённой предметной области. Они используются для решения задач искусственного интеллекта, таких как понимание естественного языка, машинное обучение, экспертные системы и т. д.

Основные принципы построения СПЗ:
1. Формализация знаний. Знания должны быть представлены в формальном виде, который может быть обработан компьютером. Для этого используются различные модели представления знаний, такие как логические, продукционные, фреймовые, семантические сети и др.
2. Непротиворечивость. Знания не должны содержать противоречий, которые могут привести к неправильным выводам. Для обеспечения непротиворечивости используются методы логического вывода, такие как дедукция, индукция и абдукция.

3. Полнота. Система должна содержать все необходимые знания для решения поставленной задачи. Полнота зависит от целей и задач системы, а также от доступных ресурсов.
4. Актуальность. Знания должны соответствовать текущему состоянию предметной области и не устаревать со временем. Актуальность обеспечивается путём обновления и пополнения системы новыми знаниями.
5. Доступность. Система должна быть доступна для пользователей, которые хотят получить доступ к её знаниям. Доступность обеспечивается с помощью интерфейсов пользователя, которые позволяют задавать вопросы и получать ответы на естественном языке.
6. Эффективность. Система должна работать быстро и эффективно, чтобы пользователи могли получать результаты в реальном времени. Эффективность достигается за счёт оптимизации алгоритмов обработки знаний и использования параллельных вычислений.
7. Гибкость. Система должна легко адаптироваться к изменениям в предметной области и требованиям пользователей. Гибкость обеспечивается за счёт модульной структуры системы и возможности добавления новых модулей.
8. Интеграция. Система должна интегрироваться с другими системами и технологиями, такими как базы данных, интернет, мобильные устройства и т. п. Интеграция позволяет расширить функциональность системы и повысить её эффективность.
9. Безопасность. Система должна обеспечивать защиту своих знаний от несанкционированного доступа и изменения. Безопасность достигается с помощью шифрования данных, аутентификации пользователей и других методов защиты информации.
10. Открытость. Система должна предоставлять возможность другим разработчикам создавать свои собственные модули и приложения на основе её знаний. Открытость способствует развитию сообщества разработчиков и повышению качества системы.

Эти принципы являются общими для всех СПЗ, но их реализация может различаться в зависимости от конкретной модели и задачи. Например, для логических моделей основное внимание уделяется формализации знаний и обеспечению непротиворечивости, а для фреймовых моделей — структурированию знаний и организации доступа к ним.

5. Средства оценки индикаторов достижения компетенций

Таблица 4

Средства оценки индикаторов достижения компетенций

Коды компетенций	**Индикаторы компетенций** *(в соотв. с Таблицей 1)*	**Средства оценки** *(в соотв. с Таблицами 5, 7)*
УК-1	ИД.УК-1.1. ИД.УК-1.2. ИД.УК-1.3. ИД.УК-1.4.	Опрос, диспут, практическое задание, контрольное задание, письменная работа (эссе)
ПК-3	ИД.ПК-3.1. ИД.ПК-3.2. ИД.ПК-3.3. ИД.ПК-3.4. ИД.ПК-3.5. ИД.ПК-3.6.	Опрос, диспут, практическое задание, контрольное задание, письменная работа (эссе)

Таблица 5

Описание средств оценки индикаторов достижения компетенций

Средства оценки *(в соотв. с Таблицами 5, 7)*	**Рекомендованный план выполнения работы**
Опрос	Магистрант в ходе подготовки и участия в опросе показывает наличие практической базы знаний в рамках дисциплины, необходимой для выполнения следующих действий в области профессиональной деятельности: – анализирует проблемную ситуацию, определяет пробелы в информации, оценивать надёжность источников информации, разрабатывает стратегию решения проблемной ситуации на основе системного и междисциплинарного подходов, строит сценарии реализации стратегии, определяя возможные риски и предлагая пути их устранения; – применяет современные методы, поиска, обработки, анализа и использования информации в рамках проведения научно-исследовательских и организационных работ в области музейных исследований и кураторских стратегий.
Диспут	Магистрант в ходе подготовки и участия в диспуте показывает наличие практической базы знаний в рамках дисциплины, необходимой для выполнения следующих действий в области профессиональной деятельности: – анализирует проблемную ситуацию, определяет пробелы в информации, оценивать надёжность источников информации, разрабатывает стратегию решения проблемной ситуации на основе системного и междисциплинарного подходов, строит сценарии реализации стратегии, определяя возможные риски и предлагая пути их устранения; – применяет современные методы, поиска, обработки, анализа и использования информации в рамках проведения научно-исследовательских и организационных работ в области музейных исследований и кураторских стратегий.
Практическое задание	Магистрант в ходе подготовки и выполнения практического задания показывает наличие практической базы знаний в рамках дисциплины, необходимой для выполнения следующих действий в области профессиональной деятельности: – анализирует проблемную ситуацию, определяет пробелы в информации, оценивать надёжность источников информации, разрабатывает стратегию решения проблемной ситуации на основе системного и междисциплинарного подходов, строит сценарии реализации стратегии, определяя возможные риски и предлагая пути их устранения; – применяет современные методы, поиска, обработки, анализа и использования информации в рамках проведения научно-исследовательских и организационных работ в области музейных исследований и кураторских стратегий.
Контрольное задание	Магистрант в ходе подготовки и выполнения контрольного задания показывает наличие практической базы знаний в рамках дисциплины, необходимой для выполнения следующих действий в области профессиональной деятельности: – анализирует проблемную ситуацию, определяет пробелы в информации, оценивать надёжность источников информации, разрабатывает стратегию решения проблемной ситуации на основе системного и междисциплинарного подходов, строит сценарии реализации стратегии, определяя возможные риски и предлагая пути их устранения; – применяет современные методы, поиска, обработки, анализа и использования информации в рамках проведения научно-исследовательских и организационных работ в области музейных исследований и кураторских стратегий.
Письменная работа (эссе)	Магистрант в ходе подготовки и написания письменной работы (эссе), показывает наличие практической базы знаний в рамках дисциплины, необходимой для выполнения следующих действий в области профессиональной деятельности: – анализирует проблемную ситуацию, определяет пробелы в информации, оценивать надёжность источников информации, разрабатывает стратегию решения проблемной ситуации на основе системного и междисциплинарного подходов, строит сценарии реализации стратегии, определяя возможные риски и предлагая пути их устранения; – применяет современные методы, поиска, обработки, анализа и использования информации в рамках проведения научно-исследовательских и организационных работ в области музейных исследований и кураторских стратегий.