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

Факультет экономики

УТВЕРЖДАЮ:
Ректор _____ В.В. Волков
«29» марта 2023 г.
Протокол Ученого Совета
№ 2 от 29 марта 2023 г.

Рабочая программа дисциплины
Текстовые данные

образовательная программа
направление подготовки
38.04.01 Экономика

направленность (профиль)
«Экономика и финансы»
программа подготовки – магистратура

язык обучения – русский
форма обучения – очная

квалификация (степень) выпускника
Магистр

Санкт-Петербург

Автор:
Браславский П.И., к.т.н., доцент факультета социологии АНООВО «ЕУСПб»

Рецензент:
Тушканова О.Н., к.т.н., доцент факультета социологии АНООВО «ЕУСПб»

Рабочая программа дисциплины **«Текстовые данные»**, входящей в образовательную программу уровня магистратуры «Экономика и финансы», утверждена на заседании Совета факультета экономики.

Протокол заседания № 9 от 27 февраля 2023 года.

АННОТАЦИЯ РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ
«Текстовые данные»

Дисциплина **«Текстовые данные»** является дисциплиной по выбору части, формируемой участниками образовательный отношений, Блока 1. «Дисциплины (модули)» основной профессиональной образовательной программы высшего образования «Финансовая экономика» по направлению подготовки 38.04.01 Экономика.

Дисциплина «Текстовые данные» дает магистрантам представление о теоретических подходах к количественному анализу текстов в общественных науках. Дисциплина также знакомит магистрантов с ключевыми источниками текстовых данных в общественных науках, дает введение в корпусные исследования и проблемы вычислительной лингвистики, магистранты развивают навыки по созданию массивов структурированных текстов из неструктурированных данных.

Программой дисциплины предусмотрены следующие виды контроля: текущий контроль успеваемости, промежуточный контроль в форме зачета с оценкой (в конце 8 модуля).

Общая трудоемкость освоения дисциплины составляет 3 зачетных единицы, 108 часов.

Содержание

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Цель освоения дисциплины **«Текстовые данные»** — изучить подходы к количественному анализу текстов в общественных науках.

Задачи:

1. Знакомство с ключевыми источниками текстовых данных в общественных науках, введение в корпусные исследования;
2. Получение навыков по созданию массивов структурированных текстов из неструктурированных данных;
3. Введение в проблемы вычислительной лингвистики.

2. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ

В результате изучения учебной дисциплины обучающийся должен овладеть следующими компетенциями: 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)
ПК-1 Способен обобщать и критически оценивать результаты, полученные отечественными и зарубежными исследователями, выявлять перспективные направления, обосновывать актуальность, теоретическую и практическую значимость избранной темы научного исследования	ИД.ПК-1.1. Осуществляет сбор основных результатов новейших исследований, опубликованных в ведущих профессиональных журналах ИД.ПК-1.2. Критически оценивает актуальность и эффективность основных результатов новейших исследований, опубликованных в ведущих профессиональных журналах в области экономики и смежных наук ИД.ПК-1.3. На основе критического анализа выявляет перспективные направления экономических исследований ИД.ПК-1.4. Обосновывает актуальность, теоретическую и практическую значимость избранной темы научного исследования	Знать: об основных результатах новейших исследований, опубликованных в ведущих профессиональных журналах, методах обобщения и анализа информации; алгоритмах опытно-поисковой деятельности; методах сбора и анализа информации по темам научных исследований З (ПК-1)
		Уметь: осваивать новые предметные области, теоретические и эмпирические методы и приемы научного исследования, осмысливать результаты исследований, делать научные обобщения и применять приобретенные знания в различных областях У (ПК-1)
		Владеть: свободно владеть понятийным аппаратом и навыками научного анализа и методологией научного подхода В (ПК-1)
ПК-4 Способен анализировать и разрабатывать методические материалы, локальные нормативные акты по управлению рисками, формулировать рекомендации по оптимизации процесса управления	ИД.ПК-4.1. Формирование методологических основ интегральной системы управления рисками, формирование основных принципов разработки локальных нормативных актов по управлению рисками на уровне крупных организаций и подразделений ИД.ПК-4.2. Разработка стандартов организации, методических и нормативных документов в сфере обеспечения функционирования и координации процесса управления рисками	Знать: национальные и международные стандарты, лучшие практики по построению систем управления рисками, законодательство Российской Федерации и отраслевые стандарты по управлению рисками З (ПК-4)
		Уметь: внедрять системы управления рисками на уровне организации, подразделения, анализировать изменения корпоративной нормативной базы по вопросам управления рисками, выявлять внешний и внутренний

Код и наименование компетенции	Индикаторы достижения компетенции	Результаты обучения (знать, уметь, владеть)
рисками, упорядочивать процесс управления рисками в целостную систему с четко определенными характеристиками и структурой	ИД.ПК-4.3. Консультирование по вопросам управления рисками в организации ИД.ПК-4.4. Поддержание и совершенствование культуры управления рисками в организации	контекст функционирования организации, разрабатывать регламентирующие документы по управлению рисками, применять термины и принципы риск-менеджмента, описывать бизнес-процессы с учетом рисков, вырабатывать рекомендации по принятию решений в сфере управления рисками У (ПК-4)
		Владеть: навыками декомпозиции стратегических целей организации в задачи подразделения на основании корпоративных нормативных документов по управлению рисками, разработки регламентов деятельности подразделения по управлению рисками и отдельных работников, реализации плана построения системы управления рисками В (ПК-4)

В результате освоения дисциплины магистрант должен:

— **знать**: понятие корпуса текстовых данных, проблемы источников данных в корпусе и способы их преодоления, основные классические задачи обработки естественных языков, разреженное и уплотненное векторные представления текстовых данных, основы поиска по представлениям текстовых данных;

— **уметь**: использовать корпусы текстовых данных в исследованиях, преодолевать проблемы источников данных в текстовых корпусах, оценивать результаты процедур лемматизации, анализировать коллокации, факторизировать матрицы, оценивать результаты семантического поиска;

— **владеть**: навыками использования корпусов текстовых данных в исследованиях, навыками выбора подходящих решений задачи извлечения именованных сущностей, навыками количественного представления текстов, навыками использования метрик и расстояний при работе с текстовыми данными.

3. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Дисциплина **«Текстовые данные»** является дисциплиной по выбору части, формируемой участниками образовательных отношений, Блока 1 «Дисциплины (модули)» образовательной программы «Экономика и финансы». Код дисциплины по Учебному плану Б1.В.ДВ.07.01. Курс читается в восьмом модуле, форма промежуточной аттестации – зачет с оценкой.

Для успешного освоения данной дисциплины требуются знания, полученные в рамках прохождения обучения на уровне бакалавриата/ специалитета.

Знания, умения и навыки, полученные при освоении данной дисциплины, применяются магистрантами в процессе изучения различных дисциплин, а также прохождения НИР.

4. ОБЪЕМ ДИСЦИПЛИНЫ

Общая трудоемкость освоения дисциплины составляет 3 (три) зачетных единицы, 108 часов.

Таблица 2

Объем дисциплины

Типы учебных занятий и самостоятельная работа		Объем дисциплины										
		Всего	Модуль									
			1	2	3	4	5	6	7	8	9	10
Очная форма обучения												
Контактная работа обучающихся с преподавателем в соответствии с УП:		**28**	-	-	-	-	-	-	-	**28**	-	-
лекционного типа (Лек)		14	-	-	-	-	-	-	-	14	-	-
практические занятия (Пр)		14	-	-	-	-	-	-	-	14	-	-
Самостоятельная работа обучающихся (СР)		**80**	-	-	-	-	-	-	-	**80**	-	-
Промежуточная аттестация	**форма**	**Зачет с оценкой**	-	-	-	-	-	-	-	**Зачет с оценкой**	-	-
	час.	**-**	-	-	-	-	-	-	-	**-**	-	-
Общая трудоемкость (час. / з.е.)		**108/3**	-	-	-	-	-	-	-	**108/3**	-	-

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) по средствам 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 соот. с Таблицей 1)	Коды ЗУВ (в соот. с Таблицей 1)
1	Текст как данные	Примеры исследований в этой традиции: Google Flu Trends, измерение идеологии, неопределенность экономической политики. Отличия от тематического, дискурс- и контент-анализа. Представление текстовых данных. Шаги в работе с текстом как данными	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4. ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-4) У (ПК-4) В (ПК-4)
2	Хранение текстовых данных	Способы хранения данных текстовых данных: файловая система, JSON, JSONL, XML, база данных. Извлечение информации из JSON. Введение в XPath. Введение в SQL.	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4. ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-4) У (ПК-4) В (ПК-4)
3	Распознавание текстов	Извлечение таблиц из PDF-файлов. Извлечение текстов и метаданных из PDF-файлов с встроенным текстом. Извлечение текстов из PDF-файлов с изображениями: локальные решения против облачных. Подготовка изображений к распознаванию. Когда распознавание текстов не работает.	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4. ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-4) У (ПК-4) В (ПК-4)
4	Извлечение сущностей из текстовых данных	Задачи обработки естественных языков: от извлечения сущностей до извлечения отношений.	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3.	З (ПК-1) У (ПК-1) В (ПК-1)

№ п/п	Наименование тем (разделов)	Содержание тем (разделов)	Коды компетенций	Индикаторы компетенций (в соот. с Таблицей 1)	Коды ЗУВ (в соот. с Таблицей 1)
		Морфоситаксическая разметка. Формат CoNLL-U. Разбор задачи извлечения именованных сущностей.		ИД.ПК-1.4. ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	З (ПК-4) У (ПК-4) В (ПК-4)
5	Разреженное векторное представление текстовых данных	«Мешок слов» (bag of words) для представления текстовых данных. Возможности и ограничения токенизации. n- и q-граммы. Закон Ципфа. Терм-документные матрицы. Способы взвешивания терм-документных матриц.	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4. ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-4) У (ПК-4) В (ПК-4)
6	Уплотненное векторное представление текстовых данных	Измерение расстояния между текстами: строковое, на базе q-грамм, на базе n-грамм. Сжатие терм-документных матриц. Латентно-семантический анализ. Латентное размещение Дирихле, тематическое моделирование. Структурное тематическое моделирование.	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4. ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-4) У (ПК-4) В (ПК-4)
7	Анализ коллокаций. Дистрибутивная семантика	Дистрибутивная гипотеза. CBOW/Skipgram-представление текстов. Матрицы совстречаемости. Вычисление семантической дистанции: косинусное расстояние. Векторизация с помощью word2vec-моделей.	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4. ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-4) У (ПК-4) В (ПК-4)

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	ПЗ		
Очная форма обучения						
Тема 1	Текст как данные	15	2	2	11	ДЗ
Тема 2	Хранение текстовых данных	15	2	2	11	ДЗ
Тема 3	Распознавание текстов	15	2	2	11	ДЗ
Тема 4	Извлечение сущностей из текстовых данных	15	2	2	11	ДЗ
Тема 5	Разреженное векторное представление текстовых данных	15	2	2	11	ДЗ
Тема 6	Уплотненное векторное представление текстовых данных	15	2	2	11	ДЗ
Тема 7	Анализ коллокаций. Дистрибутивная семантика	18	2	2	14	ДЗ
Промежуточная аттестация		-	-	-	-	Зачет с оценкой
Всего:		**108/3**	**14**	**14**	**80**	**-**

**Примечание: формы текущего контроля успеваемости: домашнее задание (ДЗ).*

6. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ

6.1 Общие положения

Знания и навыки, полученные в результате лекций и занятий семинарского типа, 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/ практическим (семинарским) занятиям. Литературу, рекомендованную в программе курса, следует, по возможности, читать в течение всего модуля, концентрируясь на обусловленных программой курса темах.

Существенную часть самостоятельной работы магистранта представляет самостоятельное изучение вспомогательных учебно-методических изданий, лекционных конспектов, интернет-ресурсов и пр. Подготовка к практическим занятиям 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. Текст как данные.

1.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 5 часов.

1.2. Подготовка к практическим занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций, выполнение домашнего задания – 6 часов.

Итого: 11 часов.

Тема 2. Хранение текстовых данных.

2.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 5 часов.

2.2. Подготовка к практическим занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций, выполнение домашнего задания – 6 часов.

Итого: 11 часов.

Тема 3. Распознавание текстов.

3.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 5 часов.

3.2. Подготовка к практическим занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций, выполнение домашнего задания – 6 часов.

Итого: 11 часов.

Тема 4. Извлечение сущностей из текстовых данных.

4.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 5 часов.

4.2. Подготовка к практическим занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций, выполнение домашнего задания – 6 часов.

Итого: 11 часов.

Тема 5. Разреженное векторное представление текстовых данных.

5.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 5 часов.

5.2. Подготовка к практическим занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций, выполнение домашнего задания – 6 часов.

Итого: 11 часов.

Тема 6. Уплотненное векторное представление текстовых данных.

6.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 5 часов.

6.2. Подготовка к практическим занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций, выполнение домашнего задания – 6 часов.

Итого: 11 часов.

Тема 7. Анализ коллокаций. Дистрибутивная семантика.

7.1. Изучение вопросов, представленных в списке тем лекций. Повторение изученного на предыдущих лекциях материала при подготовке к последующим лекциям – 7 часов.

7.2. Подготовка к практическим занятиям по предложенным темам, самостоятельное изучение рекомендованной литературы, повторение материала лекций, выполнение домашнего задания – 7 часов.

Итого: 14 часов.

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

Вопросы для самостоятельной подготовки по темам дисциплины:

1. Почему закон Ципфа имеет разные эмпирические параметры распределения при разных единицах сегментации корпуса?
2. В каких случаях модель «мешок слов» не применима для представления текстовых данных?
3. Перечислите основные проблемы при оптическом распознавании изображений отсканированных документов.
4. Какие «золотые стандарты» разметки текстов для эвалюации качества распознавания именованных сущностей на русском языке вам известны?
5. Почему расчет строкового расстояния — вычислительно емкая операция?
6. Какие алгоритмы извлечения коллокаций вы знаете?
7. В чем разница между SVD и LSA?
8. Как использовать разреженность терм-документарных матриц для ускорения

вычислений?

9. Опишите задачи, где метод ближайших соседей неприменим.

6.4 Перечень литературы для самостоятельной работы обучающегося:

1. Хименко, В.И. Случайные данные: структура и анализ / В.И. Хименко. – Москва : Техносфера, 2017. – 424 с. : ил.,табл., схем. – (Мир фотоники). – Режим доступа: по подписке. – URL: http://biblioclub.ru/index.php?page=book&id=496479
2. Dan Jurafsky и James Martin (2017). Speech and language processing. 3-е изд. Pearson. – Режим доступа: свободный. – https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

6.5 Перечень учебно-методического обеспечения для самостоятельной работы

Для обеспечения самостоятельной работы магистрантов по дисциплине «**Текстовые данные**» 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 практическому занятию, подготовку домашних заданий, активное слушание на лекциях.

Текущий контроль проводится в форме оценивания выполненных домашних заданий, демонстрирующих степень знакомства с основной и дополнительной литературой.

Таблица 5

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

Наименование тем (разделов)	**Коды компетенций**	**Индикаторы компетенций**	**Коды ЗУВ** *(в соот. с Таблицей 1)*	**Формы текущего контроля успеваемости**	**Результаты текущего контроля**
Текст как данные	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4. ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-4) У (ПК-4) В (ПК-4)	Домашнее задание 1	зачтено/ не зачтено
Хранение текстовых данных	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4.	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-4)	Домашнее задание 2	зачтено/ не зачтено

Наименование тем (разделов)	**Коды компетенций**	**Индикаторы компетенций**	**Коды ЗУВ** *(в соот. с Таблицей 1)*	**Формы текущего контроля успеваемости**	**Результаты текущего контроля**
		ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	У (ПК-4) В (ПК-4)		
Распознавание текстов	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4. ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-4) У (ПК-4) В (ПК-4)	Домашнее задание 3	зачтено/ не зачтено
Извлечение сущностей из текстовых данных	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4. ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-4) У (ПК-4) В (ПК-4)	Домашнее задание 4	зачтено/ не зачтено
Разреженное векторное представление текстовых данных	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4. ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-4) У (ПК-4) В (ПК-4)	Домашнее задание 5	зачтено/ не зачтено
Уплотненное векторное представление текстовых данных	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4. ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-4) У (ПК-4) В (ПК-4)	Домашнее задание 6	зачтено/ не зачтено
Анализ коллокаций. Дистрибутивная семантика	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4. ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-4) У (ПК-4) В (ПК-4)	Домашнее задание 7	зачтено/ не зачтено

При выполнении всех видов заданий должны быть исключены заимствования из чужих работ (за исключением тех, что снабжены ссылками).

В зависимости от сложности домашнего задания объявляется максимальный балл, который обучающийся может получить за его выполнение. Максимальные баллы за все домашние задания в сумме составляют 40 баллов.

При освоении дисциплины каждая из форм текущего контроля оценивается с использованием балльной шкалы (для каждого задания указывается максимальное число баллов) с последующим переводом в бинарную систему для получения результатов текущего контроля, фиксирующих ход образовательного процесса, согласно Таблице 6.

Таблица 6

Критерии оценивания

Формы текущего контроля успеваемости	Описание	Показатели	Количество баллов по 100-балльной шкале	Результаты текущего контроля
Домашнее 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	если дан полный и правильный ответ /решение, возможны несущественные погрешности	81–100	зачтено
		если дан правильный, но неполный ответ/решение, возможны несущественные погрешности	61–80	
		если выявлено неполное знание или частично неправильная трактовка основополагающих положений и предпосылок, присутствуют грубые ошибки	41–60	
		если выявлено незнание или неправильная трактовка основополагающих положений и предпосылок, присутствуют грубые ошибки	0–40	не зачтено
		если решалась задача, отличная от предложенной, или если ответ/решение отсутствует	0 баллов	не зачтено

7.2 Контрольные задания для текущей аттестации

Примерный материал домашних заданий:

Тема 1. Текст как данные.

Домашнее задание 1:

На данных поисковых запросов Яндекса, связанных с ковидом (URL: https://datalens.yandex.ru/c96rz1dyyf5k4-koronavirus-dannye-dlya-eksporta) вычислите корреляцию между заболеваемостью на уровне региона и самыми популярными запросами.

Тема 2. Хранение текстовых данных.

Домашнее задание 2:

Прочитать и записать текстовый корпус разыми методами: из SQLite базы, из XML, из JSON(L).

Тема 3. Распознавание текстов.

Домашнее задание 3:

Разработать конвейер по превращению заданных PDF-изображений в текстовые файлы с заданной структурой.

Тема 4. Извлечение сущностей из текстовых данных.

Домашнее задание 4:

Извлечь из заданного корпуса все именованные сущности, удовлетворяющие определенному требованию.

Тема 5. Разреженное векторное представление текстовых данных.

Домашнее задание 5:

Оценить на терм-документарной матрице вектора документов предложенными алгоритмами.

Тема 6. Уплотненное векторное представление текстовых данных.

Домашнее задание 6:

Рассчитать уплотненные вектора слов на заданном корпусе, сравнить результаты с результатами проекта RusVectōrēs.

Тема 7. Анализ коллокаций. Дистрибутивная семантика.

Домашнее задание 7

Найти в заданном корпусе, преобразованном в формат терм-документной матрицы, ближайших соседей.

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

Форма промежуточной аттестации **– зачет с оценкой**, выставляемый на основе оценки финального проекта, подготовленного магистрантом.

Перед зачетом с оценкой проводятся консультация, на которой преподаватель отвечает на вопросы обучающихся.

Критерии оценивания финального проекта представлены в таблице 7.

Таблица 7

Критерии оценивания финального проекта

Вид промежуточной аттестации	Показатели	Количество баллов
Финальный проект	Магистрант демонстрирует полную самостоятельность в подборе фактического материала и аналитическое отношении к нему, умение рассматривать примеры и факты во взаимосвязи и взаимообусловленности, отбирать наиболее существенные из них; а также показывает грамотное использование методов описания и презентации исследования	81–100
	Магистрант демонстрирует самостоятельность в подборе фактического материала и аналитическое отношение к нему, в большинстве случаев видит взаимосвязь примеров и фактов, в целом отбирает существенные из них; а также использует правильные методы описания и презентации исследования с небольшими ошибками	61–80
	Магистрант демонстрирует аналитическое отношение к материалу, видит взаимосвязь некоторых примеров и фактов; а также использует методы описания и презентации исследования с большим количеством несущественных ошибок	41–60
	Магистрант не демонстрирует аналитическое отношение к материалу, не видит взаимосвязь примеров и фактов; а также использует методы описания и презентации исследования с большим количеством существенных ошибок	0–40

В результате промежуточного контроля знаний студенты получают аттестацию по дисциплине. На основании оценки обучающегося по итогам освоения дисциплины, выраженной в 100-балльной шкале, выставляется **зачет с оценкой** в соответствии с Таблицей 8.

Таблица 8

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

Форма промежуточной аттестации/вид промежуточной аттестации	**Коды компете-нций**	**Индикаторы компетенций** *(в соот. с Таблицей 1)*	**Коды ЗУВ** *(в соот. с Таблицей 1)*	**Оценка по итогам освоения дисциплины (в 100-балльной шкале)**	**Результаты текущего контроля**
зачет с оценкой / финальный проект	ПК-1 ПК-4	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4. ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-4) У (ПК-4) В (ПК-4)	81–100	Зачтено, отлично
				61–80	Зачтено, хорошо
				41–60	Зачтено, удовлетворительно
				0–40	Не зачтено, неудовлетворительно

Результаты промежуточной аттестации по дисциплине, выраженные в оценках «зачтено, удовлетворительно», «зачтено, хорошо», «зачтено, отлично», свидетельствуют о сформированности у обучающегося компетенций по дисциплине в соответствии с картами компетенций образовательной программы «Экономика и финансы» по направлению подготовки 38.04.01 Экономика (уровень магистратуры).

Результат промежуточной аттестации по дисциплине, выраженный в оценке «не зачтено, неудовлетворительно», свидетельствует об отсутствии или критическом уровне сформированности у обучающегося компетенций по дисциплине в соответствии с картами компетенций образовательной программы «Экономика и финансы» по направлению подготовки 38.04.01 Экономика (уровень магистратуры).

7.4 Типовые задания к промежуточной аттестации

Примерные требования к проекту для промежуточной аттестации по дисциплине:

1. Тему проекта магистрант выбирает, основываясь на своих научных интересах, и согласовывает с преподавателем заранее.
2. Проект представляется в виде созданного текстового корпуса и сопроводительного документа-описания его источников, структуры и ключевых характеристик.

7.5 Средства оценки индикаторов достижения компетенций

Таблица 9

Средства оценки индикаторов достижения компетенций

Коды компетенций	**Индикаторы компетенций** *(в соот.с Таблицей 1)*	**Средства оценки** *(в соот. с Таблицами 5, 7)*
ПК-1	ИД.ПК-1.1. ИД.ПК-1.2. ИД.ПК-1.3. ИД.ПК-1.4.	домашнее задание
ПК-4	ИД.ПК-4.1. ИД.ПК-4.2. ИД.ПК-4.3. ИД.ПК-4.4.	Домашнее задание

Таблица 10

Описание средств оценки индикаторов достижения компетенций

Средства оценки *(в соот. С Таблицами 5, 7)*	**Рекомендованный план выполнения работы**
Домашнее задание	Магистрант в ходе подготовки и выполнения домашних заданий по темам дисциплины показывает способность совершать следующий набор профессиональных действий, получивший развитие в рамках данной дисциплины: 1. Выбирает тему научного исследования на основе результатов оценки отечественных и зарубежных течений в данной области, обосновывает актуальность, теоретическую и практическую значимость избранной темы научного исследования 2. Анализирует различные методические материалы по управлению рисками, формулирует перечень рекомендаций по оптимизации процесса управления рисками, учитывает необходимость построения целостной системы управления рисками
Финальный проект	Магистрант в ходе подготовки и выполнения финального проекта, показывает способность совершать следующий набор профессиональных действий, получивший развитие в рамках данной дисциплины: **1.** Анализирует корпусы текстовых данных в исследованиях, преодолевает проблемы источников данных в текстовых корпусах, оценивает результаты семантического поиска, 2. Демонстрирует навыки выбора подходящих решений задачи извлечения именованных сущностей, навыки количественного представления текстов, навыки использования метрик и расстояний при работе с текстовыми данными

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 . – Режим доступа: по подписке

2. Dan Jurafsky и James Martin (2017). Speech and language processing. 3-е изд. Pearson. – Режим доступа: свободный. – https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

8.2. Дополнительная литература

1. Matthew Gentzkow, Bryan Kelly и Matt Taddy (2019). “Text as data”. B: Journal of Economic Literature 57.3

2. Kenneth Benoit (2019). “SAGE Handbook of Research Methods in Political Science and International Relations”. B: под ред. Luigi Curini и Robert Franzese. SAGE Publishing. Гл. Text as data: An overview

3. Nikolaj Lindberg (2007). “egrep for Linguists”. https://stts.se/egrep_for_linguists/egrep_for_linguists.pdf

4. Ray Smith (2007). “An overview of the Tesseract OCR engine”. B: Ninth international conference on document analysis and recognition (ICDAR 2007). T. 2. IEEE, c. 629—633
5. Документация Yandex Vision API: https://cloud.yandex.ru/docs/vision/operations/ocr/text-detection
6. http://nlpprogress.com
7. Milan Straka и Jana Strakova´ (2017). “Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe”. B: Proceedings of the CoNLL 2017 Shared
8. Task: Multilingual Parsing from Raw Text to Universal Dependencies, c. 88— 99
9. E Bolshakova и др. (2017).“Avtomaticheskaya obrabotka tekstov na estestvennom yazike i analiz dannih”. HSE, https://www.hse.ru/data/2017/08/12/1174382135/NLP_and_DA.pdf
10. Jake Williams и др. (2015). “Zipf’s law holds for phrases, not words”. B: Scientific Reports 5, c. 12209

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. R — бесплатно
14. Python — бесплатно

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. Google. Книги: https://books.google.com
2. Internet Archive: https://archive.org
3. Koob.ru. Электронная библиотека «Куб»: http://www.koob.ru/philosophy/
4. Библиотека Гумер – гуманитарные науки: http://www.gumer.info
5. Библиотека Ихтика [ihtik.lib.ru]: http://ihtik.lib.ru/
6. Докусфера — Российская национальная библиотека: http://leb.nlr.ru
7. ЕНИП — Электронная библиотека «Научное наследие России»: http://e-heritage.ru/
8. Интелрос. Интеллектуальная Россия: http://www.intelros.ru/
9. Национальная электронная библиотека НЭБ: http://www.rusneb.ru
10. Президентская библиотека: http://www.prlib.ru
11. Российская государственная библиотека: http://www.rsl.ru/
12. 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 и корпоративную электронную почту Университета, 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 (практических занятий, лабораторных работ), 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 электронным ресурсам библиотеки Университета 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).