


Документ подписан простой электронной подписью
Информация о владельце:
ФИО: Волков В.В.
Должность: Ректор
Дата подписания: 15.07.2024 14:54:56
Уникальный программный ключ:
ed68fd4b85b778e0f0b1bfea5dbc56cf4148f1229917e799a70e51517ff6d591

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

УТВЕРЖДАЮ

Ректор  / В.В. Волков

« 28 » 28 февраля 2024 г.

Протокол Ученого Совета

№ 2 от 28 февраля 2024 г.



Рабочая программа дисциплины
«Текстовые данные»

дополнительная профессиональная программа
«Прикладной анализ данных»

вид программы
программа профессиональной переподготовки

язык обучения – русский
форма обучения – очная

Санкт-Петербург

Авторы:

Тушканова О.Н., кандидат технических наук, доцент факультета социологии АНООВО «ЕУСПб

Рабочая программа дисциплины «Текстовые данные», входящая в состав дополнительной профессиональной программы профессиональной переподготовки «Прикладной анализ данных» утверждена на заседании Ученого совета университета.

Содержание

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ. ОБЪЕМ ДИСЦИПЛИНЫ.....	4
2. ТРЕБОВАНИЯ К УРОВНЮ ОСВОЕНИЯ СОДЕРЖАНИЯ ДИСЦИПЛИНЫ.....	4
3. СОДЕРЖАНИЕ И СТРУКТУРА ДИСЦИПЛИНЫ.....	5
4. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ	7
5. ОЦЕНОЧНЫЕ СРЕДСТВА.....	8
6. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ.....	10
7. ПРОГРАММНОЕ И МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ПРОГРАММЫ.....	11

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ. ОБЪЕМ ДИСЦИПЛИНЫ

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

Задачи обучения:

- знакомство с ключевыми источниками текстовых данных в общественных науках, введение в корпусные исследования;
- получение навыков по созданию массивов структурированных текстов из неструктурированных данных;
- введение в проблемы вычислительной лингвистики.

Изучение данной дисциплины способствует формированию профессиональных навыков работы с информацией.

Отличительной особенностью реализуемого подхода к преподаванию дисциплины является разнообразных практических иллюстраций основных теоретических положений.

Общая трудоемкость дисциплины составляет 54 часа.

2. ТРЕБОВАНИЯ К УРОВНЮ ОСВОЕНИЯ СОДЕРЖАНИЯ ДИСЦИПЛИНЫ

В результате освоения дисциплины слушатель должен приобрести следующие знания и умения, необходимые для качественного изменения профессиональных компетенций:

слушатель должен знать:

- источники текстовых данных в общественных науках;
- введение в корпусные исследования;
- введение в проблемы вычислительной лингвистики;
- особенности создания массивов структурированных текстов из неструктурированных данных.

слушатель должен уметь:

- работать с ключевыми источниками текстовых данных в общественных науках;
- работать с массивами структурированных текстов, созданных из неструктурированных данных;
- использовать полученные знания и умения в профессиональной деятельности;

слушатель должен владеть:

- навыками по созданию массивов структурированных текстов из неструктурированных данных;
- сбора и обработки данных, в том числе с использованием технологий прикладного анализа данных.

В результате изучения дисциплины «Текстовые данные» слушатель приобретает следующие профессиональные компетенции (Таблица 1):

Планируемые результаты обучения по дисциплине

Таблица 1

Код и название компетенции	Содержание компетенции	Планируемые результаты обучения по дисциплине, характеризующие этапам формирования компетенций
ПК-1	способен разрабатывать методики выполнения аналитических работ	Знать: З (ПК-1) – современные методики аналитических работ в изучаемой сфере
		Уметь: У (ПК-1) – разрабатывать методики выполнения аналитических работ
		Владеть: В (ПК-1) - навыками выполнения аналитических работ в соответствии с современными методиками

Код и название компетенции	Содержание компетенции	Планируемые результаты обучения по дисциплине, характеризующие этапам формирования компетенций
ПК-3	способен управлять аналитическими ресурсами и компетенциями	Знать: З (ПК-3) – основы управления аналитическими ресурсами и компетенциями
		Уметь: У (ПК-3) – собирать и систематизировать данные необходимые для управления аналитическими ресурсами и компетенциями
		Владеть: В (ПК-3) – навыками управления аналитическими ресурсами и компетенциями
ПК-5	способен применить анализ данных к научным и общественным задачам	Знать: З (ПК-5) – основы анализа данных
		Уметь: У (ПК-5) – использовать методики анализа данных применительно к различным типам данных
		Владеть: В (ПК-5) – навыками анализа данных с учетом специфики научных и общественных задач

3. 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) в соответствии с таблицей 2.

Таблица 2

Содержание дисциплины

№ п/п	Наименование тем (разделов)	Содержание тем (разделов)	Коды компетенций	Коды ЗУВ (в соответствии с табл. 1)
1	Корпусы текстовых данных	Корпусные исследования. Описание корпусов текстовых данных: КРЯ, Wikipedia, CommonCrawl, Taiga. Разбор принципов их организации и причин создания. Примеры использования в исследованиях.	ПК-1 ПК-3 ПК-5	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-3) У (ПК-3) В (ПК-3) З (ПК-5) У (ПК-5) В (ПК-5)
2	Распознавание текстов. Создание процедуры превращения изображений или PDF в корпус	Организация корпуса. Проблемы источников данных и способы их преодоления.	ПК-1 ПК-3 ПК-5	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-3) У (ПК-3) В (ПК-3) З (ПК-5) У (ПК-5) В (ПК-5)
3	Закон Ципфа. Издержки токенизации	Введение в количественное представление текстов. Разбор процедур лемматизации, эвалюация их результатов	ПК-1 ПК-3 ПК-5	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-3) У (ПК-3) В (ПК-3) З (ПК-5) У (ПК-5)

№ п/п	Наименование тем (разделов)	Содержание тем (разделов)	Коды компетенций	Коды ЗУВ (в соответствии с табл. 1)
				В (ПК-5)
4	Извлечение сущностей из текстовых данных	Введение в классические задачи обработки естественных языков. Детальный разбор задачи извлечения именованных сущностей. Обзор существующих решений. Проблема эвалюирования.	ПК-1 ПК-3 ПК-5	3 (ПК-1) У (ПК-1) В (ПК-1) 3 (ПК-3) У (ПК-3) В (ПК-3) 3 (ПК-5) У (ПК-5) В (ПК-5)
5	Разреженное векторное представление текстовых данных	Строковые расстояния. Модель «мешок слов». Byte Pair Encoding. Анализ коллокаций.	ПК-1 ПК-3 ПК-5	3 (ПК-1) У (ПК-1) В (ПК-1) 3 (ПК-3) У (ПК-3) В (ПК-3) 3 (ПК-5) У (ПК-5) В (ПК-5)
6	Уплотненное векторное представление текстовых данных	Факторизация матриц. SVD, LSA, LDA. Дистрибутивная семантика. От word2vec к контекстуальным эмбедингам.	ПК-1 ПК-3 ПК-5	3 (ПК-1) У (ПК-1) В (ПК-1) 3 (ПК-3) У (ПК-3) В (ПК-3) 3 (ПК-5) У (ПК-5) В (ПК-5)
7	Поиск по представлениям текстовых данных	Разреженные матрицы. Виды метрик и расстояний. Метод ближайших средних, его ограничения. Эвалюация результатов семантического поиска. Возможности и ограничения семантического поиска.	ПК-1 ПК-3 ПК-5	3 (ПК-1) У (ПК-1) В (ПК-1) 3 (ПК-3) У (ПК-3) В (ПК-3) 3 (ПК-5) У (ПК-5) В (ПК-5)

Структура дисциплины

Таблица 3.

№ п/п	Наименование и содержание тем	Основные понятия (категории) и проблемы, рассматриваемые в теме	Объем дисциплины, час.				Форма текущего контроля успеваемости, промежуточной аттестации
			Всего	Аудиторная работа по видам учебных занятий		СР ¹	
				Л	СЗ ²		
1.	Текстовые данные вокруг нас	Корпусы текстовых данных: КРЯ, Wikipedia, CommonCrawl, Taiga.	7	2	2	3	опрос, диспут, практические задания
2.	Распознавание текстов	Распознавание текстов. Создание процедуры превращения изображений или PDF в корпус. Tesseract + ImageMagick.	8	2	2	4	
3.	Чем текстовые данные отличаются от других	Закон Ципфа. Издержки токенизации.	7	2	2	3	

¹ Самостоятельная работа, включает в себя часы на промежуточный контроль

² Могут включать в себя: лабораторные работы, круглые столы, мастер-классы, мастерские, деловые игры, ролевые игры, тренинги, семинары по обмену опытом, выездные занятия, консультации

№ п/п	Наименование и содержание тем	Основные понятия (категории) и проблемы, рассматриваемые в теме	Объем дисциплины, час.			Форма текущего контроля успеваемости, промежуточной аттестации	
			Всего	Аудиторная работа по видам учебных занятий			СР ¹
				Л	СЗ ²		
	типов данных						
4.	Извлечение сущностей из текстовых данных	Извлечение сущностей из текстовых данных. Natasha, Pullenti.	7	2	2	3	
5.	Разреженное векторное представление текстовых данных	Разреженное векторное представление текстовых данных. Строковые расстояния. Модель «мешок слов». Byte Pair Encoding. Анализ коллокаций.	8	2	2	4	
6.	Уплотненное векторное представление текстовых данных	Уплотненное векторное представление текстовых данных. Факторизация матриц. SVD, LSA, LDA, BigARTM. Дистрибутивная семантика. От word2vec к контекстуальным эмбедингам. RusVectoRēs.	7	2	2	3	
7.	Поиск по представлениям текстовых данных	Поиск по представлениям текстовых данных. Метод (приближенного) поиска ближайших соседей. Возможности и ограничения семантического поиска	8	2	2	4	
8.	Промежуточная аттестация	Проект	2	-	-	2	зачет
Всего:			54	14	14	26	

4. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

Общие положения.

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

Самостоятельная работа обучающегося представляет самостоятельное изучение дополнительных материалов, Интернет-ресурсов и пр. Подготовка к семинарским занятиям, создание докладов, проектов и презентаций также является важной формой работы обучающихся. Самостоятельная работа может вестись как индивидуально, так и при содействии преподавателя. Вопросы и замечания, возникшие в ходе самостоятельного внеаудиторного чтения рекомендованной литературы, обсуждаются с преподавателем и другими обучающимися. Выносятся на обсуждение, как правило, актуальные проблемы и предлагается их рассмотреть с точки зрения того или иного теоретического подхода.

На занятиях материал излагается в проблемной форме. Основной упор в преподавании делается на изучение теоретических понятий и возможности их применения на конкретных примерах, в том числе в устных выступлениях обучающихся.

Перечень основных вопросов по изучаемым темам для самостоятельной работы обучающихся.

- Почему закон Ципфа имеет разные эмпирические параметры распределения при разных единицах сегментации корпуса?
- В каких случаях модель «мешок слов» не применима для представления текстовых данных?
- Перечислите основные проблемы при оптическом распознавании изображений

- отсканированных документов.
- Какие «золотые стандарты» разметки текстов для эвалюации качества распознавания именованных сущностей на русском языке вам известны?
- Почему расчет строкового расстояния — вычислительно емкая операция?
- Какие алгоритмы извлечения коллокаций вы знаете?
- В чем разница между SVD и LSA?
- Как использовать разреженность терм-документарных матриц для ускорения вычислений?
- Опишите задачи, где метод ближайших соседей неприменим.

Источники для самостоятельной подготовки:

- 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>
- Основы научных исследований / Б.И. Герасимов, В.В. Дробышева, Н.В. Злобина и др. - М.: Форум: НИЦ Инфра-М, 2013. - 272 с. - [Электронный ресурс]. -URL: <http://znanium.com/bookread2.php?book=390595>
- Общая теория статистики: Учебное пособие / С.Н. Лысенко, И.А. Дмитриева. - Изд., испр. и доп. - М.: Вузовский учебник: НИЦ ИНФРА-М, 2014. - 219 с. - [Электронный ресурс]. URL: <http://znanium.com/bookread2.php?book=397795>
- Маркин, А.В. Построение запросов и программирование на SQL: учебное пособие / А.В. Маркин. – 3-е изд., перераб. и доп. – Москва: Диалог-МИФИ, 2014. – 384 с.: ил. – Режим доступа: по подписке. – URL: <http://biblioclub.ru/index.php?page=book&id=89077>

5. ОЦЕНОЧНЫЕ СРЕДСТВА

Проведение текущего контроля в рамках реализации данной дисциплины проходит в соответствии с Таблицей 3 данной рабочей программы дисциплины по основным понятиям (категориям) и проблемам, рассматриваемым в предложенных темах. Фиксация результатов текущего контроля в рамках реализации данной дисциплины не предусмотрена.

Типовые задания к текущей аттестации (опросы, диспуты, практические задания).

Опрос 1:

Выбрать из корпуса Taiga статьи определенного интернет-сми, упоминающие определенное слово.

Практическое задание 1:

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

Практическое задание 2:

Оценить качество сегментации разными методами, сравнить результаты.

Практическое задание 3:

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

Практическое задание 4:

Оценить на терм-документарной матрице вектора документов предложенными алгоритмами.

Практическое задание 5:

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

Практическое задание 6, диспут 1:

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

Критерии оценивания

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	выполнение практического задания с существенными ошибками или пропусками – не зачтено, полное и правильное выполнение практического задания – зачтено
Опрос	ответ отсутствует или является односложным, или содержит существенные ошибки – не зачтено слушатель 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

Форма промежуточной аттестации – зачет, выставляемый на основе подготовленного проекта.

При аттестации используются система «зачтено» и «не зачтено» в соответствии с критериями оценивания.

В результате промежуточного контроля знаний обучающиеся получают аттестацию по дисциплине.

Показатели, критерии и оценивание компетенций по уровням их формирования в процессе промежуточной аттестации

Таблица 4

Форма промежуточной аттестации/вид промежуточной аттестации	Коды компетенций	Коды ЗУВ (в соответствии с Таблицей 1)	Критерии оценивания	Оценка
зачет / проект	ПК-1 ПК-3 ПК-5	З (ПК-1) У (ПК-1) В (ПК-1) З (ПК-3) У (ПК-3) В (ПК-3) З (ПК-5) У (ПК-5) В (ПК-5)	Обучающийся демонстрирует полную самостоятельность в подборе фактического материала и аналитическое отношение 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	зачтено
			Обучающийся 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	не зачтено

Результаты промежуточного контроля по дисциплине, выраженные в бинарной системе «зачтено», показывают уровень сформированности у обучающегося компетенций.

Результаты промежуточного контроля по дисциплине, выраженные в бинарной системе «не зачтено», показывают не сформированность у обучающегося компетенций по дисциплине.

Типовые темы проектов к промежуточной аттестации.

Тему проекта слушатель выбирает, основываясь на своих научных интересах, и согласовывает с преподавателем заранее.

Проект представляется в виде созданного текстового корпуса и сопроводительного документа-описания его источников, структуры и ключевых характеристик.

Примерные темы письменной работы (эссе):

- Разработать конвейер по превращению заданных PDF-изображений в текстовые файлы с заданной структурой.
- Рассчитать уплотненные вектора слов на заданном корпусе, сравнить результаты с результатами проекта RusVectoĝes.
- Оценить на терм-документарной матрице вектора документов предложенными алгоритмами.

6. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Основная литература:

- 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>
- Основы научных исследований / Б.И. Герасимов, В.В. Дробышева, Н.В. Злобина и др. - М.: Форум: НИЦ Инфра-М, 2013. - 272 с. - [Электронный ресурс]. -URL: <http://znanium.com/bookread2.php?book=390595>

Дополнительная литература:

- Общая теория статистики: Учебное пособие / С.Н. Лысенко, И.А. Дмитриева. - Изд., испр. и доп. - М.: Вузовский учебник: НИЦ ИНФРА-М, 2014. - 219 с. - [Электронный ресурс]. URL: <http://znanium.com/bookread2.php?book=397795>
- Маркин, А.В. Построение запросов и программирование на SQL: учебное пособие / А.В. Маркин. – 3-е изд., перераб. и доп. – Москва: Диалог-МИФИ, 2014. – 384 с.: ил. – Режим доступа: по подписке. – URL: <http://biblioclub.ru/index.php?page=book&id=89077>

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

Информационно-справочные системы:

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

Тематические системы:

- Google. Книги: <https://books.google.com>
- Internet Archive: <https://archive.org>
- Koob.ru. Электронная библиотека «Куб»: <http://www.koob.ru/philosophy/>

- Библиотека Ихтика [ihtik.lib.ru]: <http://ihtik.lib.ru/>
- Докусфера — Российская национальная библиотека: <http://leb.nlr.ru>
- ЕНИП — Электронная библиотека «Научное наследие России»: <http://e-heritage.ru/>
- Интелрос. Интеллектуальная Россия: <http://www.intelros.ru/>
- Национальная электронная библиотека НЭБ: <http://www.rusneb.ru>
- Неприкосновенный запас: <http://magazines.russ.ru/nz/>
- Президентская библиотека: <http://www.prlib.ru>
- Российская государственная библиотека: <http://www.rsl.ru/>
- Российская национальная библиотека: <http://www.nlr.ru/poisk/>

7. ПРОГРАММНОЕ И МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ПРОГРАММЫ

В ходе реализации образовательного процесса используются многофункциональные аудитории для проведения занятий лекционного типа, занятий семинарского типа, укомплектованные специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Проведение занятий лекционного типа и семинарского типа обеспечивается демонстрационным оборудованием.

Для обучающихся с ограниченными возможностями здоровья и инвалидов (в случае необходимости) могут быть созданы специальные условия для получения образования.

Программное обеспечение

При осуществлении образовательного процесса в рамках Университета слушателям рекомендовано использовать следующее лицензионное программное обеспечение:

- OS Microsoft Windows (OVS OS Platform)
- MS Office (OVS Office Platform)
- Adobe Acrobat Professional 11.0 MLP AOO License RU
- Adobe CS5.5 Design Standart Win IE EDU CLP
- ABBYY FineReader 11 Corporate Edition
- ABBYY Lingvo x5
- Adobe Photoshop Extended CS6 13.0 MLP AOO License RU
- Adobe Acrobat Reader DC /Pro – бесплатно
- Google Chrome – бесплатно
- Opera – бесплатно
- Mozilla – бесплатно
- VLC – бесплатно