МИНОБРНАУКИ РОССИИ

Федеральное государственное автономное образовательное учреждение
высшего образования
«Российский государственный гуманитарный университет»
(ФГАОУ ВО «РГГУ»)

ИНСТИТУТ ЛИНГВИСТИКИ
Учебно-научный центр компьютерной лингвистики

ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА ДЛЯ ИСТОРИКА

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

46.04.01 История
Код и наименование направления подготовки/специальности

Искусственный интеллект и цифровые технологии в исторических исследованиях
Наименование направленности (профиля)/ специализации

Уровень высшего образования: магистратура

Форма обучения: *очная*

РПД адаптирована для лиц
с ограниченными возможностями
здоровья и инвалидов

Москва 2026

Обработка естественного языка для историка
Рабочая программа дисциплины

Составители:
старший преподаватель, вр.и.о. директора УНЦ компьютерной лингвистики И.Е. Пинхасик
преподаватель УНЦ компьютерной лингвистики Ф.А. Тучак

УТВЕРЖДЕНО
Протокол заседания УНЦ компьютерной лингвистики
№ 6 от 12.12.2025

ОГЛАВЛЕНИЕ

1. Пояснительная записка

1.1. Цель и задачи дисциплины

Цель дисциплины: освоение студентами базовых понятий компьютерной лингвистики и автоматической обработки естественного языка с последующим развитием прикладных навыков и умений в области автоматизированной обработки научных текстов и работы с корпусными данными и электронными текстовыми ресурсами, в том числе применительно к историческим текстам и архивным данным.

Задачи дисциплины:

- освоение основных понятий и терминов по дисциплинам корпусной и компьютерной лингвистики;
- освоение основных библиотек языка программирования Python, используемых для обработки естественного языка;
- получение опыта работы с различными корпусами текстов и базами текстовых данных, в том числе относящихся к историческим материалам.

1.2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с индикаторами достижения компетенций

Компетенция (код и наименование)	**Индикаторы компетенций** (код и наименование)	**Результаты обучения**
ПК-4. Способен ориентироваться в программном обеспечении информационных систем и баз данных историко-ориентированного профиля; создавать историко-ориентированные информационные системы и базы данных; способен использовать в конкретно-исторических исследованиях, основанных на информации массовых исторических источников, методы и технологии математической статистики и компьютерного моделирования, современной науки о данных	ПК-4.1. Умеет ориентироваться в программном обеспечении информационных систем и баз данных, умеет создавать историко-ориентированные информационные системы и базы данных, использовать в конкретно-исторических исследованиях методы и технологии математической статистики и компьютерного моделирования, современной науки о данных	*Знать*: основные понятия и методы современной компьютерной лингвистики; основные понятия и методы современной корпусной лингвистики; базовые принципы лингвистической разметки. *Уметь*: совершать очистку текстовых данных для автоматической обработки и анализа; применять базовые функции и методы библиотек обработки естественного языка Python для решения практических задач по обработке и анализу языковых данных. *Владеть*: навыками отбора текстовых документов по критериям релевантным для решения исследовательской задачи; базовыми методами статистического анализа и инструментами их реализации; базовыми методами параметрической оценки корпуса, включая его взвешенность и репрезентативность.
ПК-5. Способен применять цифровые	ПК-5.1. Владеет цифровыми технологиями	*Знать*: исторические и современные аспекты обработки естественного

технологии анализа данных нарративных, изобразительных, картографических, аудиовизуальных исторических источников; способен использовать методы и технологии 3D-моделирования для виртуальной реконструкции объектов историко-культурного наследия	анализа данных нарративных, изобразительных, картографических, аудиовизуальных исторических источников, методами и технологиями 3D моделирования для виртуальной реконструкции объектов историко-культурного наследия	языка с использованием правиловых, статистических и нейросетевых моделей. *Уметь*: применять базовые функции и методы библиотек визуализации данных и релевантных модулей библиотек обработки естественного языка Python для решения практических задач по визуализации статистических корпусных и языковых данных; применять базовый функционал основных корпусных менеджеров и инструментов автоматического распознавания текста. *Владеть*: основным понятийным аппаратом и навыками реализации визуальных решений в области векторной семантики; навыками создания облаков слов на материале анализируемого корпуса; базовыми навыками извлечения именованных сущностей.

1.3. Место дисциплины в структуре образовательной программы

Дисциплина «Обработка естественного языка для историка» относится к части блока дисциплин учебного плана, формируемой участниками образовательной программы. Необходимы знания, умения и владения, сформированные в результате обучения по дисциплине “Программирование для гуманитарных наук” или аналогичные базовые навыки цифровой грамотности и базового владения языком программирования Python, полученные из иных курсов. В результате освоения дисциплины формируются знания, умения и владения, необходимые для изучения последующих дисциплин как обязательной части учебного плана, так и части, формируемой участниками образовательных отношений.

2. Структура дисциплины

Общая трудоёмкость дисциплины составляет 3 з.е., 108 академических часов.

Структура дисциплины для очной формы обучения

Объем дисциплины в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении учебных занятий:

Семестр	Тип учебных занятий	Количество часов
1	Лекции	20
1	Семинары	22
	Всего:	42

Объем дисциплины в форме самостоятельной работы обучающихся составляет 66 академических часов.

3. Содержание дисциплины

№	Наименование раздела дисциплины	Содержание
1.	**Тема 1. Введение. Компьютерная лингвистика и автоматическая обработка естественного языка**	Компьютерная лингвистика и автоматическая обработка естественного языка. Лингвистический и инженерный подход к компьютерной лингвистике. Задачи компьютерной лингвистики. Краткая история компьютерной лингвистики: правиловые, статистические и нейросетевые модели. Сложности при обработке естественного языка: омонимия, синонимия, проблемы с пониманием прагматики и т. п.
2.	**Тема 2. Основы корпусной лингвистики.**	Корпус как лингвистический объект. Ключевые характеристики корпуса как объекта изучения: существование в реальном мире, потенциально неограниченный объем, разнообразие типов. Сбалансированность и репрезентативность корпуса.
3.	**Тема 3. Работа с онлайн-ресурсами по представлению исторических данных.**	Корпусные ресурсы исторических данных. Цифровая работа с эго-документами. Платформы “Прожито”, “Пишу тебе”. Опыт проекта “Пушкин ⟨цифровой⟩”. Обработка исторических текстов. Платформа Transkribus и основы компьютерного зрения. Особенности работы с историческими подкорпусами НКРЯ. Оцифровка берестяных грамот.
4.	**Тема 4. Работа с корпусными менеджерами.**	Основной функционал корпусных менеджеров (SketchEngine, Voyant Tools, LancsBox, AntConc). Scraping с помощью корпусных менеджеров. Извлечение ключевых слов с помощью существующих электронных ресурсов. Составление списков стоп-слов. Создание облаков слов (word clouds). Частотные списки как основа для сравнения корпусов.
5.	**Тема 5. Основы дистрибутивной семантики и введение в обработку естественного языка (Natural Language Processing, NLP).**	Дистрибутивная гипотеза, закон Ципфа и представление языковых единиц в многомерном пространстве (метод векторизации); принципы и сферы применения NLP; знакомство с основными библиотеками Python, используемыми для обработки естественного языка.
6.	**Тема 6. Основы компьютерной лингвистики и морфологический анализ.**	Сегментация текста на токены и предложения. Неравнозначность токена и слова. Проблемы токенизации и деления на предложения в языках с различными системами графики. Токенизаторы, сентенайзеры, лемматизаторы и стеммеры.
7.	**Тема 7. Основы программирования на Python для обработки естественного языка.**	Основы работы с регулярными выражениями. Модуль Python re. Scraping с помощью инструментов Python. NLTK как классическая библиотека для обработки естественного языка. Обработка морфологии русского

		языка. Natasha и razdel. Spacy. Сравнение работы методов изученных модулей на одинаковых текстах и анализ данных. Частеречная разметка (POS-tagging). Извлечение именованных сущностей.
8.	**Тема 8. Основы работы с открытыми библиотеками обработки естественного языка и статистическая обработка.**	Библиотеки numpy и pandas для обработки данных. Метод describe() в pandas. Среднее арифметическое, медиана, стандартное отклонение, квантили. Хи-квадрат, точный тест Фишера. Встроенные библиотеки python и модули NLTK для токенизации, удаления знаков препинания и стоп-слов, подсчета частотности и т.д.
9.	**Тема 9. Universal Dependencies и автоматический синтаксический анализ естественного языка.**	Проект Universal Dependencies, особенности его документации и теоретические основания. Формат файлов .conllu. Библиотека pyconll. Визуализация данных UD. Пайплайн обработки SpaCy nlp. Оболочка spacy_udpipe.
10.	**Тема 10. Обработка и визуализация результатов лингвистического анализа.**	Принципы визуализации лингвистического анализа. Библиотека matplotlib. Построение двух- и трехмерных графиков. Работа с аргументами функций matplotlib для визуального оформления данных. Работа с изображениями и картами.

4. Образовательные технологии

Для проведения учебных занятий по дисциплине используются различные образовательные технологии. Для организации учебного процесса может быть использовано электронное обучение и (или) дистанционные образовательные технологии.

5. Оценка планируемых результатов обучения

5.1 . Система оценивания

Текущий контроль

При оценивании активности участия в семинаре (максимальная оценка – 2 балла за каждый семинар, в сумме 22 балла) учитываются:

- степень вовлеченности в дискуссию или, альтернативно, практического применения изучаемых методов (1 балл);
- количество решенных заданий на семинаре (1 балл).

При оценивании домашних заданий (максимальная оценка в сумме за курс – 38 баллов) учитываются:

- способность прокомментировать и объяснить свой код (для заданий, требующих программирования) или ход решения;
- качество и читаемость кода (для заданий, требующих программирования);
- количество решенных заданий.

Промежуточная аттестация (экзамен)

При проведении промежуточной аттестации студент должен представить свой исследовательский проект, представляющий собой либо проект, написанный на Python, решающий исследовательскую задачу на материале текстового датасета с применением

изученных методов и последующим представлением результатов, либо проект по созданию собственного корпуса с обоснованием и выполнением исследовательской задачи на его материале и последующим представлением результатов.

При оценивании проекта учитываются на каждый из теоретических вопросов учитывается:
- качество сбора и предобработки данных (максимум – 10 баллов);
- аргументированность обоснования исследовательской задачи и полнота ее выполнения (максимум – 10 баллов);
- использование пройденных технологий и методов (максимум – 10 баллов);
- презентация материала (максимум – 10 баллов).

Полученный совокупный результат конвертируется в традиционную шкалу оценок и в шкалу оценок Европейской системы переноса и накопления кредитов (European Credit Transfer System; далее – ECTS) в соответствии с таблицей:

<table>
<tr><td>100-балльная шкала</td><td colspan="2">Традиционная шкала</td><td>Шкала ECTS</td></tr>
<tr><td>95 – 100</td><td rowspan="2">отлично</td><td rowspan="5">зачтено</td><td>A</td></tr>
<tr><td>83 – 94</td><td>B</td></tr>
<tr><td>68 – 82</td><td>хорошо</td><td>C</td></tr>
<tr><td>56 – 67</td><td rowspan="2">удовлетворительно</td><td>D</td></tr>
<tr><td>50 – 55</td><td>E</td></tr>
<tr><td>20 – 49</td><td rowspan="2">неудовлетворительно</td><td rowspan="2">не зачтено</td><td>FX</td></tr>
<tr><td>0 – 19</td><td>F</td></tr>
</table>

5.2 Критерии выставления оценки по дисциплине

Баллы/ Шкала ECTS	**Оценка по дисциплине**	**Критерии оценки результатов обучения по дисциплине**
100-83/ A,B	отлично	Выставляется обучающемуся, если он глубоко и прочно усвоил теоретический и практический материал, может продемонстрировать это на занятиях и в ходе промежуточной аттестации. Обучающийся исчерпывающе и логически стройно излагает учебный материал, умеет увязывать теорию с практикой, справляется с решением задач профессиональной направленности высокого уровня сложности, правильно обосновывает принятые решения. Свободно ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «высокий».
82-68/ C	хорошо	Выставляется обучающемуся, если он знает теоретический и практический материал, грамотно и по существу излагает его на занятиях и в ходе промежуточной аттестации, не допуская существенных неточностей. Обучающийся правильно применяет теоретические положения при решении практических задач профессиональной направленности разного уровня сложности, владеет необходимыми для этого навыками и приёмами. Достаточно хорошо ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «хороший».
67-50/ D,E	удовлетвори тельно	Выставляется обучающемуся, если он знает на базовом уровне теоретический и практический материал, допускает отдельные ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает определённые затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, владеет необходимыми для этого базовыми навыками и

Баллы/ Шкала ECTS	Оценка по дисциплине	Критерии оценки результатов обучения по дисциплине
		приёмами. Демонстрирует достаточный уровень знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «достаточный».
49-0/ F,FX	неудовлетворительно	Выставляется обучающемуся, если он не знает на базовом уровне теоретический и практический материал, допускает грубые ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает серьёзные затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, не владеет необходимыми для этого навыками и приёмами. Демонстрирует фрагментарные знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции на уровне «достаточный», закреплённые за дисциплиной, не сформированы.

5.3 Оценочные средства (материалы) для текущего контроля успеваемости, промежуточной аттестации обучающихся по дисциплине

Примеры заданий для аудиторной и домашней работы:

1. Возьмите любой достаточно большой текст (в текстовом файле!), желательно минимум рассказ, а лучше полноценный роман. Поделите его на слова и посчитайте для него статистику по количеству слов: сколько раз встретилось какое слово. Напишите функцию, которая будет отрисовывать на графике (столбцовая диаграмма подойдет) частоты для n самых частотных слов, при этом добавьте функции интерактивность и возможность выбирать n.

2. Пишем программу - помощник обнаружения тавтологии: ищем в тексте повторяющиеся слова, которые находятся на расстоянии от 2 до 10 слов друг от друга. В простом случае считаем, что пунктуации у нас нет, в сложном можете попробовать ее тоже учесть.

3. Попробуем себя в решении задачи определения темы текста. Будем считать, что два текста похожи по теме, если у них больше общих слов (только не предлогов с союзами), чем у других текстов. У нашей программы для определения темы будет несколько готовых текстов (достаточно больших!) с уже известной темой в базе: выберите тексты (и темы) самостоятельно, 5-6 будет достаточно. Что должна делать программа? При запуске вы ей сообщаете название нового файла с текстом, который нужно классифицировать, она его открывает, обрабатывает и сравнивает с текстами в своей базе. С которым из текстов оказалось больше всего общих слов, того и тема! Очевидно, вам понадобится какие-то слова из текстов отбрасывать (подумайте, каким образом это сделать - здесь на самом деле несколько вариантов концепций), а еще лемматизировать.

6. Учебно-методическое и информационное обеспечение дисциплины

6.1 . Список источников и литературы

Основной учебник

Jurafsky, Dan & James H. Martin. 2017. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 3nd edition draft. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Дополнительные учебники

1. Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. Handbook of Natural Language Processing. 2nd ed. Boca Raton: Chapman & Hall/CRC.
2. Авраменко А.П. Большие языковые модели в лингвистике и лингводидактике. 2 изд. М., 2024.
3. Компьютерная лингвистика и автоматическая обработка текстов : учебное пособие / Н.В. Лукашевич, А.А. Сорокин. – М. : МАКС Пресс, 2025. – 608 с.
4. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных. М., 2017.
5. Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). 2016. Прикладная и компьютерная лингвистика. М.: URSS.

6.2 . Перечень ресурсов информационно-телекоммуникационной сети «Интернет».

1) Перечень электронных ресурсов
2) Национальный корпус русского языка (НКРЯ): http://ruscorpora.ru/
3) Regular Expression Cheat Sheet https://www.cheatography.com/davechild/cheatsheets/regular expressions17
4) Universal Dependencies http://universaldependencies.org/
5) Syntactic treebanks https://en.wikipedia.org/wiki/Treebank#Syntactic_treebanks

7. Материально-техническое обеспечение дисциплины

Занятия по курсу можно проводить с максимальной эффективностью в компьютерном классе или аудитории с доступом в Интернет, проектором и экраном для презентаций. Необходимо также наличие доски или флипчарта, чтобы преподаватель мог разбирать примеры по ходу объяснения и записывать задания. Для самостоятельной работы студентам необходимо рабочее место, оборудованное персональным компьютером с доступом в Интернет, аудио- и видеоплеером (Windows Media Player, MPC, WinAmp, VLC и т.п.) а также офисными программами (Microsoft Office, OpenOffice, LibreOffice, Zoho Office и т.п.).

8. Обеспечение образовательного процесса для лиц с ограниченными возможностями здоровья и инвалидов

В ходе реализации дисциплины используются следующие дополнительные методы обучения, текущего контроля успеваемости и промежуточной аттестации обучающихся в зависимости от их индивидуальных особенностей:

●для слепых и слабовидящих: лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением; письменные задания выполняются на компьютере со специализированным программным обеспечением или могут быть заменены устным ответом; обеспечивается индивидуальное равномерное освещение не менее 300 люкс; для выполнения задания при необходимости предоставляется увеличивающее устройство; возможно также использование собственных увеличивающих устройств; письменные задания оформляются увеличенным шрифтом; экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.

●для глухих и слабослышащих: лекции оформляются в виде электронного документа, либо предоставляется звукоусиливающая аппаратура индивидуального пользования; письменные задания выполняются на компьютере в письменной форме; экзамен и зачёт проводятся в письменной форме на компьютере; возможно проведение в форме тестирования.

●для лиц с нарушениями опорно-двигательного аппарата: лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением; письменные задания выполняются на компьютере со специализированным программным обеспечением; экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.

При необходимости предусматривается увеличение времени для подготовки ответа.

Процедура проведения промежуточной аттестации для обучающихся устанавливается с учётом их индивидуальных психофизических особенностей. Промежуточная аттестация может проводиться в несколько этапов.

При проведении процедуры оценивания результатов обучения предусматривается использование технических средств, необходимых в связи с индивидуальными особенностями обучающихся. Эти средства могут быть предоставлены университетом, или могут использоваться собственные технические средства.

Проведение процедуры оценивания результатов обучения допускается с использованием дистанционных образовательных технологий.

Обеспечивается доступ к информационным и библиографическим ресурсам в сети Интернет для каждого обучающегося в формах, адаптированных к ограничениям их здоровья и восприятия информации:

●для слепых и слабовидящих: в печатной форме увеличенным шрифтом, в форме электронного документа, в форме аудиофайла.

●для глухих и слабослышащих: в печатной форме, в форме электронного документа.

●для обучающихся с нарушениями опорно-двигательного аппарата: в печатной форме, в форме электронного документа, в форме аудиофайла.

Учебные аудитории для всех видов контактной и самостоятельной работы, научная библиотека и иные помещения для обучения оснащены специальным оборудованием и учебными местами с техническими средствами обучения:

●для слепых и слабовидящих: устройством для сканирования и чтения с камерой SARA CE; дисплеем Брайля PAC Mate 20; принтером Брайля EmBraille ViewPlus;

●для глухих и слабослышащих: автоматизированным рабочим местом для людей с нарушением слуха и слабослышащих; акустический усилитель и колонки;

●для обучающихся с нарушениями опорно-двигательного аппарата: передвижными, регулируемыми эргономическими партами СИ-1; компьютерной техникой со специальным программным обеспечением.

9. Методические материалы

9.1 . Планы семинарских занятий

Семинар 1. Работа с онлайн-ресурсами по представлению исторических данных. (2 ч.).

Вопросы для обсуждения:

Корпусные ресурсы исторических данных. Цифровая работа с эго-документами. Платформы “Прожито”, “Пишу тебе”. Опыт проекта “Пушкин ⟨цифровой⟩”. Обработка исторических текстов. Платформа Transkribus и основы компьютерного зрения. Особенности работы с историческими подкорпусами НКРЯ. Оцифровка берестяных грамот.

Список литературы:

1. Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. Handbook of Natural Language Processing. 2nd ed. Boca Raton: Chapman & Hall/CRC.
2. Jurafsky, Dan & James H. Martin. 2017. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 3nd edition draft. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf
3. Авраменко А.П. Большие языковые модели в лингвистике и лингводидактике. 2 изд. М., 2024.
4. Компьютерная лингвистика и автоматическая обработка текстов : учебное пособие / Н.В. Лукашевич, А.А. Сорокин. – М. : МАКС Пресс, 2025. – 608 с.
5. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных. М., 2017.
6. Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). 2016. Прикладная и компьютерная лингвистика. М.: URSS.

Материально-техническое обеспечение занятия:

В ходе лабораторных работ студенты должны быть обеспечены компьютерами с лицензионным программным обеспечением и выходом в Интернет.

Тема 4. Работа с корпусными менеджерами.	Основной функционал корпусных менеджеров (SketchEngine, Voyant Tools, LancsBox, AntConc). Scraping с помощью корпусных менеджеров. Извлечение ключевых слов с помощью существующих электронных ресурсов. Составление списков стоп-слов. Создание облаков слов (word clouds). Частотные списки как основа для сравнения корпусов.
Тема 5. Основы дистрибутивной семантики и введение в обработку естественного языка (Natural Language Processing, NLP).	Дистрибутивная гипотеза, закон Ципфа и представление языковых единиц в многомерном пространстве (метод векторизации); принципы и сферы применения NLP; знакомство с основными библиотеками Python, используемыми для обработки естественного языка.
Тема 6. Основы компьютерной лингвистики и морфологический анализ.	Сегментация текста на токены и предложения. Неравнозначность токена и слова. Проблемы токенизации и деления на предложения в языках с различными системами графики. Токенизаторы, сентенайзеры, лемматизаторы и стеммеры.
Тема 7. Основы программирования на Python для обработки естественного языка.	Основы работы с регулярными выражениями. Модуль Python re. Scraping с помощью инструментов Python. NLTK как классическая библиотека для обработки естественного языка. Обработка морфологии русского языка. Natasha и razdel. Spacy. Сравнение работы методов изученных модулей на одинаковых текстах и анализ данных. Частеречная разметка (POS-tagging). Извлечение именованных сущностей.
Тема 8. Основы работы с открытыми библиотеками	Библиотеки numpy и pandas для обработки данных. Метод describe() в pandas. Среднее арифметическое, медиана,

обработки естественного языка и статистическая обработка.	стандартное отклонение, квантили. Хи-квадрат, точный тест Фишера. Встроенные библиотеки python и модули NLTK для токенизации, удаления знаков препинания и стоп-слов, подсчета частотности и т.д.
Тема 9. Universal Dependencies и автоматический синтаксический анализ естественного языка.	Проект Universal Dependencies, особенности его документации и теоретические основания. Формат файлов .conllu. Библиотека pyconll. Визуализация данных UD. Пайплайн обработки SpaCy nlp. Оболочка spacy_udpipe.
Тема 10. Обработка и визуализация результатов лингвистического анализа.	Принципы визуализации лингвистического анализа. Библиотека matplotlib. Построение двух- и трехмерных графиков. Работа с аргументами функций matplotlib для визуального оформления данных. Работа с изображениями и картами.

Семинар 2. Работа с корпусными менеджерами. (2 ч.)

Вопросы для обсуждения:

Основной функционал корпусных менеджеров (SketchEngine, Voyant Tools, LancsBox, AntConc). Scraping с помощью корпусных менеджеров. Извлечение ключевых слов с помощью существующих электронных ресурсов. Составление списков стоп-слов. Создание облаков слов (word clouds). Частотные списки как основа для сравнения корпусов.

Список литературы:

Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. Handbook of Natural Language Processing. 2nd ed. Boca Raton: Chapman & Hall/CRC.

Jurafsky, Dan & James H. Martin. 2017. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 3nd edition draft. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Авраменко А.П. Большие языковые модели в лингвистике и лингводидактике. 2 изд. М., 2024.

Компьютерная лингвистика и автоматическая обработка текстов : учебное пособие / Н.В. Лукашевич, А.А. Сорокин. – М. : МАКС Пресс, 2025. – 608 с.

Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных. М., 2017.

Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). 2016. Прикладная и компьютерная лингвистика. М.: URSS.

Материально-техническое обеспечение занятия:

В ходе лабораторных работ студенты должны быть обеспечены компьютерами с лицензионным программным обеспечением и выходом в Интернет.

Семинар 3. Основы дистрибутивной семантики и введение в обработку естественного языка (Natural Language Processing, NLP). (2 ч.)

Вопросы для обсуждения:

Дистрибутивная гипотеза, закон Ципфа и представление языковых единиц в многомерном пространстве (метод векторизации); принципы и сферы применения NLP; знакомство с основными библиотеками Python, используемыми для обработки естественного языка.

Список литературы:

Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. Handbook of Natural Language Processing. 2nd ed. Boca Raton: Chapman & Hall/CRC.

Jurafsky, Dan & James H. Martin. 2017. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 3nd edition draft. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Авраменко А.П. Большие языковые модели в лингвистике и лингводидактике. 2 изд. М., 2024.

Компьютерная лингвистика и автоматическая обработка текстов : учебное пособие / Н.В. Лукашевич, А.А. Сорокин. – М. : МАКС Пресс, 2025. – 608 с. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных. М., 2017.

Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). 2016. Прикладная и компьютерная лингвистика. М.: URSS.

Материально-техническое обеспечение занятия:

В ходе лабораторных работ студенты должны быть обеспечены компьютерами с лицензионным программным обеспечением и выходом в Интернет.

Семинар 4. Основы компьютерной лингвистики и морфологический анализ. (2 ч.)

Вопросы для обсуждения:

Сегментация текста на токены и предложения. Неравнозначность токена и слова. Проблемы токенизации и деления на предложения в языках с различными системами графики. Токенизаторы, сентенайзеры, лемматизаторы и стеммеры.

Список литературы:

Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. Handbook of Natural Language Processing. 2nd ed. Boca Raton: Chapman & Hall/CRC.

Jurafsky, Dan & James H. Martin. 2017. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 3nd edition draft. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Авраменко А.П. Большие языковые модели в лингвистике и лингводидактике. 2 изд. М., 2024.

Компьютерная лингвистика и автоматическая обработка текстов : учебное пособие / Н.В. Лукашевич, А.А. Сорокин. – М. : МАКС Пресс, 2025. – 608 с.

Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных. М., 2017.

Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). 2016. Прикладная и компьютерная лингвистика. М.: URSS.

Материально-техническое обеспечение занятия:

В ходе лабораторных работ студенты должны быть обеспечены компьютерами с лицензионным программным обеспечением и выходом в Интернет.

Семинары 5-7. Основы программирования на Python для обработки естественного языка. (6 ч.)

Вопросы для обсуждения:

Основы работы с регулярными выражениями. Модуль Python re. Scraping с помощью инструментов Python. NLTK как классическая библиотека для обработки естественного языка. Обработка морфологии русского языка. Natasha и razdel. Spacy. Сравнение работы методов

изученных модулей на одинаковых текстах и анализ данных. Частеречная разметка (POS-tagging). Извлечение именованных сущностей.

Список литературы:

Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. Handbook of Natural Language Processing. 2nd ed. Boca Raton: Chapman & Hall/CRC.

Jurafsky, Dan & James H. Martin. 2017. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 3nd edition draft. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Авраменко А.П. Большие языковые модели в лингвистике и лингводидактике. 2 изд. М., 2024.

Компьютерная лингвистика и автоматическая обработка текстов : учебное пособие / Н.В. Лукашевич, А.А. Сорокин. – М. : МАКС Пресс, 2025. – 608 с.

Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных. М., 2017.

Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). 2016. Прикладная и компьютерная лингвистика. М.: URSS.

Материально-техническое обеспечение занятия:

В ходе лабораторных работ студенты должны быть обеспечены компьютерами с лицензионным программным обеспечением и выходом в Интернет.

Семинары 8-9. Основы работы с открытыми библиотеками обработки естественного языка и статистическая обработка. (4 ч.)

Вопросы для обсуждения:

Библиотеки numpy и pandas для обработки данных. Метод describe() в pandas. Среднее арифметическое, медиана, стандартное отклонение, квантили. Хи-квадрат, точный тест Фишера. Встроенные библиотеки python и модули NLTK для токенизации, удаления знаков препинания и стоп-слов, подсчета частотности и т.д.

Список литературы:

Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. Handbook of Natural Language Processing. 2nd ed. Boca Raton: Chapman & Hall/CRC.

Jurafsky, Dan & James H. Martin. 2017. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 3nd edition draft. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Авраменко А.П. Большие языковые модели в лингвистике и лингводидактике. 2 изд. М., 2024.

Компьютерная лингвистика и автоматическая обработка текстов : учебное пособие / Н.В. Лукашевич, А.А. Сорокин. – М. : МАКС Пресс, 2025. – 608 с.

Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных. М., 2017.

Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). 2016. Прикладная и компьютерная лингвистика. М.: URSS.

Материально-техническое обеспечение занятия:

В ходе лабораторных работ студенты должны быть обеспечены компьютерами с лицензионным программным обеспечением и выходом в Интернет.

Семинар 10. Universal Dependencies и автоматический синтаксический анализ естественного языка. (2 ч.)

Вопросы для обсуждения:

Проект Universal Dependencies, особенности его документации и теоретические основания. Формат файлов .conllu. Библиотека pyconll. Визуализация данных UD. Пайплайн обработки SpaCy nlp. Оболочка spacy_udpipe.

Список литературы:

Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. Handbook of Natural Language Processing. 2nd ed. Boca Raton: Chapman & Hall/CRC.

Jurafsky, Dan & James H. Martin. 2017. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 3nd edition draft. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Авраменко А.П. Большие языковые модели в лингвистике и лингводидактике. 2 изд. М., 2024.

Компьютерная лингвистика и автоматическая обработка текстов : учебное пособие / Н.В. Лукашевич, А.А. Сорокин. – М. : МАКС Пресс, 2025. – 608 с.

Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных. М., 2017.

Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). 2016. Прикладная и компьютерная лингвистика. М.: URSS.

Материально-техническое обеспечение занятия:

В ходе лабораторных работ студенты должны быть обеспечены компьютерами с лицензионным программным обеспечением и выходом в Интернет.

Семинар 11. Обработка и визуализация результатов лингвистического анализа. (2 ч.)

Вопросы для обсуждения:

Принципы визуализации лингвистического анализа. Библиотека matplotlib. Построение двух- и трехмерных графиков. Работа с аргументами функций matplotlib для визуального оформления данных. Работа с изображениями и картами.

Список литературы:

Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. Handbook of Natural Language Processing. 2nd ed. Boca Raton: Chapman & Hall/CRC.

Jurafsky, Dan & James H. Martin. 2017. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 3nd edition draft. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Авраменко А.П. Большие языковые модели в лингвистике и лингводидактике. 2 изд. М., 2024.

Компьютерная лингвистика и автоматическая обработка текстов : учебное пособие / Н.В. Лукашевич, А.А. Сорокин. – М. : МАКС Пресс, 2025. – 608 с.

Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных. М., 2017.

Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). 2016. Прикладная и компьютерная лингвистика. М.: URSS.

Материально-техническое обеспечение занятия:

В ходе лабораторных работ студенты должны быть обеспечены компьютерами с лицензионным программным обеспечением и выходом в Интернет.

Приложение 1. Аннотация
рабочей программы дисциплины

АННОТАЦИЯ РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ

Цель дисциплины "Обработка естественного языка для историка": освоение студентами базовых понятий компьютерной лингвистики и автоматической обработки естественного языка с последующим развитием прикладных навыков и умений в области автоматизированной обработки научных текстов и работы с корпусными данными и электронными текстовыми ресурсами, в том числе применительно к историческим текстам и архивным данным.

Задачи дисциплины:

- освоение основных понятий и терминов по дисциплинам корпусной и компьютерной лингвистики;
- освоение основных библиотек языка программирования Python, используемых для обработки естественного языка;
- получение опыта работы с различными корпусами текстов и базами текстовых данных, в том числе относящихся к историческим материалам.

В результате освоения дисциплины обучающийся должен:

Знать: основные понятия и методы современной компьютерной лингвистики; основные понятия и методы современной корпусной лингвистики; базовые принципы лингвистической разметки; исторические и современные аспекты обработки естественного языка с использованием правиловых, статистических и нейросетевых моделей.

Уметь: совершать очистку текстовых данных для автоматической обработки и анализа; применять базовые функции и методы библиотек обработки естественного языка Python для решения практических задач по обработке и анализу языковых данных; применять базовые функции и методы библиотек визуализации данных и релевантных модулей библиотек обработки естественного языка Python для решения практических задач по визуализации статистических корпусных и языковых данных; применять базовый функционал основных корпусных менеджеров и инструментов автоматического распознавания текста.

Владеть: навыками отбора текстовых документов по критериям релевантным для решения исследовательской задачи; базовыми методами статистического анализа и инструментами их реализации; базовыми методами параметрической оценки корпуса, включая его взвешенность и репрезентативность. основным понятийным аппаратом и навыками реализации визуальных решений в области векторной семантики; навыками создания облаков слов на материале анализируемого корпуса; базовыми навыками извлечения именованных сущностей.