Минобрнауки России

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

(ФГБОУ ВО «ВГУ»)

УТВЕРЖДАЮ

Заведующий кафедрой

Сирота Александр Анатольевич

Кафедра технологий обработки и защиты информации

05.03.2025

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Б1.В.02 Компьютерная лингвистика

1. Код и наименование направления подготовки/специальности:

09.04.02 Информационные системы и технологии

2. Профиль подготовки/специализация:

Системы прикладного искусственного интеллекта

3. Квалификация (степень) выпускника:

Магистратура

4. Форма обучения:

Очная

5. Кафедра, отвечающая за реализацию дисциплины:

Кафедра технологий обработки и защиты информации

6. Составители программы:

Гаршина Вероника Викторовна, к.т.н., доцент

7. Рекомендована:

№ 5 05.03.2025

8. Учебный год:

2025-2026

9. Цели и задачи учебной дисциплины:

Изучить формальные, программно-реализуемые подходы к изучению структур и закономерностей естественных языков, ознакомится с основными прикладными практическими задачами компьютерной лингвистики

Основные задачи дисциплины:

изучить основные принципы и методов обработки естественного языка, получение навыков разработки и интеграции программных систем обработки естественного языка в программные продукты.

знакомство с принципами проектирования программных систем, ориентированных на обработку естественно-языковых текстов

10. Место учебной дисциплины в структуре ООП:

Дисциплина относится к вариативной части учебного плана Б1.В.

Для ее изучения требуются входные знания из курсов: математические методы в современных информационных технологиях, искусственный интеллект, программирование и теория алгоритмов.

11. Планируемые результаты обучения по дисциплине/модулю (знания, умения, навыки), соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями выпускников) и индикаторами их достижения:

Код и название компетенции	Код и название индикатора компетенции	Знания, умения, навыки
ПК-5 Способен разрабатывать и исследовать модели объектов профессиональной деятельности, предлагать и адаптировать методики решения научно-исследовательских задач, планировать и проводить исследования	ПК-5.1 Знает методы исследования предметной области, математические модели описания предметной области, методы оптимизации прикладных задач, современные методики тестирования ИС, методики описания и моделирования бизнес-процессов, средства моделирования бизнес-процессов	**Знает** базовые понятия, математические и алгоритмические методы обработки текстовой информации, применяемые для проектирования прикладных программных систем.
ПК-5 Способен разрабатывать и исследовать модели объектов профессиональной деятельности, предлагать и адаптировать методики решения научно-исследовательских задач, планировать и проводить исследования	ПК-5.2 Умеет проводить и организовывать проведение исследований, направленных на решение исследовательских задач в рамках реализации научного (научно-технического, инновационного) проекта с использованием моделей объектов профессиональной деятельности	**Умеет** проектировать, разрабатывать и интегрировать программные системы обработки естественного языка в программные продукты Применять программные библиотеки и инструменты для разработки систем, ориентированных на обработку текста и звучащей речи.
ПК-3 Способен определять варианты структур программного обеспечения информационных систем (программного средства), необходимые информационные потоки и исследовать варианты структур с использованием моделей различного уровня	ПК-3.1 Умеет проводить анализ внешнесистемных требований, возможностей их реализации, определяет концептуальный и функциональный облик системы (программного средства), выявление и анализ известных аналогов	**Знает** терминологию, базовые понятия, математические и алгоритмические методы обработки текстовой информации, этапы разработки лингвистически ориентированных программных продуктов, технологии представления и обработки текстовой информации, формальные модели представления естественного языка.

12. Объем дисциплины в зачетных единицах/час:

2/72

Форма промежуточной аттестации:

Зачет с оценкой

13. Трудоемкость по видам учебной работы

Вид учебной работы	Семестр 1	Семестр 2	Всего
Аудиторные занятия	0	48	48
Лекционные занятия		32	32
Практические занятия			0
Лабораторные занятия		16	16
Самостоятельная работа	0	24	24
Курсовая работа			0
Промежуточная аттестация	0	0	0
Часы на контроль			0
Всего	0	72	72

13.1. Содержание дисциплины

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
1	Задачи компьютерной лингвистики в изучении естественного языка (ЕЯ).	**Лекционные занятия по разделу** 1. Компьютерная лингвистика: задачи, направления исследований. Проблемы моделирования естественного языка. Лингвистические ресурсы, используемые для обработки текста и речи: дата сеты, текстовые и речевые корпуса, словари. Стандарты разметки лингвистических данных. Цикл обработки данных в data science. **Лабораторные занятия по разделу** не предусмотрены	Создан электронный онлайн - курс, размещены материалы к практикческим и лабораторным работам.

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
2	Алгоритмы лингвистического представления и анализа текста.	**Лекционные занятия по разделу** 2. Уровни текстового анализа: графематический, фонетический, морфологический, синтаксический, семантический. Основные задачи, их взаимосвязь. Лингвистический процессор - функциональная структура. Графематический анализ: выделение структурных элементов в тексте: границы предложений, слов, словари сокращений. Стемминг и лематизация. Стемминг: классификация алгоритмов стемминга. Алгоритмы стемминга для русского языка: Портера, Stemka, MyStem. Ошибки стемминга. Применение регулярных выражений. 3. Методы морфологического анализа, используемые в лингвистических процессорах. Морфологические словари. Программные библиотеки для работы с морфологией естественного языка. 4. Синтаксический анализ в компьютерной лингвистике. Формальные модели представления синтаксиса: деревья составляющих, грамматики составляющих, деревья зависимостей. Способы представления синтаксического разбора: синтаксическое дерево, размеченное предложение. Форматы синтаксической разметки в парсерах (на примере CONLLU). 5. КС - грамматики. Примеры синтаксических парсеров для английского и русского языка. Парсинг на основе Томита и Yargy парсеров. **Лабораторные занятия по разделу** **1.** Лабораторная работа № 1,2 Предобработка текста: нормализация и стандартизация, токенизация, **с**темминг и лематизация, очистка с выделение словаря стоп-слов, морфоанализаторы. Знакомство на основе библиотеки NLTK (английский язык) и Наташа (русский язык). 2. Лабораторная работа № 3 Извлечение фактов из текста на основе контекстно-свободных грамматик, реализуемых в Томита парсере, Yargy парсере.	Создан электронный онлайн - курс, размещены материалы к практикческим и лабораторным работам.

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
3	Статистические методы анализа текстов и исследования структуры ЕЯ текста	**Лекционные занятия по разделу** 6.Статистические методы анализа структур ЕЯ текста на морфологическом, синтаксическом, семантическом уровнях. Биграммы. Методы позиционных статистик в исследовании текстов. 1 и 2 законы Ципфа. Оценки частоты встречаемости слов, меры лексического разнообразия текстов, распределение встречаемости слов по тектсту. Нахождение ключевых стов и статистическая оценка тематики текста. Статистические характеристики авторства и стилистики. Проверка гипотез. **Лабораторные занятия по разделу** 1. Лабораторная работа № 4 Исследование текстового корпуса. Сравнительный статистический анализ текстовых корпусов.	Создан электронный онлайн - курс, размещены материалы к практикческим и лабораторным работам.

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
4	Векторная модель текста и классификация полнотекстовых документов	**Лекционные занятия по разделу** 7. Векторное представление слов. One-hot кодирование. Векторное представление текста на основе модели мешка слов (bag of words, BoW). Алгоритм построения BoW. Представление слов эмбеддингами. Размерность. Операции над векторами слов. 8. Модель текста Word2Vec , GloVe. Оценка семантического расстояния между объектами, представленными векторами. Контекстные модели на основе Word2Vec - CBOW (Continious Bag-of-words). Контекстные модели на основе Word2Vec - Skip-Grams. Алгоритм. 9. Классификация текстов на основе выделенных эмбеддингов моделью BoW и TF- IDF. Применения известных методов (NB, SVM, LogReg, RF) к классификации текстов в сочетании с выделением n-грамм. Метрики оценки качества классификации текстов. Accuracy, Precision, Recall, F1-score. Задачи рубрицирования текстов. **Лабораторные занятия по разделу** 1. Лабораторная работа № 5 Представления текста как BoW, TF-IDF, Co- occurrence matrix Представление очищенных данных в виде BoW, TF-IDF, обучение нескольких моделей классического машинного обучения и Co- occurrence matrix. 2. Лабораторная работа № 6 Word2Wec, GloVe. Обучение word2vec и glove, визуализация признаков. Получение эмбеддингов, обучение нескольких моделей машинного обучения.	Создан электронный онлайн - курс, размещены материалы к практикческим и лабораторным работам.

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
5	Архитектуры НС для работы с текстами	**Лекционные занятия по разделу** 10. Развитие архитектур НС для работы с текстами (ретроспектива). Transformer models. Обзор архитектуры трансформеров (Attention is all you need). 11. ELMo, BERT, XLNet. Краткий обзор моделей. Дообучение моделей на предметную область текстов, язык. 12. Обзор больших языковых моделей (Large Language Models): GPT, LLaMA, Mistral AI, DeepSeek, Gemini, YandexGPT, Grok. Галлюцинации, проблемы надежности выводов в LLM. Интеграция LLM с графами знаний. **Лабораторные занятия по разделу** Лабораторная работа № 7 Получение эмбеддингов для разных моделей, визуализация признаков. Обучение классических моделей машинного обучения.	Создан электронный онлайн - курс, размещены материалы к практикческим и лабораторным работам.

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
6	Прикладные задачи компьютерной лингвистики	**Лекционные занятия по разделу** 13. Извлечение фактов из текстов - распознавание именованных сущностей и ключевых слов - Named Entity Recognition(NER) и установление между ними взаимосвязей (отношений)..Типы именованных сущностей и способы извлечения их из текстов. Применение в системах обработки текстов. Проблемы разрешения омонимии, анафоры и кореферентности. Автоматические системы извлечения знаний из разнородных текстовых источников. Задачи структурирования текстовых данных. 14. Задачи генерации текстов. Методы генерации. Шаблонные системы генерации. Семантические, морфологические, синтаксические проблемы синтеза текстов. Автоматическое аннотирование и реферирование. Применение генеративно-состязательных сетей (Generative Adversarial Networks, GAN) для генерации текстов - SeqGAN. 15. Вопросно-ответные системы: индексирование в информационно-поисковых системах, архитектура, способы обработки запросов, генерация различных типов ответов. Генерация диалогов в вопросно-ответных системах - чат-боты. RAG (Retrieval Augmented Generation). 16. Проблема семантического анализа для автоматических систем обработки текстов. DataMining и TextMining. Семантическая модель текста, моделирование нарратива. **Лабораторные занятия по разделу** Лабораторная работа № 8 Реализация информационного поиска и вопросно-ответной системы на основе технологий langchain, huggingface models, RAG (Retrieval Augmented Generation).	Создан электронный онлайн - курс, размещены материалы к практикческим и лабораторным работам.

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела)	Лекционные занятия	Практические занятия	Лабораторные занятия	Самостоятельная работа	Всего
1	Задачи компьютерной лингвистики в изучении естественного языка (ЕЯ).	2		0	2	4
2	Алгоритмы лингвистического представления и анализа текста.	8		6	6	20
3	Статистические методы анализа текстов и исследования структуры ЕЯ текста	2		2	4	8
4	Векторная модель текста и классификация полнотекстовых документов	6		4	4	14
5	Архитектуры НС для работы с текстами	6		2	4	12
6	Прикладные задачи компьютерной лингвистики	8		2	4	14
		32	0	16	24	72

14. Методические указания для обучающихся по освоению дисциплины

1) При изучении дисциплины рекомендуется использовать следующие средства:

- рекомендуемую основную и дополнительную литературу;
- методические указания и пособия;
- контрольные задания для закрепления теоретического материала;
- электронные версии учебников и методических указаний для выполнения лабораторно практических работ (при необходимости материалы рассылаются по электронной почте).

2) Для максимального усвоения дисциплины рекомендуется проведение письменного опроса (тестирование, решение задач) студентов по материалам лекций и лабораторных работ. Подборка вопросов для тестирования осуществляется на основе изученного теоретического материала. Такой подход позволяет повысить мотивацию студентов при конспектировании лекционного материала.

3) При проведении лабораторных занятий обеспечивается максимальная степень соответствия с материалом лекционных занятий и осуществляется экспериментальная проверка методов, алгоритмов и технологий, применяемых в интеллектуальной обработке информации, излагаемых в рамках лекций.

4) При переходе на дистанционный режим обучения для создания электронных курсов, чтения лекций он-лайн и проведения лабораторно- практических занятий используется информационные ресурсы Образовательного портала "Электронный университет ВГУ (https:\\edu.vsu.ru), базирующегося на системе дистанционного обучения Moodle, развернутой в университете.

15. Перечень основной и дополнительной литературы, ресурсов интернет, необходимых

для освоения дисциплины

№ п/п	Источник
1	Корягин, С. В. Методы анализа естественно-языковых текстов : учебно-методическое пособие / С. В. Корягин, А. М. Русаков. — Москва : РТУ МИРЭА, 2023. — 86 с. — ISBN 978-5-7339-1737-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/331649 (дата обращения: 15.06.2025). — Режим доступа: для авториз. пользователей. Скопировать в буфер
2	Гаврилова, И. В. Основы искусственного интеллекта : учебное пособие / И. В. Гаврилова, О. Е. Масленникова. — 3-е изд., стер. — Москва : ФЛИНТА, 2019. — 283 с. — ISBN 978-5-9765-1602-1. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/115839 (дата обращения: 15.06.2025). — Режим доступа: для авториз. пользователей. Скопировать в буфер
3	Гаврилова, Т. А. Инженерия знаний. Модели и методы / Т. А. Гаврилова, Д. В. Кудрявцев, Д. И. Муромцев. — 6-е изд., стер. — Санкт-Петербург : Лань, 2023. — 324 с. — ISBN 978-5-507-46580-4. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/312842 (дата обращения: 15.06.2025). — Режим доступа: для авториз. пользователей.
4	Муромцев, Д. И. Структурирование, разметка и обогащение данных : учебно-методическое пособие / Д. И. Муромцев, И. А. Шилин, И. В. Исаев. — Санкт-Петербург : НИУ ИТМО, 2024. — 72 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/460262 (дата обращения: 15.06.2025). — Режим доступа: для авториз. пользователей.
5	Митяков, Е. С. Искусственный интеллект и машинное обучение : учебное пособие для вузов / Е. С. Митяков, А. Г. Шмелева, А. И. Ладынин. — Санкт-Петербург : Лань, 2025. — 252 с. — ISBN 978-5-507-51465-6. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/450827 (дата обращения: 15.06.2025). — Режим доступа: для авториз. пользователей.

б) дополнительная литература:

№ п/п	Источник
1	Ганегедара, Т. Обработка естественного языка с TensorFlow : руководство / Т. Ганегедара ; перевод с английского В. С. Яценкова. — Москва : ДМК Пресс, 2020. — 382 с. — ISBN 978-5-97060-756-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/140584 (дата обращения: 15.06.2025). — Режим доступа: для авториз. пользователей.

№ п/п	Источник
2	Бенгфорт Бенджамин, Билбро Ребекка, Охеда Тони Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. — СПб.: Питер, 2019.
3	Джонс, М. Т. Программирование искусственного интеллекта в приложениях / М. Т. Джонс. — Москва : ДМК Пресс, 2011. — 312 с. — ISBN 978-5-94074-746-8. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/1244 (дата обращения: 15.06.2025). — Режим доступа: для авториз. пользователей.
4	Теофили, Т. Глубокое обучение для поисковых систем : руководство / Т. Теофили ; перевод с английского Д. А. Беликова. — Москва : ДМК Пресс, 2020. — 318 с. — ISBN 978-5-97060-776-3. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/140574 (дата обращения: 15.06.2025). — Режим доступа: для авториз. пользователей.

в) информационные электронно-образовательные ресурсы:

№ п/п	Источник
1	ЭБС «Университетская библиотека online» . Контракт №3010-06/28-24 от 28.12.2024. Срок действия контракта: с даты его подписания до 10.02.2026 г.
2	Информационно-телекоммуникационная система «Контекстум» (Национальный цифровой ресурс «РУКОНТ») . Договор ДС-208 от 01.02.2021 пролонгирован до 01.02.2027.
3	Электронная библиотека ВГУ. Договор №ДС-208 от 01.02.2021 (с ООО «ЦКБ «БИБКОМ» и ООО «Агентство «Книга-Сервис» о создании Электронной библиотеки ВГУ). с 01.02.2021 по 31.01.2027.
4	ЭБС Лань (лицензионный договор №3010, (с 01/03/2024 по 28.02.2025). Пролонгация договора.

16. Перечень учебно-методического обеспечения для самостоятельной работы

№ п/п	Источник
1	Боярский К. К. Введение в компьютерную лингвистику. Учебное пособие. – СПб: НИУ ИТМО, 2013. – 72 с.
2	Информационные ресурсы Образовательного портала "Электронный университет ВГУ (https:\\edu.vsu.ru)

17. Образовательные технологии, используемые при реализации учебной дисциплины, включая дистанционные образовательные технологии (ДОТ), электронное обучение (ЭО),

смешанное обучение):

Для реализации учебного процесса используются:

ПО Microsoft в рамках подписки "Imagine/Azure Dev Tools for Teaching", договор №3010-16/96-18 от 29 декабря 2018г.

Персеры русского языка ТОМИТА, Yargy (Свободно-распространяемое ПО)

Язык программирования Python, IDE Pysharm.

Библиотеки Python - Наташа, NLTK и др для обработки естественного языка (Свободно-распространяемое ПО).

Библиотеки Python - для задач машинного обучения и представления данных (Свободно-распространяемое ПО).

Фреймворк LangChain. для разработки приложений LLM на Python и JavaScript. (Свободно-распространяемое ПО)

При проведении занятий в дистанционном режиме обучения используются технические и информационные ресурсы Образовательного портала "Электронный университет ВГУ (https:\\edu.vsu.ru), базирующегося на системе дистанционного обучения Moodle, развернутой в университете, а также другие доступные ресурсы сети Интернет.

18. Материально-техническое обеспечение дисциплины:

1. 394018, г. Воронеж, площадь Университетская, д. 1, корпус 1а, аудитория 292 Учебная аудитория: специализированная мебель, компьютер преподавателя Pentium- G3420-3,2ГГц, монитор с ЖК 17", мультимедийный проектор, экран. Система для видеоконференций Logitech ConferenceCam Group и ноутбук 15.6'' FHD Lenovo V155-15API ПО: ОС Windows v.7, 8, 10, Набор утилит (архиваторы, файл-менеджеры), LibreOffice v.5-7, Foxit PDF Reader/ Специализированная мебель: доска меловая 1 шт., столы 31 шт., стулья 64 шт.; выход в Интернет, доступ к фондам учебно-методической документации и электронным изданиям.

2. Компьютерный класс (один из №1-4 корп. 1а, ауд. № 382-385), Учебная аудитория: специализированная мебель, персональные компьютеры на базе i5-9600KF-3,7ГГц, мониторы ЖК 24'' (16 шт.), специализированная мебель: доска маркерная 1 шт., столы 16 шт., стулья 33 шт.; доступ к фондам учебно-методической документации и электронным изданиям, доступ к электронным библиотечным системам, выход в Интернет.ПО: ОС Windows v.7, 8, 10, Набор утилит (архиваторы, файл-менеджеры), LibreOffice v.5-7, Foxit PDF Reader.

19. Оценочные средства для проведения текущей и промежуточной аттестаций

Порядок оценки освоения обучающимися учебного материала определяется содержанием следующих разделов дисциплины:

№ п/п	Разделы дисциплины (модули)	Код компетенции	Код индикатора	Оценочные средства для текущей аттестации
1	Разделы 1-6	ПК-5	ПК-5.1	Устный опрос, собеседование. Практико-ориентированные задания по соответствующим разделам, лабораторные работы.

№ п/п	Разделы дисциплины (модули)	Код компетенции	Код индикатора	Оценочные средства для текущей аттестации
2	Разделы 1-6	ПК-5	ПК-5.2	Устный опрос, собеседование. Практико-ориентированные задания по соответствующим разделам, лабораторные работы.
3	Разделы 1-6	ПК-3	ПК-3.1	Устный опрос, собеседование. Практико-ориентированные задания по соответствующим разделам, лабораторные работы.

Промежуточная аттестация

Форма контроля - Зачет с оценкой

Оценочные средства для промежуточной аттестации

Перечень вопросов, лабораторные работы

20 Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

20.1 Текущий контроль успеваемости

Текущая аттестация проводится в соответствии с Положением о текущей аттестации обучающихся по программам высшего образования Воронежского государственного университета. Текущая аттестация проводится в формах устного опроса (индивидуальный опрос, фронтальная беседа) и письменных работ (контрольные, лабораторные работы). При оценивании могут использоваться количественные или качественные шкалы оценок.

Текущий контроль успеваемости по дисциплине осуществляется с помощью следующих оценочных средств:

- Устный опрос на лабораторных занятиях;
- Практико-ориенторованное задание;
- Лабораторные работы.

№ п/п	Наименование оценочного средства	Представление оценочного средства в фонде	Критерии оценки
1	Устный опрос на лабораторных занятиях	Вопросы по темам/разделам дисциплины	Правильный ответ – зачтено, неправильный или принципиально неточный ответ - не зачтено
2	Практико - ориентированное задание по разделам дисциплины	Теоретические вопросы по темам \ разделам дисциплины	Шкала оценивания соответствует приведенной ниже

3	Лабораторная работа	Содержит 8 лабораторных заданий, предусматривающие разработку систем обработки текстов на основе различных алгоритмов с использованием программных средств разработки.	При успешном выполнении работ в течение семестра фиксируется возможность оценивания только теоретической части дисциплины в ходе промежуточной аттестации (зачета), в противном случае проверка задания по лабораторным работам выносится на зачет.

Пример задания для выполнения лабораторной работы

Лабораторная работа

Работа с библиотекой Natural Language Toolkit (NLTK) для решения задач

Статистической обработки текстов (английский язык)

Изучение материала на основе документации по NLTK: http://www.nltk.org/ Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit Steven Bird, Ewan Klein, and Edward Loper, http://www.nltk.org/book_1ed/

Эксперименты проводить в https://colab.research.google.com/

Ход выполнения работы

Задание для выполнения:

1 – этап экспериментов (Анализ структуры отдельного корпуса (текста))

1. Выбрать корпус для экспериментов (из NLTK или загрузить свой)
2. Провести статистический анализ текста:
   - Длина текста, словарь текста, число различных слов в словаре, рассчитать параметр лексического разнообразия текста.
   - Определить число предложений, слов (провести токенизацию).
   - Убрать стоп слова (предлоги, союзы, управляющие слова) и построить частотный график встречаемости слов в тексте. Кумулятивный график частотного распределения слов.
   - Выделить частотные слова, относящиеся к одной леме (провести лематизацию)
   - На основе результатов лематизации вывести на печать слова, определяющие тематику текста (претенденты на ключевые слова). Выделить по частоте и длине.
   - Провести исследование тематической структуры текста (в каких частях текста о чем говориться) – исследовать частотное расположение слов в тексте - построить график дисперсии.
   - Распечатать ключевые слова (частотные слова), относящиеся к наиболее тематически важному разделу текста (определить по графику дисперсии). Для них построить частотный график встречаемости слов в тексте. Кумулятивный график частотного распределения слов.
   - Для ключевых слов найти им соответствующие биграммы и коллокации в тексте, оценить их частотность. Экспертным методом проверить соответствуют ли определенные словосочетания важными для уточнения тематики текста.

2 – этап экспериментов (сравнительный анализ нескольких корпусов)

1. Выбрать 2-3 корпуса для экспериментов (из NLTK или загрузить свои)
2. Провести статистический анализ этих корпусов по плану задания 1.

- Провести сравнение по статистическим параметрам: словарь текста, число различных слов в словаре, рассчитать параметр лексического разнообразия те (насколько) текста.
- Исследовать тематические структуры текстов (в каких частях текстов о чем говориться) исследовать частотное расположение слов в тексте - построить график дисперсии.

Отчетность по лабораторной работе

- По проведенным исследованиям сделать отчет с заключением о статистике, стилистике, тематике исследуемых текстов.
- Отчет оформляется в Word с описанием проведенного исследования с питоновскими скриптами и комментариями в блокноте, разработанном в https://colab.research.google.com/.
- Дайте название своему проекту с экспериментами (щелкнув на имя блокнота в правом верхнем углу, переименуйте). Все ваши эксперименты сохраняются на вашем гугл диске.
- В отчет нужно прикрепить ссылку на созданный блокнот с экспериментами – Поделиться (в правом верхнем углу colab.research).
- Отчет выложить в ответах на задание на мудле.

Приведённые ниже задания рекомендуется использовать при проведении диагностических работ для оценки остаточных знаний по дисциплине

Компетенция ПК-8

Вопросы с выбором 1-балл

1. Что относится к лингвистическим ресурсам для разработки программного обеспечения систем компьютерной лингвистики (множественный выбор):

1. Базы словосочетаний

2. Тезаурусы

3. Онтологии

4. Текстовые корпуса

5. Базы данных

6. Компьютерные словари

Ответ: 1,2,3,4,6.

2. Поставить правильное соответствие:

1. Графематический анализ - a) Выделение грамматической основы слова, определение частей речи, приведение слова к словарной форме.

2. Фонетический анализ - b) Выявление смысловых связей между словами и группами, извлечение семантических отношений.

3. Морфологический анализ - c) Выявление синтаксических связей между словами в предложении, построение синтаксической структуры предложения.

4. Синтаксический анализ - d) анализ звукового состава слова, позволяет вычленить в слове звуки и определить их характеристики.

5. Семантический анализ - e) Выделение из массива данных предложений и слов (токенов).

Ответы: 1-e, 2-d, 3-a, 4-c, 5-b. 3.

3. Поставить соответствие:

1. Токенизация a) Процесс использует словарь и морфологический анализ, чтобы в итоге привести слово к его канонической форме (словарной форме).

2. Стемминг b) Процесс разделения текста на предложения-компоненты или процесс разделения предложений на слова компоненты.

3. Лемматизация c) Процесс отсечения от слова окончаний и суффиксов, чтобы оставшаяся часть была одинаковой для всех грамматических форм слова.

Ответы: 1-b, 2-c, 3-a.

4. .Какие формальные модели используются в компьютерной лингвистике? (множественный выбор)

a.Контекстно-свободные грамматики

b.Марковские модели

c.Мультиагентные модели

d.Графовые модели

e.Автоматные модели

f.Динамические модели

Ответы: a, b, d, e.

5. Какие утверждения верны для задания Контекстно-свободной грамматики (множественный выбор):

a. Конечное множество A - алфавит. Его элементы называются символами. Конечные последовательности символов образуют слова в данном алфавите.

b. Алфавит разделяется на терминальные (“окончательные”) и нетерминальные (“промежуточные”) символы.

c. Среди нетерминальных символов может быть выбран один – начальный.

d. Правила грамматики имеют вид K->X, где K-нетерминальный символ, а X- слово, в которое могут входить и терминальные, и нетерминальные символы.

e. Правила грамматики имеют вид aKb->aXb, где K нетерминальный символ, окруженный как нетерминальными, так и терминальными символами a,b. X- слово, в которое могут входить и терминальные, и нетерминальные символы, окруженное нетерминальными и терминальными символами.

Ответы: a, b, c, d.

Компетенция ПК-15

Вопросы с коротким ответом 2-балла

1. Приведите название закона, отражающего эмпирическую закономерность распределения частоты слов естественного языка: “если все слова языка (или просто длинного текста) упорядочить по убыванию частоты их использования, то частота n-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру n.”

Ответ: Закон Ципфа.

Вопросы с развернутым ответом 3-балла

1. Опишите реализацию стеминга для русского языка на основе алгоритма Портера. Опишите шаги алгоритма.

Ответ:

Идея алгоритма: существует ограниченное количество словообразующих суффиксов, и стемминг слова происходит без использования каких-либо баз основ: только множество существующих суффиксов и вручную заданные правила.

Алгоритм состоит из пяти шагов.

На каждом шаге отсекается словообразующий суффикс и оставшаяся часть проверяется на соответствие правилам (например, для русских слов основа должна содержать не менее одной гласной). Если полученное слово удовлетворяет правилам, происходит переход на следующий шаг.

Если нет — алгоритм выбирает другой суффикс для отсечения.

На первом шаге отсекается максимальный формообразующий суффикс, на втором — буква «и», на третьем — словообразующий суффикс, на четвертом — суффиксы превосходных форм, «ь» и одна из двух «н».

20.2 Промежуточная аттестация

Промежуточная аттестация может включать в себя проверку теоретических вопросов, а также, при необходимости (в случае не выполнения в течение семестра), проверку выполнения установленного перечня лабораторных заданий, позволяющих оценить уровень полученных знаний и/или практическое (ие) задание(я), позволяющее (ие) оценить степень сформированности умений и навыков.

Для оценки теоретических знаний используется перечень контрольно-измерительных материалов. Каждый контрольно-измерительный материал для проведения промежуточной аттестации включает два задания - вопросов для контроля знаний, умений и владений в рамках оценки уровня сформированности компетенции. При оценивании используется количественная шкала. Критерии оценивания представлены в приведенной ниже таблице 1.

Для оценивания результатов обучения на экзамене используются следующие содержательные показатели (формулируется с учетом конкретных требований дисциплины)

- знание теоретических основ учебного материала, основных определений, понятий и используемой терминологии;
- владение навыками проведения компьютерного эксперимента, тестирования компьютерных алгоритмов обработки информации.
- владение навыками программирования и экспериментирования с компьютерными моделями алгоритмов обработки информации в рамках выполняемых лабораторных заданий;
- умение обосновывать свои суждения и профессиональную позицию по излагаемому вопросу;
- умение связывать теорию с практикой, иллюстрировать ответ примерами, в том числе, собственными, умение выявлять и анализировать основные закономерности, полученные, в том числе, в ходе выполнения лабораторно-практических заданий;
- умение проводить обоснование и представление основных теоретических и практических результатов (теорем, алгоритмов, методик) с использованием математических выкладок, блок-схем, структурных схем и стандартных описаний к ним;

Различные комбинации перечисленных показателей определяют ***критерии оценивания*** результатов обучения (сформированности компетенций) на зачете: высокий (углубленный) уровень сформированности компетенций; повышенный (продвинутый) уровень сформированности компетенций; пороговый (базовый) уровень сформированности компетенций.

Для оценивания результатов обучения на зачете с оценкой используется 4-балльная шкала: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно». Для оценивания результатов обучения на зачете используется – зачтено, не зачтено по результатам тестирования.

Соотношение показателей, критериев и шкалы оценивания результатов обучения на зачете с оценкой представлено в следующей таблице.

Критерии оценивания компетенций и шкала оценок на зачете

Критерии оценивания компетенций	Уровень сформированности компетенций	Шкала оценок
Обучающийся демонстрирует полное соответствие знаний, умений, навыков по приведенным критериям свободно оперирует понятийным аппаратом и приобретенными знаниями, умениями, применяет их при решении практических задач. Успешно выполнены лабораторные работы в соответствии с установленным перечнем.	Повышенный уровень	отлично
Ответ на контрольно-измерительный материал не полностью соответствует одному из перечисленных выше показателей, но обучающийся дает правильные ответы на дополнительные вопросы. При этом обучающийся демонстрирует соответствие знаний, умений, навыков приведенным в таблицах показателям, но допускает незначительные ошибки, неточности, испытывает затруднения при решении практических задач. Успешно выполнены лабораторные работы в соответствии с установленным перечнем.	Базовый уровень	хорошо
Обучающийся демонстрирует неполное соответствие знаний, умений, навыков приведенным в таблицах показателям, допускает значительные ошибки при решении практических задач. При этом ответ на контрольно- измерительный материал не соответствует любым двум из перечисленных показателей, обучающийся дает неполные ответы на дополнительные вопросы. Успешно выполнены лабораторные работы в соответствии с установленным перечнем.	Пороговый уровень	удовлетворительно
Ответ на контрольно-измерительный материал не соответствует любым трем из перечисленных показателей. Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки. Не выполнены лабораторные работы в соответствии с установленным перечнем.	Ниже порогового уровня	не зачтено

Пример контрольно- измерительного материала

УТВЕРЖДАЮ
Заведующий кафедрой технологий
обработки и защиты информации
_____А.А. Сирота
05.03.2025

Направление подготовки *09.04.02 Информационные системы и технологии*
программа *Системы прикладного искусственного интеллекта*
Дисциплина *Б.1.В.02 Компьютерная лингвистика*
Форма обучения *Очное*
Вид контроля *Зачет с оценкой*
Вид аттестации *Промежуточная*

Контрольно-измерительный материал № 1

1. Стемминг и лематизация. Стемминг: классификация алгоритмов стемминга. Алгоритмы стемминга для русского языка: Портера, Stemka, MyStem. Ошибки стемминга.

2. Автоматические системы извлечения знаний из разнородных текстовых источников. Задачи структурирования текстовых данных. Извлечение именованных сущностей и отношений между ними - подходы.

Преподаватель _____В.В.Гаршина

Примерный перечень вопросов к зачету

Компьютерная лингвистика – общие вопросы

1. Компьютерная лингвистика как междисциплинарная область. Основные задачи компьютерной лингвистики, направления исследований. Классификация прикладных систем в области компьютерной лингвистики. Проблемы моделирования естественного языка в компьютерной лингвистике. Лингвистические ресурсы, используемые для обработки текста и речи.

Лингвистические уровни анализа текстовых данных

2. Уровни текстового анализа: графематический, фонетический, морфологический, синтаксический, семантический. Основные задачи, их взаимосвязь. Графематический анализ: задачи, методы реализации (примеры), выделение структурных элементов в тексте, границы предложений, слов, словари сокращений.
3. Морфологический анализ. Задачи, методы реализации, примеры морфоанализаторов и инструментов разработки.
4. Стемминг и лематизация. Стемминг: классификация алгоритмов стемминга. Алгоритмы стемминга для русского языка: Портера, Stemka, MyStem. Ошибки стемминга. Лематизация: используемые методы, примеры для русского языка, инструменты разработки.
5. Вероятностно-статистические характеристики текста, его элементов. Их применение в задачах лингвистики
6. Синтаксический анализ в компьютерной лингвистике. Способы представления синтаксического разбора: синтаксическое дерево, размеченное предложение. Примеры синтаксических парсеров и инструменты разработки. Формальная модель представления синтаксиса: деревья составляющих. Грамматики составляющих.
7. Формальная модель представления синтаксиса: деревья зависимостей, КС - грамматики. Извлечение фактов на основе контекстно-свободных грамматик, реализуемых в Томита парсере, Yargy парсере.

Компьютерные модели и технологии обработки текстовых данных

8. В чем заключается предобработка текстового дата сета? Как провести проверку сбалансированности текстового корпуса? Как провести нормализацию, стемминг и лематизацию, очистку с выделение словаря стоп-слов?
9. Как реализовать поиск значимых слов, претендующих на роль ключевых в текстовом корпусе? Какие преобразования корпуса необходимо провести, какие гипотезы проверить, какими методами графического анализа можно воспользоваться? Законы Зипфа.
10. Векторное представление слов. One-hot кодирование. Векторное представление текста на основе модели мешка слов (bag of words, BoW). Алгоритм построения BoW.
11. Векторное представление слов эмбеддингами. Размерность. Операции над векторами слов.
12. Векторная модель текста Word2Vec. Оценка семантического расстояния между объектами, представленными векторами.
13. Контекстные модели на основе Word2Vec - CBOW (Continious Bag-of-words).
14. Контекстные модели на основе Word2Vec - Skip-Grams. Алгоритм.

15. Классификация текстов на основе выделенных эмбеддингов моделью BoW и применения известных методов классификации. Метрики оценки качества классификации текстов. Accuracy, Precision, Recall, F1-score.
16. Как выделить признаки из корпуса текстов с помощью TF- IDF. Какова эффективность

классификации разными моделями на основе метода TF-IDF с использованием биграмм и по отдельным словам.

Прикладные системы обработки и хранения текстовых данных

17. Проблемы автоматизации синтеза (генерации) текста. Этапы генерации (схема). Методы генерации. Шаблонные системы генерации. Генерация текстов на основе БД - простой отчет, связанный отчет. ЕЯ запросы к БД.
18. Алгоритмы синтеза текста для вербализации заданного содержания. Семантические, морфологические, синтаксические проблемы синтеза текстов.
19. Автоматическое аннотирование: архитектура построения систем, используемые методы, прикладное использование, примеры действующих систем.
20. Автоматическое реферирование: архитектура построения систем, используемые методы, прикладное использование, примеры действующих систем.
21. Информационный поиск: архитектура, модели представления документов, обработка поисковых запросов, извлечение документов.
22. Модели информационного поиска: инвертированная индексация, Булева и векторная модели. Метрики оценки близости документов. Оценка качества поиска: tf-idf, точность, полнота.
23. Вопросно-ответные системы: индексирование в информационно-поисковых системах, архитектура, способы обработки запросов, генерация различных типов ответов. Генерация диалогов в вопросно-ответных системах - чат-боты.
24. DataMining и TextMining. Извлечение фактов из текстов, установление взаимосвязей. Проблемы разрешения омонимии, анафоры и кореферентности.
25. Автоматические системы извлечения знаний из разнородных текстовых источников. Задачи структурирования текстовых данных. Извлечение именованных сущностей и отношений между ними. Подходы.
26. Автоматическое выделение именованных сущностей и ключевых слов. Типы именованных сущностей и способы извлечения из текстов. Применение в системах обработки текстов.Извлечение фактов на основе контекстно-свободных грамматик, реализуемых в Томита парсере, Yargy парсере. Примеры грамматик.