ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

**УТВЕРЖДЕНА**

Решением Ученого совета
АНО ВО «Центральный университет»
«07» марта 2024 г.
Протокол №1

**Рабочая программа дисциплины (модуля)**
**«Обработка естественного языка»**

**Направление подготовки:** *02.03.01 Математика и компьютерные науки*

**Направленность (профиль) подготовки:** *Математика и искусственный интеллект*

**Квалификация (степень) выпускника:** *бакалавр*

**Форма обучения:** *очная*

**Срок освоения программы:** *4 года*

**Год набора:** *2024*

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 0215B0C000D1B3FEB44A37EE631E0C27D4
Владелец: Ивашкевич Евгений Васильевич
Действителен: с 13.01.2026 до 13.04.2027

Электронный документ

**Москва**
**2024**

**Содержание**

1. **Краткая характеристика дисциплины (модуля)** ..... 3
2. **Перечень планируемых результатов обучения** ..... 5
3. **Тематический план** ..... 7
4. **Содержание дисциплины (модуля)** ..... 7
5. **Учебно-методическое обеспечение** ..... 8
6. **Материально-техническое обеспечение** ..... 8
7. **Методические и оценочные материалы** ..... 10

## 1. Краткая характеристика дисциплины (модуля)

Рабочая программа дисциплины (модуля) «Обработка естественного языка» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – бакалавриат по направлению 02.03.01 «Математика и компьютерные науки», профиль «Математика и искусственный интеллект», утвержденный приказом Министерства науки и высшего образования Российской Федерации № 807 от 23.08.2017 года.

Изучение дисциплины (модуля) «Обработка естественного языка» позволяет эффективно решать задачи автоматизации взаимодействия между человеком и машиной в реальных проектах. Обработка естественного языка является ключевой технологией в разработке интеллектуальных систем: чат-ботов, голосовых помощников, поисковых систем и аналитических платформ.

**Место дисциплины (модуля) в структуре образовательной программы**

Настоящая дисциплина (модуль) включена в учебный план по программе подготовки бакалавриата по направлению 02.03.01 «Математика и компьютерные науки», профиль «Математика и искусственный интеллект» и входит в вариативную часть Блока 1, формируемую участниками образовательных отношений, как дисциплина по выбору.

Дисциплина (модуль) изучается на 3 или 4 курсе в 5, 6, 7 семестре на выбор, доступна для прохождения при условии успешного завершения дисциплины (модуля) «Глубокое обучение».

**Цель изучения дисциплины (модуля):** формирование систематических знаний, практических умений и навыков в области анализа, обработки и генерации естественного языка с использованием современных методов машинного обучения и нейросетевых архитектур для решения прикладных задач в области искусственного интеллекта.

**Задачи изучения дисциплины (модуля):**

— формирование знаний принципов представления и предобработки текстовой информации в задачах обработки естественного языка;

— формирование знаний архитектур ключевых моделей обработки текста, включая RNN, LSTM, Transformer и современные языковые модели (BERT, GPT и др.);

— формирование знаний методов решения стандартных задач NLP: классификация текста, извлечение сущностей, машинный перевод, суммаризация, вопросно-ответные системы;

— формирование умений и навыков обучения, настройки и оценки качества моделей обработки естественного языка с использованием современных библиотек и фреймворков;

— формирование умений интегрировать языковые модели в сложные пайплайны и многомодальные системы, включая взаимодействие с визуальной информацией.

**В результате освоения дисциплины (модуля) обучающийся должен:**

***знать:***

— основные задачи области NLP и типы текстовых данных, включая однодокументные и многодокументные задачи;

— принципы токенизации, векторных представлений текста и языкового моделирования;

— архитектуру Трансформера, механизм внимания и различия между энкодерными, декодерными и encoder-decoder моделями;

— подходы к использованию промптинга, обучения с учителем и адаптации моделей под домен;

— принципы построения систем с дополненной генерацией (RAG), метрики качества и основные риски NLP-систем.

***уметь:***

— формулировать NLP-задачу, выбирать подходящий бейзлайн и корректные метрики оценки;

— строить и оценивать системы текстовой классификации и семантического поиска на основе эмбеддингов;

— извлекать структурированную информацию из текстов по заданной схеме с использованием современных моделей;

— дообучать энкодерные модели с учителем для решения прикладных задач;

— проектировать пайплайн с дополненной генерацией, анализировать ошибки и обосновывать выбор архитектуры.

***владеть:***

— стеком Python и библиотеками: pandas, scikit-learn, Hugging Face Transformers/Datasets, sentence-transformers, PyTorch;

— методами построения TF-IDF и эмбеддинговых бейзлайнов для поиска и классификации;

— практиками воспроизводимых экспериментов и автопроверяемых систем оценивания;

— техниками структурированного извлечения данных и проектирования прототипов RAG-систем;

— навыками анализа компромиссов между качеством, латентностью, стоимостью и безопасностью NLP-решений.

## 2. Перечень планируемых результатов обучения

Компетенции, формируемые в результате освоения дисциплины (модуля) при проведении учебных занятий в форме контактной работы обучающихся с педагогическими работниками Университета и в форме самостоятельной работы обучающихся:

| Компетенция | Содержание компетенции | Индикатор компетенции | Перечень планируемых результатов обучения по дисциплине (модулю) |
|---|---|---|---|
| УК-6. | Способен управлять своим временем, выстраивать и реализовывать траекторию саморазвития на основе принципов образования в течение всей жизни | УК-6.1. | Знает основные принципы самовоспитания и самообразования, профессионального и личностного развития, исходя из этапов карьерного роста и требований рынка труда |
| | | УК-6.2. | Умеет планировать свое рабочее время и время для саморазвития. формулировать цели личностного и профессионального развития и условия их достижения, исходя из тенденций развития области профессиональной деятельности, индивидуально-личностных особенностей |
| | | УК-6.3. | Имеет практический опыт получения дополнительного образования, изучения дополнительных образовательных программ |
| ОПК-1. | Способен находить, формулировать и решать актуальные и значимые проблемы прикладной и компьютерной математики | ОПК-1.1. | Знает основные методы и подходы к решению задач прикладной и компьютерной математики, включая алгоритмы, математическое моделирование и теорию оптимизации, а также современные инструменты и технологии, используемые в этой области |
| | | ОПК-1.2. | Умеет анализировать и формулировать математические задачи, применять соответствующие методы и алгоритмы для их решения, а также интерпретировать и представлять результаты в понятной и доступной форме |
| | | ОПК-1.3. | Имеет практический опыт работы над проектами или исследованиями в области прикладной и компьютерной математики, включая участие в конкурсах, олимпиадах или научных публикациях, где были решены актуальные и значимые задачи |
| ПК-1. | Способен определять общие формы и закономерности области машинного обучения | ПК-1.1. | Знает основные теоретические концепции и принципы, относящиеся к области машинного обучения, а также ключевые закономерности и модели, которые помогают в анализе и интерпретации данных |
| | | ПК-1.2. | Умеет проводить систематический анализ области разработки, выявлять и формулировать общие закономерности и тенденции, а также применять |

| | | | |
|---|---|---|---|
| | | | методы исследования для получения новых знаний и понимания |
| | | ПК-1.3. | Имеет практический опыт работы в области машинного обучения, включая участие в научных проектах, исследованиях или практических заданиях, где были выявлены и описаны общие формы и закономерности |
| ПК-2. | Способен решать типовые задачи профессиональной деятельности в области профессиональной деятельности, опираясь на информационную и библиографическую культуру, используя информационно-коммуникационные технологии и учитывая основные требования информационной безопасности | ПК-2.1. | Знает основы информационной и библиографической культуры, а также принципы информационной безопасности и применения информационно-коммуникационных технологий в профессиональной деятельности |
| | | ПК-2.2. | Умеет эффективно использовать информационно-коммуникационные технологии для решения стандартных задач профессиональной деятельности, учитывая требования информационной безопасности |
| | | ПК-2.3. | Имеет опыт работы с информационными ресурсами и технологиями в области профессиональной деятельности, включая соблюдение норм информационной безопасности |
| ПК-3. | Способен применять методы математического и алгоритмического моделирования для решения как теоретических, так и практических задач в рамках профессиональной деятельности | ПК-3.1. | Знает основные методы математического и алгоритмического моделирования, а также их применение для решения теоретических и прикладных задач |
| | | ПК-3.2. | Умеет применять методы математического и алгоритмического моделирования для анализа и решения различных задач в области математики и компьютерных наук |
| | | ПК-3.3. | Имеет опыт использования методов математического и алгоритмического моделирования при решении теоретических и прикладных задач в профессиональной деятельности |

## 3. Тематический план

| №п/п | Наименование раздела дисциплины (модуля) | Трудоемкость, академические часы | | | | ТКУ (текущий контроль успеваемости) |
|---|---|---|---|---|---|---|
| | | *Очная форма* | | | | |
| | | Аудиторная работа | | Контроль | Самостоятельная работа | |
| | | Лекции | Практические занятия | | | |
| 1 | Основы NLP | 10 | 10 | | 40 | Домашние задания, Тесты |
| 2 | Трансформеры и LLM | 8 | 8 | 2 | 34 | Домашние задания, Тесты |
| 3 | Прикладной NLP | 12 | 12 | 2 | 48 | Домашние задания, Тесты |
| | ***Экзамен*** | | | 4 | | |
| ***Итого:*** | | ***30*** | ***30*** | ***8*** | ***122*** | |
| ***Объем дисциплины (модуля) (в ак. ч.)*** | | ***190*** | | | | |
| ***Объем дисциплины (модуля) (в зач. ед.)*** | | ***5*** | | | | |

## 4. Содержание дисциплины (модуля)

| №п/п | Наименование раздела дисциплины (модуля) | Содержание дисциплины (модуля) по темам |
|---|---|---|
| 1 | Основы NLP | Введение в область NLP. Основы обработки текста. Текстовые представления и семантическая близость. Извлечение структурированной информации из текста. Нейросетевые подходы в области NLP |
| 2 | Трансформеры и LLM | Архитектура трансформера. Трансформеры в дискриминативных и генеративных задачах. Адаптация и пост-обучение больших языковых моделей |
| 3 | Прикладной NLP | Поиск и ранжирование. Основы методов дополненной генерации (RAG). Оценка качества в области NLP. Оптимизация больших языковых моделях NLP-системы. |

## 5. Учебно-методическое обеспечение

Университет располагает полным набором лицензионного и свободно распространяемого программного обеспечения, включая продукты отечественного производства.

Каждый студент в течение всего периода обучения получает индивидуальный неограниченный доступ к электронно-библиотечной системе и электронной информационно-образовательной среде университета. Эти системы предоставляют возможность доступа к ресурсам из любой точки, где есть подключение к сети Интернет, как на территории университета, так и за его пределами.

Студентам обеспечен удаленный доступ к современным профессиональным базам данных и информационным справочным системам.

***Основная литература:***

1. Бессмертный, И. А. Интеллектуальные системы : учебник и практикум для среднего профессионального образования / И. А. Бессмертный, А. Б. Нугуманова, А. В. Платонов. — 2-е изд. — Москва : Издательство Юрайт, 2026. — 250 с. — (Профессиональное образование). — ISBN 978-5-534-20730-9. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/586775.

2. Кревецкий, А. В. Основы технологий искусственного интеллекта : учебное пособие / А. В. Кревецкий, Н. И. Роженцова, Ю. А. Ипатов ; под общ. ред. А. В. Кревецкого. - Йошкар-Ола : Поволжский государственный технологический университет, 2023. - 272 с. - ISBN 978-5-8158-2358-7. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2133953.

***Дополнительная литература:***

1. Платонов, А. В. Машинное обучение : учебник для вузов / А. В. Платонов. — 2-е изд. — Москва : Издательство Юрайт, 2026. — 78 с. — (Высшее образование). — ISBN 978-5-534-22200-5. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/600893.

2. Гасанов, Э. Э. Интеллектуальные системы. Теория хранения и поиска информации : учебник для вузов / Э. Э. Гасанов, В. Б. Кудрявцев. — 2-е изд., испр. и доп. — Москва : Издательство Юрайт, 2026. — 271 с. — (Высшее образование). — ISBN 978-5-534-08684-3. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/584574.

## 6. Материально-техническое обеспечение

Университет располагает материально-технической базой, соответствующей действующим противопожарным правилам и нормам и обеспечивающей проведение всех видов дисциплинарной и междисциплинарной подготовки, практической и научно-исследовательской работ обучающихся, предусмотренных учебным планом.

Помещения, которые представляют собой учебные аудитории для проведения занятий лекционного типа, занятий семинарского (практического) типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы и помещения для хранения и профилактического обслуживания учебного оборудования. Помещения укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Изучение дисциплины (модуля) обеспечивается в учебных аудиториях, оснащенных:

— столами и стульями;

— компьютерной техникой;

— механическими калькуляторами;

— специализированным оборудованием, включая демонстрационное оборудование.

Помещения для самостоятельной работы обучающихся, в том числе приспособленные для использования инвалидами и лицами с ограниченными возможностями здоровья, оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду Университета.

Обучающимся предоставляется доступ (в том числе удаленный) к ресурсам информационно-телекоммуникационной сети «Интернет», электронным ресурсам (в том числе электронным библиотечным системам, современным профессиональным базам данных и информационным справочным системам):

| № | Наименование портала (издания, курса, документа) | Ссылка |
|---|---|---|
| 1. | Научная электронная библиотека elibrary.ru библиотека | https://elibrary.ru/defaultx.asp |
| 2. | База данных для IT-специалистов | https://habr.com |
| 3. | База данных ScienceDirect | https://www.sciencedirect.com |
| 4. | Официальный сайт Министерства науки и высшего образования Российской Федерации | https://minobrnauki.gov.ru/ |
| 5. | Федеральный портал «Российское образование" | https://www.edu.ru/ |
| 6. | Информационная система "Единое окно доступа к образовательным ресурсам" | http://window.edu.ru/ |
| 7. | Единая коллекция цифровых образовательных ресурсов | http://school-collection.edu.ru/ |
| 8. | Федеральный центр информационно - образовательных ресурсов | http://fcior.edu.ru/ |

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), в том числе комплект лицензионного программного обеспечения, современные профессиональные базы данных и информационные справочные системы:

| Наименование ПО | Производство | Лицензионное / свободно распространяемое |
|---|---|---|
| **Операционные системы:** | | |
| Microsoft Imagine (Windows Client, Server) | зарубежное | лицензионное |
| **Браузеры:** | | |
| Яндекс.Браузер | отечественное | свободно распространяемое |
| Google Chrome | зарубежное | свободно распространяемое |
| **Офисные приложения:** | | |
| Microsoft Imagine (Visio, OneNote) | зарубежное | лицензионное |
| TeXstudio | зарубежное | свободно распространяемое |
| Adobe Acrobat Reader | зарубежное | свободно распространяемое |
| **Программное обеспечение для планирования и учета времени:** | | |
| Toggle app | зарубежное | свободно распространяемое |
| **Системы управления проектами:** | | |
| Microsoft Imagine (Project) | зарубежное | лицензионное |
| **Системы управления базами данных:** | | |
| Microsoft Imagine (SQL Server) | зарубежное | лицензионное |
| **Системы резервного копирования (backup):** | | |
| Acronis Backup Advanced for HyperV | зарубежное | лицензионное |
| **Справочно-правовые системы:** | | |

| | | |
|---|---|---|
| КонсультантПлюс: справочно-правовая система | отечественное | лицензионное |
| **Средства антивирусной защиты:** | | |
| Kaspersky Endpoint Security для бизнеса Стандартный Russian Edition | отечественное | лицензионное |
| **Среды разработки:** | | |
| Visual Studio Code | зарубежное | свободно распространяемое |
| Bash (Unix shell) | зарубежное | свободно распространяемое |
| Anaconda | зарубежное | свободно распространяемое |
| Robotic Operating System | зарубежное | свободно распространяемое |
| CopelliaSim | зарубежное | свободно распространяемое |
| Google Colaboratory | зарубежное | свободно распространяемое |
| **Пакеты программных средств и библиотек:** | | |
| AutoPsy | зарубежное | свободно распространяемое |
| Interactive Disassembler (IDA) | зарубежное | свободно распространяемое |
| **Системы управления библиографической информацией:** | | |
| Zotero | зарубежное | свободно распространяемое |
| **Сервисы и службы:** | | |
| Bind | зарубежное | свободно распространяемое |
| Docker | зарубежное | свободно распространяемое |

## 7. Методические и оценочные материалы

### Методические указания для обучающихся по освоению дисциплины (модуля)

В процессе изучения дисциплины (модуля) «Обработка естественного языка» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, семинары и домашние задания, тесты, а также различные виды самостоятельной работы обучающихся по заданию преподавателя, направленные на развитие навыков профессиональной лексики, закрепление практических профессиональных компетенций, поощрение инициатив.

*Лекция* – систематическое, последовательное, монологическое изложение преподавателем учебного материала, как правило, теоретического характера.

В процессе лекций рекомендуется вести конспект лекций: кратко и схематично фиксировать основные идеи, выводы и обобщения лекции; выделять важные мысли, ключевые слова и термины. Необходимо отметить вопросы или материалы, которые вызывают затруднения, и попытаться найти ответы в рекомендованной литературе. Если разобраться в материале не удается, следует сформулировать вопрос и задать его преподавателю на консультации или во время семинарского (практического) занятия.

*Семинар* – это форма учебной деятельности, проводимая в учебном заведении под руководством преподавателя, где студенты активно участвуют в обсуждениях, практических заданиях и других формах взаимодействия.

Для успешной подготовки к семинару рекомендуется заранее ознакомиться с темой занятия и основными материалами, чтобы иметь возможность активно участвовать в обсуждении. Также полезно подготовить вопросы и идеи для обсуждения, что поможет глубже понять материал и продемонстрировать заинтересованность.

*Домашнее задание* – набор задач по темам недели.

При работе над домашними заданиями важно внимательно ознакомиться с требованиями и сроками выполнения. Рекомендуется разбивать задания на этапы, чтобы избежать перегрузки и лучше усвоить материал. Использовать различные источники информации, включая учебники и онлайн-ресурсы, для более глубокого понимания темы.

*Тест* – особая форма проверки знаний. Проводится после освоения одной или нескольких тем и свидетельствует о качестве понимания основных понятий изучаемого материала. Тестовые задания составлены к ключевым понятиям, основным разделам, важным терминологическим категориям изучаемой дисциплины (модуля).

Для подготовки к тесту необходимо знать терминологический аппарат дисциплины (модуля), понимать смысл научных категорий и уметь их использовать в профессиональной лексике. Владение понятийным аппаратом, включённым в тестовые задания, позволяет преподавателю быстро проверить уровень понимания студентами важных методологических категорий.

*Соревнование* – организованное мероприятие, в рамках которого участники соперничают друг с другом для достижения определенной цели, демонстрируя свои навыки, знания или способности в заданной области.

*Бонусные баллы* – это оценки, которые студенты могут получить за выполнение дополнительных заданий.

Формат бонусных баллов позволяет студентам улучшить общую оценку по дисциплине (модулю) и стимулирует углубленное изучение материала.

*Самостоятельная работа* – работа студентов, направленная на углубленное изучение отдельных тем и вопросов учебной дисциплины (модуля).

В процессе самостоятельной работы студенты взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи студента включают работу с конспектами лекций (обработка текста), повторное изучение учебных материалов планов и тезисов ответов, изучение дополнительных тем, выполнение учебно-исследовательских заданий и другое.

## Система оценивания результатов обучения по дисциплине (модулю)

**Критерии получения уровня и оценивания сформированности компетенций по дисциплине (модулю) «Обработка естественного языка»**

Оценивание уровня учебных достижений, обучающихся по дисциплине (модулю), осуществляется в виде текущего контроля успеваемости и промежуточной аттестации.

**Промежуточная аттестация** по дисциплине (модулю) осуществляется в форме ***экзамена***, при этом проводится оценка компетенций, сформированных по дисциплине.

Для оценивания текущего контроля успеваемости и промежуточной аттестации используется десятибалльная шкала оценивания, которая соотносится с традиционной пятибалльной шкалой следующим образом, если иное не указано в силлабусе дисциплины:

| Десятибалльная оценка | Пятибалльная оценка | Общая характеристика результата обучения по дисциплине (модулю) |
|---|---|---|
| 10 | Отлично | Студент полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину. Он самостоятельно и логически последовательно отвечает на все вопросы, акцентируя внимание на наиболее важном. Умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя ключевые моменты и устанавливая причинно-следственные связи. Четко формулирует ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные |
| 9 | Отлично | |
| 8 | Отлично | |

| Десятибалльная оценка | Пятибалльная оценка | Общая характеристика результата обучения по дисциплине (модулю) |
|---|---|---|
| | | задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами. |
| 7 | Хорошо | Студент обладает знаниями предмета почти в полном объеме рабочей программы и самостоятельно, логически последовательно и всесторонне отвечает на все вопросы, акцентируя внимание на наиболее значимых моментах. Он умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя его ключевые аспекты и устанавливая причинно-следственные связи. Формулирует свои ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные ситуационные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами. |
| 6 | Хорошо | |
| 5 | Удовлетворительно | Студент обладает базовыми знаниями по дисциплине (модулю), но испытывает трудности при самостоятельных ответах и использует неточные формулировки. В ходе ответов он допускает ошибки, касающиеся сути вопросов. Студент способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования. |
| 4 | Удовлетворительно | |
| 3 | Не сдан | Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы. |
| 2 | Не сдан | |
| 1 | Не сдан | |

Дисциплина (модуль) «Обработка естественного языка» оценивается следующим образом:

| Активность | Вес | Количество | Описание |
|---|---|---|---|
| Домашние задания | 40% | 5 | Набор задач по темам недели |
| Тест | 30% | 15 | Форма проверки знаний студентов через ответы на вопросы с выбором из нескольких вариантов ответов |
| Экзамен | 30% | 1 | Письменная или устная работа над заданием, направленным на проверку полученных знаний и навыков по дисциплине (модулю) |

В рамках изучения дисциплины (модуля) возможно получение бонусных баллов.

**Формула расчёта итоговой оценки по дисциплине (модулю) «Обработка естественного языка»:** «0,4 х среднее за домашние задания + 0,3 х среднее за тесты + 0,3 х Экзамен».

**Текущий контроль успеваемости обучающихся по дисциплине (модулю)**

**Примерные домашние задания**

**Домашнее задание 1.**

**Задание 1.**

1. Объясните, что такое цифровой текст и опишите основные этапы его предобработки: токенизация, нормализация, удаление стоп-слов.
2. Реализуйте на Python пайплайн предобработки текста для набора коротких отзывов (например, из IMDb или Yelp) с использованием библиотек nltk или spaCy.
3. Проведите анализ: сравните количество токенов до и после предобработки, визуализируйте самые частотные слова.

**Задание 2.**

1. Опишите, как текст может быть представлен в виде векторов: методы Bag-of-Words, TF-IDF, эмбеддинги (Word2Vec, GloVe).
2. Реализуйте вычисление сходства между текстами с помощью TF-IDF и косинусной меры. Протестируйте на парах предложений (похожие/разные).
3. Проведите эксперимент: сравните результаты семантического поиска с использованием TF-IDF и предобученных эмбеддингов (например, sentence-transformers). Прокомментируйте различия.

**Задание 3.**

1. Опишите задачу извлечения структурированной информации из текста (например, извлечение имён, дат, организаций).
2. Реализуйте извлечение именованных сущностей (NER) с использованием модели spaCy или transformers на наборе новостных текстов.
3. Оцените точность извлечения вручную на 10 примерах, проанализируйте типичные ошибки.

**Домашнее задание 2.**

**Задание 1.**

1. Объясните архитектуру Трансформера: механизм внимания, multi-head attention, позиционные кодировки.
2. Сравните энкодерные (BERT), декодерные (GPT) и encoder-decoder (T5) архитектуры: в каких задачах каждая из них применяется?
3. Визуализируйте веса внимания в предобученной модели (например, BERT) с помощью библиотеки BertViz на примере простого предложения.

**Задание 2.**

1. Изучите подходы к адаптации больших языковых моделей: fine-tuning, prompt tuning, LoRA.
2. Выберите модель (например, distilbert-base-uncased или bloomz) и дообучите её на задаче классификации текстов (например, SST-2) с использованием Hugging Face Transformers.
3. Сравните качество до и после дообучения, опишите используемый пайплайн и гиперпараметры.

**Задание 3.**

1. Реализуйте генерацию текста с помощью предобученной LLM (например, gpt2, Llama-3-8b через Hugging Face).

2. Проведите эксперимент с разными стратегиями генерации: greedy, beam search, sampling (с разными temperature, top_k, top_p).

3. Проанализируйте, как параметры влияют на разнообразие и качество генерации.

**Домашнее задание 3.**

**Задание 1.**

1. Опишите принцип работы систем семантического поиска: индексация, эмбеддинги, поиск по векторам.

2. Реализуйте простой поисковый движок на основе эмбеддингов (sentence-transformers) для набора документов (например, абстракты научных статей).

3. Оцените качество поиска: проверьте, насколько релевантны топ-3 результата для нескольких запросов.

**Задание 2.**

1. Объясните концепцию дополненной генерации (RAG): как внешние данные используются для улучшения генерации.

2. Реализуйте простую RAG-систему: используйте векторный поиск для извлечения контекста и LLM (например, gpt-3.5-turbo или Llama-3) для генерации ответа.

3. Протестируйте на вопросах, требующих знаний вне модели (например, "Какие меры были приняты на G7 в 2023 году?"). Сравните ответы с и без контекста.

**Задание 3.**

1. Опишите ключевые метрики оценки NLP-моделей: точность, F1, BLEU, ROUGE, perplexity, а также современные подходы (LLM-as-a-Judge).

2. Выберите задачу (например, суммаризация) и оцените два варианта выхода модели по нескольким метрикам.

3. Проведите анализ: сравните автоматические метрики с субъективной оценкой (по шкале 1–5). Сделайте вывод о надёжности метрик.

**Примерные задания для теста**
**Тест 1.**

**Вопрос 1.**
Что из перечисленного является основной целью токенизации?

A. Уменьшение размера текста
B. Преобразование текста в последовательность числовых векторов
C. Разбиение текста на слова, подслова или символы
D. Удаление стоп-слов и пунктуации

Ответ: C.

**Вопрос 2.**
Какой метод представления текста учитывает частоту слова в документе и его редкость в корпусе?

A. Bag-of-Words
B. One-Hot Encoding
C. TF-IDF
D. Word2Vec

Ответ: C.

**Вопрос 3.**
Что такое эмбеддинг слова?

A. Словарная статья из википедии
B. Числовой вектор, представляющий семантику слова

C. Индекс слова в словаре
D. Частота слова в тексте

Ответ: B.

**Вопрос 4.**

Какой из следующих подходов позволяет определить, что слова "автомобиль" и "машина" близки по смыслу?

A. TF-IDF
B. Эмбеддинги (например, Word2Vec или Sentence-BERT)
C. One-Hot Encoding
D. Подсчёт символов

Ответ: B.

**Вопрос 5.**

Какая задача заключается в извлечении имен, дат и организаций из текста?

A. Суммаризация
B. Классификация текста
C. Извлечение именованных сущностей (NER)
D. Машинный перевод

Ответ: C.

## Тест 2.

**Вопрос 1.**

Что является ключевой идеей механизма внимания в архитектуре Трансформера?

A. Учёт всех слов в последовательности при вычислении представления каждого слова
B. Использование рекуррентных связей между токенами
C. Классификация слов по частям речи
D. Преобразование текста в изображение

Ответ: A.

**Вопрос 2.**

Какая архитектура используется в моделях типа BERT?

A. Декодерная (decoder-only)
B. Энкодерная (encoder-only)
C. Encoder-decoder
D. CNN + RNN

Ответ: B.

**Вопрос 3.**

Какая модель является примером генеративной архитектуры с декодерной структурой?

A. T5
B. BERT
C. RoBERTa
D. GPT-3

Ответ: D.

**Вопрос 4.**

Что такое fine-tuning языковой модели?

A. Генерация текста с использованием промптов
B. Обучение модели с нуля на новом корпусе
C. Дообучение предобученной модели на задаче с размеченными данными
D. Извлечение эмбеддингов без обучения

Ответ: C.

**Вопрос 5.**

Какой метод позволяет адаптировать LLM с меньшим количеством параметров для обновления?

A. LoRA (Low-Rank Adaptation)
B. TF-IDF
C. PCA
D. Bag-of-Words

Ответ: A.

**Кейс от ООО «КНС ГРУПП»**

Студентам предлагается разработать локальный ИИ-сервис транскрибации и диаризации речи для EdTech-аналитики (Voice Detection).

**1. Описание задания:**

Создание высокопроизводительного локального бэкенд-сервиса на FastAPI для автоматической обработки аудиозаписей образовательных лекций и уроков. Пайплайн системы включает пакетную почанковую (chunk-by-chunk) обработку аудио, распознавание речи (ASR) с помощью библиотеки faster-whisper и выделение спикеров (Voice Diarization) с сохранением сквозного контекста голосов на протяжении всей записи.

*Проблема:* EdTech-платформы и методисты обрабатывают терабайты аудиозаписей занятий. Готовые облачные решения накладывают ограничения на конфиденциальность данных и неэффективны по стоимости. При этом стандартные open-source решения часто требуют загрузки всего аудио целиком в оперативную память, что приводит к падению сервисов на длинных лекциях. Почанковая обработка решает проблему утилизации ресурсов, но создаёт сложный инженерный вызов: как сопоставить и не потерять ID уникальных спикеров (учителя и разных студентов) между изолированными чанками аудио.

*Почему важно для бизнеса:* Создание легко масштабируемого, автономного (on-premise) сервиса аналитики аудио позволяет EdTech-компаниям бесшовно размечать структуру уроков, оценивать вовлечённость студентов и контролировать качество преподавания без передачи данных на сторонние API.

**2. Этапы работы:**

— Разработать архитектуру FastAPI сервиса, принимающего аудиофайл и выполняющего его пакетную нарезку на чанки.

— Интегрировать модуль транскрибации на базе faster-whisper для почанкового перевода речи в текст.

— Реализовать модуль Voice Detection (Diarization) для распознавания спикеров внутри чанков.

— Разработать механизм сквозного трекинга спикеров (сопоставление эмбеддингов голосов) между чанками, чтобы уникальные ID студентов сохранялись едиными для всей лекции.

— Создать демо-интерфейс для визуализации таймлайна урока с разметкой ролей.

**3. Используемые ресурсы:**

— реальные аудиозаписи лекций и практических занятий с качественной звуковой дорожкой;

— базовый скрипт-пример интеграции модуля транскрибации faster-whisper.

**4. Итоговый результат:**

Полностью локальный работающий бэкенд-сервис на FastAPI, который принимает на вход аудиозапись занятия, производит пакетную обработку по чанкам и возвращает структурированный таймлайн с разметкой спикеров (кто и когда говорил) и соответствующим текстом транскрипции. К сервису прилагается базовый веб-интерфейс (демо) для визуального отображения цветной ленты таймлайна.

**5. По итогам выполнения проекта студенты должны представить:**

Демо работающей системы — живой показ прототипа на тестовых аудиозаписях с демонстрацией кода, разбором пайплайна, бэкенд-архитектуры и результатов сквозного трекинга спикеров.

*Удовлетворительно*: Сервис на FastAPI стабильно принимает аудио, делит его на чанки и транскрибирует через faster-whisper без падения по памяти. Реализовано бинарное разделение речи на Преподавателя (Teacher) и Студентов (Students). Система точно отделяет голос лектора от общей массы реплик учащихся (все студенты могут иметь единый маркер ролей). Демо-интерфейс визуализирует этот таймлайн.

*Хорошо:* Сервис на FastAPI стабильно принимает аудио, делит его на чанки и транскрибирует через faster-whisper. Реализовано разделение речи на Преподавателя и индивидуальных Студентов (более 2 уникальных спикеров). Внедрена кластеризация голосовых профилей для различения учащихся, однако сквозное сопоставление идентичности спикера на протяжении всей записи реализовано частично (ID спикера может меняться между удаленными сегментами). Демо-интерфейс визуализирует таймлайн с несколькими участниками.

*Отлично* (Продвинутый Voice Detection): Реализован продвинутый пайплайн: система способна самостоятельно в динамическом режиме определять точное количество уникальных спикеров на записи (N студентов + 1 учитель). Внедрён алгоритм сквозного сопоставления векторов/профилей голосов (эмбеддингов) между изолированными чанками, благодаря чему ID конкретного студента не «бьётся» и остаётся единым на протяжении всего хронометража лекции. Сервис выдаёт чистый и точный итоговый лог-таймлайн.

**Задания для промежуточной аттестации по дисциплине (модулю)**

| № п/п | Задание | Ответ | Компетенция |
|---|---|---|---|
| 1 | Назовите основные компоненты в архитектуре модели BERT. | энкодера | УК-6 |
| 2 | Назовите метод предобучения, используемый в BERT для маскированного языкового моделирования. | MLM | УК-6 |
| 3 | Укажите задачи, используемые для предобучения BERT. | задачи | УК-6 |
| 4 | Определите основное преимущество архитектуры кодировщика трансформера по сравнению с RNN. | параллелизм | ПК-2 |
| 5 | Назовите механизм, отвечающий за внимание в трансформерах. | attention | ПК-3 |
| 6 | Укажите основные метрики для оценки качества ранжирования в информационном поиске. | метрики | ПК-2 |
| 7 | Определите способ доменной адаптации предобученных моделей для специфических задач. | fine-tuning | ОПК-1 |
| 8 | Назовите ограничение encoder-based моделей при работе с длинными текстами. | длина | ОПК-1 |
| 9 | Укажите основные шаги в процессе классификации текстов с использованием BERT. | шага | ОПК-1 |
| 10 | Назовите метод векторизации текста, используемый в encoder-based моделях. | embedding | ПК-1 |
| 11 | Определите количество параметров в больших языковых моделях, где применяется Mixture-of-Experts. | миллиарды | ПК-1 |
| 12 | Укажите тип параллельных вычислений, используемых при обучении LLM. | data | ПК-1 |
| 13 | Назовите метод оптимизации LLM для снижения размера модели. | квантизация | ПК-2 |

| | | | |
|---|---|---|---|
| 14 | Определите основные задачи, где fine-tuning LLM дает улучшения. | генерации | ПК-2 |
| 15 | Укажите механизм работы Efficient Transformers для оптимизации вычислений. | sparsity | ПК-2 |
| 16 | Назовите преимущество метода PEFT в дообучении LLM. | эффективность | ПК-3 |
| 17 | Определите модальности, поддерживаемые мультимодальными LLM. | текста | ПК-3 |
| 18 | Укажите тип генерации, используемой в Retrieval Augmented Generation. | augmented | ПК-3 |
| 19 | Назовите применение кодовых LLM в программировании. | генерация | ОПК-1 |
| 20 | Определите альтернативные архитектуры LLM помимо трансформеров. | RNN | ПК-1 |