

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО
Проректор по учебной работе

А.А. Воронов

	Рабочая программа дисциплины (модуля)
по дисциплине:	Анализ данных в научной литературе
по направлению:	Прикладные математика и физика
профиль подготовки:	Системная и синтетическая биология Физтех-школа Биологической и Медицинской Физики учебно-научный центр гуманитарных и социальных наук
курс:	1
квалификация:	магистр

Семестр, формы промежуточной аттестации: 2 (весенний) - Экзамен

Аудиторных часов: 30 всего, в том числе:

лекции: 30 час.

семинары: 0 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 75 час.

Подготовка к экзамену: 30 час.

Всего часов: 135, всего зач. ед.: 3

Количество контрольных работ, заданий: 2

Программу составил: А.А. Костин, канд. филос. наук, доцент

Программа обсуждена на заседании учебно-научного центра гуманитарных и социальных наук 04.04.2025

Аннотация

Большие массивы текстовых данных и общая информационная перегруженность современных областей знания затрудняют работу с текстовой базой. Принципиальным этот вопрос является для областей знания, в которых текст является основным источником знания. Однако и такие эмпирические области, как физика и биология, требуют тщательного исследования источниковедческой базы, объем которой часто превышает человеческие возможности.

Вместе с тем, достижения в таких областях, как обработка естественного языка и глубокое обучение нейронных сетей, делают возможной продуктивную работу с текстами любого объема и сложности. Это и предполагается продемонстрировать на материале курса.

На примере анализа научных публикаций - статей и монографий - рассматриваются основные методы обработки, анализа и визуализации данных: векторное представление текстовых данных, решение задач на их семантическое сходство, анализ и визуализация обработанных данных в виде сложных сетей, автоматизация обработки текстовых данных с помощью больших языковых моделей.

1. Цели и задачи

Цель дисциплины

Цель курса — формирование у слушателей практических навыков работы с научными текстовыми данными с применением передовых методов анализа данных и глубокого обучения

Задачи дисциплины

- Познакомить с базовыми принципами в области машинного обучения и обработки естественного языка.
- Обучить практическому применению инструментов для извлечения, обработки и анализа данных из научных текстов.
- Развить навыки адаптации и настройки больших языковых моделей для анализа содержания научных статей и монографий.
- Дать практические знания для реализации систем семантического поиска в научных исследованиях.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	УК-1.1 Анализирует проблемную ситуацию как систему, выявляя ее составляющие и связи между ними
	УК-1.2 Осуществляет поиск вариантов решения поставленной проблемной ситуации на основе доступных источников информации
	УК-1.3 Разрабатывает стратегию достижения поставленной цели как последовательность шагов, предвидя результат каждого из них и оценивая их влияние на внешнее окружение планируемой деятельности и на взаимоотношения участников этой деятельности
УК-5 Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия	УК-5.1 Способен выявлять специфику философских и научных традиций основных мировых культур
	УК-5.2 Способен определять теоретическое и практическое значение культурно-языкового фактора при взаимодействии различных философских и научных традиций

УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки	УК-6.1 Умеет решать задачи собственного личностного и профессионального развития, определять и реализовывать приоритеты совершенствования собственной деятельности
	УК-6.2 Оценивает свою деятельность, соотносит цели, способы и средства выполнения деятельности с её результатами

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные концепции и принципы анализа данных, машинного обучения и обработки естественного языка;
- архитектуру и функциональные особенности больших языковых моделей (GPT, BERT и др.);
- современные методы информационного поиска, тематического моделирования и семантической обработки текстов.

уметь:

- применять инструменты программирования (Python, библиотеки TensorFlow, PyTorch, Hugging Face) для реализации алгоритмов анализа данных и NLP;
- использовать методологию Retrieval Augmented Generation (RAG) для повышения качества извлечения информации из научных текстов;
- разрабатывать алгоритмы предобработки, векторизации и тематического анализа текстовых корпусов.

владеть:

- навыками обработки больших объёмов текстовых данных из научных источников;
- методиками построения конвейеров анализа данных для применения в задачах научной литературы;
- опытом настройки и оптимизации больших языковых моделей под конкретные задачи анализа научной литературы.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Введение в анализ данных в научной литературе: обзор инструментов и методов	2			5
2	Особенности научной литературы: структура статей и монографий, метаданные, специфика различных дисциплин	2			5
3	История обработки естественного языка: от чат-ботов на правилах до больших языковых моделей	2			5
4	Базы данных научных публикаций: PubMed, arXiv, Scopus, Web of Science, открытые репозитории	2			5
5	Обзор существующих инструментов анализа научной литературы (Semantic Scholar и другие)	2			5
6	Установка и настройка среды для работы с инструментами анализа данных и большими языковыми моделями (Python, Google Colab, Hugging Face)	2			5
7	Сбор и предобработка научных данных: извлечение текста из PDF/XML, удаление шума, аннотирование метаданных	2			5
8	Методы визуализации результатов анализа данных: инструменты визуализации (Matplotlib, Seaborn, Plotly), построение интерактивных отчетов и дашбордов	2			5
9	Кластеризация и тематическое моделирование: поиск трендов с помощью библиотеки BERTopic	2			5
10	Поиск связей между исследованиями: построение графов знаний (Knowledge Graphs) из текста и метаданных научных публикаций	2			5
11	Дообучение больших языковых моделей (SFT) под работу с научными текстами	2			5
12	Наборы данных (датасеты) с инструкциями для отдельных доменов знания	2			5

13	Генерация с дополненной выборкой (Retrieval Augmented Generation, RAG): повышение точности и интерпретируемости генеративных моделей	2			5
14	Мультимодальные модели для анализа текста и изображений (таблиц и графиков)	2			5
15	Разработка телеграм-бота для автоматизации анализа научной литературы	2			5
Итого часов		30			75
Подготовка к экзамену		30 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 2 (Весенний)

1. Введение в анализ данных в научной литературе: обзор инструментов и методов

Обзор предметной области:

- Роль анализа научных публикаций в формировании новых гипотез и поиске трендов.
- Примеры успешных кейсов автоматизированного анализа (мета-анализ, систематические обзоры).

Инструменты и платформы:

- Язык программирования: Python и базовые библиотеки (pandas, numpy, scikit-learn).
- Обзор специализированных библиотек для анализа текста: NLTK, SpaCy, Gensim.

Практикум:

- Обзор примеров кода для анализа небольшого корпуса статей.
- Демонстрация работы с API (например, Semantic Scholar API) для получения научных данных.

2. Особенности научной литературы: структура статей и монографий, метаданные, специфика различных дисциплин

Структура публикаций:

- Формат IMRaD (Introduction, Methods, Results, Discussion) для статей.
- Структура монографий, сборников, тезисов конференций.
- Методы автоматизированного «парсинга» структуры документа (PDF/XML).

Метаданные:

- Роль метаданных для индексации, поиска и анализа; стандарты (BibTeX, RIS).
- Использование DOI, ORCID, CrossRef и других систем идентификации.

Специфика дисциплин:

- Отличия в терминологии различных областей (биология, физика, компьютерные науки).
- Примеры применения специализированных NLP-моделей, таких как SciBERT для биологических текстов.

Практикум:

- Разбор статей и монографий с выделением структурных элементов.

- Автоматическая идентификация ключевых разделов текста с помощью моделей на основе Transformer.

3. История обработки естественного языка: от чат-ботов на правилах до больших языковых моделей

Ранний этап (правило-ориентированные системы):

- Примеры: ELIZA, системы на основе регулярных выражений.
- Ограничения и проблемы традиционных систем.

Статистические модели и традиционный ML:

- N-gram модели для машинного перевода.
- Применение методов кластеризации для тематического моделирования.

Переход к глубоким нейронным сетям:

- Введение рекуррентных нейронных сетей (RNN, LSTM).
- Эволюция Seq2Seq и механизм внимания.

Эра трансформеров и SOTA:

- Архитектура Transformer, самообучение и предварительное обучение (pre-training) с последующим fine-tuning.
- Обзор моделей: GPT, BERT

Практикум:

- Сравнительный анализ производительности моделей на небольшом корпусе данных.
- Использование библиотек Hugging Face для загрузки и тестирования моделей.

4. Базы данных научных публикаций: PubMed, arXiv, Scopus, Web of Science, открытые репозитории

Обзор баз данных:

- Краткие характеристики PubMed, arXiv, Scopus, Web of Science.
- Доступ к открытым репозиториям: CORE, Directory of Open Access Journals (DOAJ).

Способы доступа к данным:

- Использование API (например, API для PubMed и arXiv).
- Возможности сбора данных из Интернета (с соблюдением правил и лицензий).

Методы интеграции данных:

- Стандартизация форматов метаданных.
- Примеры объединения данных из нескольких источников для создания корпуса.

Практикум:

- Демонстрация работы с API одной из баз данных (например, получение и обработка результатов поиска из arXiv).
- Создание конвейера (pipeline) для автоматического обновления набора данных.

5. Обзор существующих инструментов анализа научной литературы (Semantic Scholar и другие)

Ключевые платформы:

- Semantic Scholar, Google Scholar
- Специфика каждого инструмента: алгоритмы ранжирования, поиск по цитируемости, извлечение ключевых идей.

Техническая архитектура и API:

- Обзор используемых технологий (NLP-пайплайны).
- Возможности интеграции с внешними сервисами для кастомизации аналитики.

Практикум:

- Работа с API Semantic Scholar: получение метаданных и анализ цитирования.
- Построение небольшого дашборда для визуализации результатов поиска и анализа трендов.

6. Установка и настройка среды для работы с инструментами анализа данных и большими языковыми моделями (Python, Google Colab, Hugging Face)

Установка Python и создание виртуального окружения:

- Использование conda/venv для управления зависимостями.
- Основы работы с Jupyter Notebook и Google Colab.

Подключение необходимых библиотек:

- Установка Hugging Face Transformers, Datasets, PyTorch.
- Дополнительные библиотеки: spaCy, scikit-learn, pandas, matplotlib.

Работа с Google Colab:

- Подключение к GPU для ускоренного обучения моделей.
- Примеры использования Colab.

Практикум:

- Пошаговое руководство по созданию и настройке окружения.
- Демонстрационная работа по запуску и тестированию базовых примеров моделей с использованием библиотеки Hugging Face.

7. Сбор и предобработка научных данных: извлечение текста из PDF/XML, удаление шума, аннотирование метаданных

Извлечение текста:

- Инструменты: PDFMiner для PDF; lxml и BeautifulSoup для XML.
- Преобразование неструктурированных данных в удобный формат (JSON, CSV).

Предобработка и очистка данных:

- Удаление «шума»: артефакты разметки, ненужные символы, OCR-ошибки.
- Токенизация, стемминг/лемматизация с использованием spaCy, NLTK и современных токенизаторов от Hugging Face.

Аннотирование метаданных:

- Автоматическое выделение элементов (авторы, год публикации, ключевые слова) с помощью моделей NER (spaCy, SciBERT).
- Стандартизация метаданных для ведения базы данных.

Практикум:

- Построение ETL-пайплайна для обработки научных публикаций.
- Лабораторная работа: извлечение текста и аннотирование метаданных с применением моделей на основе Transformer.

8. Методы визуализации результатов анализа данных: инструменты визуализации (Matplotlib, Seaborn, Plotly), построение интерактивных отчетов и дашбордов

Обзор библиотек визуализации:

- Matplotlib и Seaborn для базовых графиков (гистограммы, тепловые карты, boxplot).

- Plotly для создания интерактивных графиков и дашбордов (Dash, Streamlit).

Методы визуализации аналитических результатов:

- Построение визуальных представлений для кластеризации, тематического моделирования, графов знаний.
- Принципы рассказа данных (data storytelling) с акцентом на интерпретируемость аналитики.

Современные SOTA-подходы:

- Визуализация сложных машинных моделей (например, t-SNE, UMAP для векторных представлений).
- Интеграция аналитики в BI-платформы (Tableau, PowerBI).

Практикум:

- Практический проект по созданию интерактивного дашборда для анализа трендов в научной литературе.
- Задание: визуализировать результаты тематического моделирования с использованием Plotly и Dash.

9. Кластеризация и тематическое моделирование: поиск трендов с помощью библиотеки BERTopic

Теоретическая база:

- Обзор методов кластеризации (K-means, DBSCAN, HDBSCAN) и тематического моделирования (LDA, NMF).
- Преимущества использования контекстных эмбедингов (Sentence Transformers).

Библиотека BERTopic:

- Архитектура: сочетание Sentence Transformers, HDBSCAN и оптимизированного кластерного анализа.
- Пошаговое руководство по установке, настройке и интерпретации результатов.
- Использование BERTopic для анализа больших корпусов текстов, улучшение интерпретируемости тем с помощью визуальных инструментов.

Практикум:

- Практическое задание по кластеризации научных публикаций с помощью BERTopic.
- Анализ полученных тем и выявление ключевых трендов в выбранном корпусе данных.

10. Поиск связей между исследованиями: построение графов знаний (Knowledge Graphs) из текста и метаданных научных публикаций

Graphs) из текста и метаданных научных публикаций

Извлечение сущностей и отношений:

- Обзор моделей NER (например, SciBERT, SciSpacy) для автоматического выделения ключевых сущностей.
- Методы Relation Extraction (например, с использованием dependency parsing или transformer-based классификации).

Построение графов знаний:

- Технологии и базы данных для графов: Neo4j, ArangoDB.
- Построение графов с применением Graph Neural Networks (GNN) и Graph Attention Networks (GAT) для выявления ключевых связей.
- Интеграция полученных знаний в структуру графа с динамическим обновлением.
- Визуализация и интерактивный поиск по Knowledge Graph.

Практикум:

- Проект: извлечение сущностей и создание графа знаний из набора научных публикаций.
- Демонстрация визуализации графа с помощью специализированных библиотек (py2neo, NetworkX + Plotly).

11. Дообучение больших языковых моделей (SFT) под работу с научными текстами

Основы дообучения:

- Разбор этапов SFT: подготовка датасета, настройка гиперпараметров, выбор метрик качества.
- Проблема переобучения и методы регуляризации.

Инструменты:

- Использование PyTorch/TensorFlow и библиотеки Hugging Face Transformers для дообучения.
- Примеры SOTA-практик: адаптация моделей вроде BERT, RoBERTa или GPT для специализированных задач (например, извлечение информации из научных текстов).

Оценка качества:

- Примеры fine-tuning на корпусах научных публикаций (например, ACL Anthology, PubMed Central).
- Оценка результатов с помощью метрик F1-score, ROC-AUC и других.

Практикум:

- Лабораторная работа по дообучению выбранной языковой модели на корпусе научных текстов.
- Проведение экспериментов с различными гиперпараметрами и оценка интерпретируемости модели.

12. Наборы данных (датасеты) с инструкциями для отдельных доменов знания

Анализ существующих датасетов:

- Обзор доменных датасетов (биология, физика, компьютерные науки) на таких платформах, как Kaggle, UCI, Hugging Face Datasets.
- Критерии качества: баланс, полнота и валидность данных.

Методы аннотации и документирования:

- Составление инструкций для аннотаторов и стандартов оформления документации (data sheets for datasets).
- Автоматизированные методы валидации и очистки датасетов с использованием NLP-инструментов.
- Создание конвейера данных (пайплайна) для динамического обновления датасета с метаданными.

Практикум:

- Проект по сбору и аннотированию небольшого доменного датасета.
- Документирование процесса сборки набора данных с примерами инструкций.

13. Генерация с дополненной выборкой (Retrieval Augmented Generation, RAG): повышение точности и интерпретируемости генеративных моделей

Концепция RAG:

- Принцип работы: использование внешнего хранилища знаний для дополнения входных данных.

- Компоненты системы: модули (например, Dense Passage Retrieval, DPR) и генеративные модели.

SOTA-архитектуры и реализации:

- Пример: интеграция LLM с RAG-компонентами для повышения качества ответов.
- Обзор существующих решений на базе Hugging Face (RAG-интерфейсы, предобученные модели).

Практикум:

- Лабораторная работа по созданию RAG-системы с использованием открытых наборов данных и моделей.
- Сравнительный анализ с традиционными генеративными подходами.

14. Мультимодальные модели для анализа текста и изображений (таблиц и графиков)

Введение в мультимодальность:

- Обзор применения мультимодальных моделей: анализ научных публикаций, где текст сопровождается графами, диаграммами и таблицами.
- Примеры SOTA: CLIP (Contrastive Language–Image Pretraining), VisualBERT, LayoutLM для документов.

Методы интеграции разных данных:

- Предобработка изображений и таблиц для передачи в модели: OCR, сегментация изображений.
- Создание общего представления multimodal embeddings.

Применение в анализе научной литературы:

- Анализ инфографики, диаграмм и таблиц для извлечения ключевых фактов.
- Построение интерактивных систем для визуального анализа данных.

Практикум:

- Проект: построение мультимодальной модели для анализа публикации с комбинированными данными (текст + графики).
- Эксперимент с загрузкой предобученных мультимодальных моделей и их дообучением на специализированном датасете.

15. Разработка телеграм-бота для автоматизации анализа научной литературы

Архитектура телеграм-бота:

- Выбор инструментов: библиотека python-telegram-bot или aiogram для Python.
- Основы построения backend'a: вебхуки, polling, обработка сообщений.

Интеграция аналитических модулей:

- Подключение модулей анализа: извлечение данных из академических баз, NLP-анализ (например, кластеризация, RAG, Knowledge Graphs).
- Использование облачных сервисов для хостинга и масштабирования (Yandex Cloud и другие).

Пользовательский интерфейс и функциональность:

- Возможности: запрос к базе данных, визуализация результатов, прогнозирование трендов.
- Обработка естественного языка для распознавания пользовательских команд и поиска по запросам.
- Интеграция телеграм-бота с современными LLM, позволяющими вести диалог с пользователем на естественном языке (например, gpt4free API).
- Автоматическое обновление данных и вывод рекомендаций на основе анализа новейших публикаций.

Практикум:

- Разработка прототипа телеграм-бота с базовыми функциями: поиск публикаций, визуализация тематических кластеров, ответы на типичные вопросы.
- Финальный проект: интеграция бота с RAG-системой и Knowledge Graph для комплексного анализа научной литературы.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Учебная аудитория, оснащенная компьютером и мультимедийным оборудованием (проектор, звуковая система). Принтер и бумага для распечатки материалов к лекциям.

6.Перечень рекомендуемой литературы

Основная литература

Рекомендуемая литература для самостоятельного изучения:

Анализ данных : учебник для вузов / под редакцией В. С. Мхитаряна. — Москва : Издательство Юрайт, 2025. — 448 с. — (Высшее образование). — ISBN 978-5-534-19964-2. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://urait.ru/bcode/560311> (дата обращения: 10.04.2025).

Дополнительная литература

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

1. Валентин Малых. Краткая история NLP URL: <https://www.youtube.com/watch?v=Fnm-68F1kbw>

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

На лекционных занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

В ходе изучения дисциплины студент должен самостоятельно пополнять свои знания и изучить основополагающие работы в области изучаемой дисциплины.

Успешное освоение курса требует напряжённой работы студента непосредственно на лекции, а также самостоятельной работы для усвоения пройденного материала и решение задаваемых теоретических задач.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Прикладные математика и физика
профиль подготовки:	Системная и синтетическая биология Физтех-школа Биологической и Медицинской Физики учебно-научный центр гуманитарных и социальных наук
курс:	<u>1</u>
квалификация:	магистр
Семестр, формы промежуточной аттестации: 2 (весенний) - Экзамен	
Разработчик:	А.А. Костин, канд. филос. наук, доцент

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	УК-1.1 Анализирует проблемную ситуацию как систему, выявляя ее составляющие и связи между ними
	УК-1.2 Осуществляет поиск вариантов решения поставленной проблемной ситуации на основе доступных источников информации
	УК-1.3 Разрабатывает стратегию достижения поставленной цели как последовательность шагов, предвидя результат каждого из них и оценивая их влияние на внешнее окружение планируемой деятельности и на взаимоотношения участников этой деятельности
УК-5 Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия	УК-5.1 Способен выявлять специфику философских и научных традиций основных мировых культур
	УК-5.2 Способен определять теоретическое и практическое значение культурно-языкового фактора при взаимодействии различных философских и научных традиций
УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки	УК-6.1 Умеет решать задачи собственного личностного и профессионального развития, определять и реализовывать приоритеты совершенствования собственной деятельности
	УК-6.2 Оценивает свою деятельность, соотносит цели, способы и средства выполнения деятельности с её результатами

2. Показатели оценивания компетенций

В результате изучения дисциплины «Анализ данных в научной литературе» обучающийся должен:

знать:

- основные концепции и принципы анализа данных, машинного обучения и обработки естественного языка;
- архитектуру и функциональные особенности больших языковых моделей (GPT, BERT и др.);
- современные методы информационного поиска, тематического моделирования и семантической обработки текстов.

уметь:

- применять инструменты программирования (Python, библиотеки TensorFlow, PyTorch, Hugging Face) для реализации алгоритмов анализа данных и NLP;
- использовать методологию Retrieval Augmented Generation (RAG) для повышения качества извлечения информации из научных текстов;
- разрабатывать алгоритмы предобработки, векторизации и тематического анализа текстовых корпусов.

владеть:

- навыками обработки больших объёмов текстовых данных из научных источников;
- методиками построения конвейеров анализа данных для применения в задачах научной литературы;
- опытом настройки и оптимизации больших языковых моделей под конкретные задачи анализа научной литературы.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

Вопросы для текущего контроля:

1. Как анализ научных публикаций способствует формированию новых гипотез и выявлению трендов в науке?
2. Чем отличаются инструменты для работы с текстовыми данными от классических библиотек для анализа численных данных?
3. Что представляет собой формат IMRaD (Introduction, Methods, Results, Discussion) и какие ключевые разделы выделяются в научных статьях?
4. Как стандарты метаданных (BibTeX, RIS, DOI, ORCID, CrossRef) помогают в создании баз данных научной литературы?
5. Какие ограничения имели традиционные системы чат-ботов на правилах и как они были преодолены с появлением глубокого обучения?
6. Как осуществляется доступ к данным через API и какие возможности открытых репозиторий (CORE, DOAJ) применяются на практике?
7. Какие алгоритмы ранжирования и методы обработки цитируемости используются в таких системах, как Semantic Scholar и Google Scholar?
8. Какие библиотеки следует установить для начала работы с NLP и большими языковыми моделями?
9. Каковы основные этапы очистки данных от «шума», токенизации, стемминга и лемматизации?
10. Как применяется построение дашбордов для представления результатов тематического моделирования и кластеризации данных?

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

Вопросы для подготовки к экзамену:

1. Опишите поэтапно процесс установки Python, создания виртуального окружения и настройки Google Colab для работы с GPU.
2. Какие подходы применяются для удаления шумовых данных и ошибок OCR?
3. В чем разница между статическими графиками (histograms, boxplot, heatmap) и динамическими дашбордами?
4. Как подготовить датасет для дообучения языковой модели на научных текстах?
5. Опишите методы стандартизации и объединения данных из таких источников, как PubMed, arXiv, Scopus.
6. Каким образом происходит извлечение сущностей и отношений из текста?
7. Как осуществляется объединение текстовых данных с визуальной информацией (графы, диаграммы, таблицы)?
8. Какие ключевые элементы следует выделять при аннотировании научных публикаций?
9. Объясните концепцию RAG и принципы работы системы, объединяющей модуль retrieval (Dense Passage Retrieval, DPR) с генеративной моделью.
10. Опишите архитектуру телеграм-бота, его взаимодействие с базами данных и NLP-модулями.

Примеры билетов для проведения экзамена.

Экзаменационный билет №1. Анализ научной литературы и визуализация результатов

Вопрос 1. Извлечение и предобработка научных данных

- Опишите последовательность действий по извлечению текста из PDF и XML документов, укажите используемые инструменты.
- Как осуществляется предобработка извлечённого текста: удаление артефактов, токенизация, стемминг/лемматизация, аннотирование метаданных?

Вопрос 2. Построение интерактивных дашбордов

- Какие библиотеки используются для визуализации аналитических результатов?
- Приведите пример интеграции данных, полученных в результате тематического моделирования, в интерактивный дашборд.

Критерии оценивания

Оценка «отлично (10)» – заслуживает студент, обнаруживший всестороннее, систематическое и глубокое знание учебного программного материала, самостоятельно выполнивший все предусмотренные программой задания, глубоко усвоивший основную и дополнительную литературу, рекомендованную программой, активно работавший на занятиях, разбирающийся в основных научных концепциях по изучаемой дисциплине, проявивший творческие способности и научный подход в понимании и изложении учебного программного материала, чей ответ отличается богатством и точностью использованных терминов, а изложение материала в нем последовательно и логично;

Оценка «отлично (9)» – заслуживает студент, обнаруживший всестороннее, систематическое знание учебного программного материала, самостоятельно выполнивший все предусмотренные программой задания, глубоко усвоивший основную литературу и знакомый с дополнительной литературой, рекомендованной программой, активно работавший на занятиях, показавший систематический характер знаний по дисциплине, достаточный для дальнейшей учебы, а также способность к их самостоятельному пополнению, чей ответ отличается точностью использованных терминов, а изложение материала в нем последовательно и логично;

Оценка «отлично (8)» – заслуживает студент, обнаруживший полное знание учебно-программного материала, не допускающий в ответе существенных неточностей, самостоятельно выполнивший все предусмотренные программой задания, усвоивший основную литературу, рекомендованную программой, активно работавший на занятиях, показавший систематический характер знаний по дисциплине, достаточный для дальнейшей учебы, а также способность к их самостоятельному пополнению.

Оценка «хорошо (7)» – заслуживает студент, обнаруживший достаточно полное знание учебно-программного материала, не допускающий в ответе существенных неточностей, самостоятельно выполнивший все предусмотренные программой задания, усвоивший основную литературу, рекомендованную программой, активно работавший на занятиях, показавший систематический характер знаний по дисциплине, достаточный для дальнейшей учебы, а также способность к их самостоятельному пополнению;

Оценка «хорошо (6)» – заслуживает студент, обнаруживший достаточно полное знание учебно-программного материала, не допускающий в ответе существенных неточностей, самостоятельно выполнивший основные предусмотренные программой задания, усвоивший основную литературу, рекомендованную программой, отличавшийся достаточной активностью на занятиях, показавший систематический характер знаний по дисциплине, достаточный для дальнейшей учебы;

Оценка «хорошо (5)» – заслуживает студент, обнаруживший знание основного учебно-программного материала в объеме, необходимом для дальнейшей учебы и предстоящей работы по профессии, не отличавшийся активностью на занятиях, самостоятельно выполнивший основные предусмотренные программой задания, усвоивший основную литературу, рекомендованную программой, однако допустивший некоторые погрешности при их выполнении и в ответе на зачете, но обладающий необходимыми знаниями для самостоятельного устранения допущенных погрешностей;

Оценка «удовлетворительно (4)» – заслуживает студент, обнаруживший знание основного учебно-программного материала в объеме, необходимом для дальнейшей учебы и предстоящей работы по профессии, не отличавшийся активностью на занятиях, самостоятельно выполнивший основные предусмотренные программой задания, усвоивший основную литературу, рекомендованную программой, однако допустивший некоторые погрешности при их выполнении и в ответе на зачете, но обладающий необходимыми знаниями для устранения под руководством преподавателя допущенных погрешностей;

Оценка «удовлетворительно (3)» – заслуживает студент, обнаруживший знание основного учебно-программного материала в объеме, необходимом для дальнейшей учебы и предстоящей работы по профессии, не отличавшийся активностью на занятиях, самостоятельно выполнивший основные предусмотренные программой задания, однако допустивший погрешности при их выполнении и в ответе на зачете, но обладающий необходимыми знаниями для устранения под руководством преподавателя наиболее существенных погрешностей;

Оценка «неудовлетворительно (2)» – выставляется студенту, обнаружившему пробелы в знаниях или отсутствие знаний по значительной части основного учебно-программного материала, не выполнившему самостоятельно предусмотренные программой основные задания, допустившему принципиальные ошибки в выполнении предусмотренных программой заданий, допускающему существенные ошибки при ответе, и не способному продолжить обучение или приступить к профессиональной деятельности без дополнительных занятий по соответствующей дисциплине;

Оценка «неудовлетворительно (1)» – нет ответа (отказ от ответа) или представленный ответ полностью не соответствует существу содержащихся в задании вопросов.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Процедура оценки знаний, умений и навыков включает в себя прохождение экзамена в форме устного опроса по билетам. На подготовку к ответу студенту дается 30 минут. Вопросы включают проверку знаний теории и методов исследования, а также практические задания для анализа реального случая. Преподаватель также может задавать дополнительные вопросы сверх имеющихся в билете.