МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»
(РУТ (МИИТ)

Рабочая программа дисциплины (модуля),
как компонент образовательной программы
базового высшего образования
по направлению подготовки
09.03.02 Информационные системы и технологии,
утвержденной первым проректором РУТ (МИИТ)
Тимониным В.С.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Обработка естественного языка

Направление подготовки: 09.03.02 Информационные системы и технологии

Направленность (профиль): Технологии искусственного интеллекта в транспортных системах

Форма обучения: Очная

1. Общие сведения о дисциплине (модуле).

Дисциплина посвящена методам обработки естественного языка как направлению прикладного искусственного интеллекта. В ходе изучения рассматриваются подготовка текстовых корпусов, очистка и нормализация текста, токенизация, морфологический и синтаксический анализ, векторные представления слов и документов, классификация текстов, извлечение именованных сущностей, тематическое моделирование, архитектура Transformer, языковые модели, поиск по смысловой близости, RAG и оценка качества текстовых решений. На лабораторных занятиях обучающиеся последовательно создают воспроизводимое программное решение на Python для анализа русскоязычных текстов и оформляют техническую документацию.

Целью освоения дисциплины является формирование способности разрабатывать, проверять и документировать программные решения обработки естественного языка, обеспечивающие подготовку текстовых данных, извлечение признаков, классификацию, поиск сведений, работу с языковыми моделями и оценку качества результата в прикладных задачах искусственного интеллекта.

Для достижения поставленной цели в рамках дисциплины решается комплекс задач, направленных на формирование у обучающихся способности – анализировать прикладную задачу обработки текста, собирать и подготавливать текстовый корпус, выполнять нормализацию и морфологический анализ, строить числовые представления текстов, обучать модели классификации и извлечения сущностей, применять архитектуру Transformer и языковые модели, реализовывать поиск по смысловой близости, проектировать RAG, оценивать качество текстового решения и готовить техническую документацию.

2. Планируемые результаты обучения по дисциплине (модулю).

Перечень формируемых результатов освоения образовательной программы (компетенций) в результате обучения по дисциплине (модулю):

ПК-10 - Способен разрабатывать программные решения с использованием технологий компьютерного зрения, обработки естественного языка и мультиагентных систем.

Обучение по дисциплине (модулю) предполагает, что по его результатам обучающийся будет:

Знать:

- предметную область обработки естественного языка и ее связь с задачами искусственного интеллекта в транспортных и информационных системах
- виды текстовых данных, включая документы, обращения пользователей, сообщения, технические описания, журналы событий и нормативные тексты
- принципы формирования текстового корпуса, включая источник, разметку, метаданные, качество данных, баланс классов и правовые ограничения использования текста
- методы предварительной обработки текста, включая очистку, нормализацию регистра, удаление служебных символов, разбиение на предложения и токены
- морфологические свойства русского языка, включая лемматизацию, часть речи, грамматические признаки и неоднозначность словоформ
- основы синтаксического анализа текста, включая зависимости между словами, словосочетания, границы сущностей и связь синтаксиса с извлечением сведений
- способы представления текста в числовом виде, включая мешок слов, частотные признаки, TF-IDF и n-граммы
- принципы векторного представления слов, предложений и документов, включая смысловую близость, размерность пространства и ограничения интерпретации
- методы классификации текстов, включая постановку задачи, признаки, целевую переменную, обучение с учителем и контроль качества модели
- методы извлечения именованных сущностей, включая разметку персон, организаций, мест, дат, номеров, маршрутов и технических объектов
- методы тематического моделирования и кластеризации текстов, включая поиск скрытых тем, группировку документов и интерпретацию тематических признаков
- устройство архитектуры Transformer, включая механизм внимания, позиционное представление, кодировщик, декодировщик и предобучение языковой модели
- принципы применения предобученных языковых моделей для классификации, извлечения сущностей, поиска, суммаризации и ответа на вопросы
- основы дообучения языковой модели на прикладном наборе данных, включая подготовку выборок, функцию потерь, параметры обучения и контроль переобучения

- принципы поиска по смысловой близости, включая разбиение документов, векторный индекс, метаданные, ранжирование и проверку найденных фрагментов

- принципы RAG, включая извлечение релевантного контекста, передачу фрагментов в языковую модель, указание источников и отказ при недостаточности данных

- методы оценки качества решений обработки текста, включая точность, полноту, F-меру, матрицу ошибок, ручную проверку и анализ типовых ошибок

- требования к технической документации текстового решения, включая описание корпуса, обработки, моделей, параметров, ограничений, метрик и результатов проверки

Уметь:

- уметь формализовать задачу обработки текста при помощи описания источников, целевых сущностей, классов и критериев качества в условиях прикладного сценария искусственного интеллекта

- уметь формировать текстовый корпус при помощи Python и Pandas в условиях неоднородных источников, метаданных, дублей и неполной разметки

- уметь выполнять очистку и нормализацию текста при помощи регулярных выражений, Natasha, spaCy или NLTK в условиях русскоязычных документов разного качества

- уметь выполнять морфологический анализ при помощи pymorphy3 или Natasha в условиях лемматизации, определения частей речи и устранения служебных слов

- уметь строить частотные признаки при помощи scikit-learn в условиях классификации коротких и средних текстов

- уметь обучать модель классификации текста при помощи scikit-learn или CatBoost в условиях контроля качества на проверочной выборке

- уметь извлекать именованные сущности при помощи Natasha, DeepPavlov или модели Transformer в условиях поиска персон, организаций, мест, дат и технических объектов

- уметь строить векторные представления документов при помощи sentence-transformers в условиях поиска смысловой близости и группировки текстов

- уметь реализовывать тематический анализ при помощи scikit-learn или BERTopic в условиях выявления групп обращений, документов или сообщений

- уметь применять предобученную модель Transformer при помощи PyTorch и Hugging Face Transformers в условиях классификации, извлечения сущностей или ответа на вопрос

- уметь проектировать RAG при помощи LangChain, Qdrant или FAISS в условиях поиска фрагментов, передачи контекста и проверки источников

- уметь оценивать качество текстовой модели при помощи метрик точности, полноты, F-меры и анализа ошибок в условиях несбалансированных классов и неоднозначных текстов

- уметь готовить техническую документацию по решению обработки естественного языка при помощи описания корпуса, обработки, моделей, метрик, ограничений и результатов проверки

Владеть:

- навыком подготовки и структурирования русскоязычного текстового корпуса средствами Python и Pandas

- навыком очистки, нормализации, токенизации и морфологического анализа текста

- навыком построения частотных и векторных представлений текстовых документов

- навыком обучения и проверки моделей классификации текстов

- навыком извлечения именованных сущностей и анализа ошибок разметки

- навыком применения предобученных языковых моделей и архитектуры Transformer

- навыком реализации поиска по смысловой близости и RAG для ответа с опорой на источники

- навыком подготовки технической документации по программному решению обработки естественного языка

3. Объем дисциплины (модуля).

3.1. Общая трудоемкость дисциплины (модуля).

Общая трудоемкость дисциплины (модуля) составляет 3 з.е. (108 академических часа(ов).

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

Тип учебных занятий	Количество часов	
	Всего	Семестр №7
Контактная работа при проведении учебных занятий (всего):	32	32
В том числе:		
Занятия лекционного типа	16	16
Занятия семинарского типа	16	16

3.3. Объем дисциплины (модуля) в форме самостоятельной работы обучающихся, а также в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении промежуточной аттестации составляет 76 академических часа (ов).

3.4. При обучении по индивидуальному учебному плану, в том числе при ускоренном обучении, объем дисциплины (модуля) может быть реализован полностью в форме самостоятельной работы обучающихся, а также в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении промежуточной аттестации.

4. Содержание дисциплины (модуля).

4.1. Занятия лекционного типа.

№ п/п	Тематика лекционных занятий / краткое содержание
1	Предметная область обработки естественного языка Рассматриваемые вопросы: - прикладные задачи анализа документов, обращений пользователей, сообщений и технических текстов; - особенности русского языка как объекта программной обработки; - структура решения от текстового корпуса до проверенной модели.
2	Текстовый корпус и предварительная обработка текста Рассматриваемые вопросы: - источники текстовых данных, метаданные, разметка, дубли и правовые ограничения; - очистка, нормализация, разбиение на предложения и токены; - контроль качества корпуса и предотвращение утечки данных между выборками.
3	Лингвистический анализ русского текста Рассматриваемые вопросы: - лемматизация, части речи, грамматические признаки и неоднозначность словоформ; - синтаксические зависимости, словосочетания и границы сущностей; - применение Natasha, pymorphy3, spaCy и NLTK в задачах анализа текста.
4	Числовые представления текста и классификация Рассматриваемые вопросы: - мешок слов, n-граммы, частотные признаки и TF-IDF;

№ п/п	Тематика лекционных занятий / краткое содержание
	- постановка задачи классификации текстов и выбор целевой переменной; - метрики качества классификации и анализ ошибок модели.
5	Извлечение сведений из текста Рассматриваемые вопросы: - именованные сущности, даты, номера, маршруты, организации и технические объекты; - правила, словари и обучаемые модели извлечения сущностей; - тематическое моделирование, кластеризация и интерпретация групп документов.
6	Архитектура Transformer и языковые модели Рассматриваемые вопросы: - механизм внимания, позиционное представление, кодировщик и декодировщик; - предобучение, дообучение и применение языковой модели к прикладной задаче; - ограничения языковых моделей, вычислительные требования и контроль переобучения.
7	Смысловой поиск и RAG Рассматриваемые вопросы: - векторные представления предложений и документов, индексы сходства и метаданные; - разбиение документов, извлечение релевантных фрагментов и ранжирование результатов; - ответ с опорой на источники, проверка фрагментов и отказ при недостаточности данных.
8	Оценка качества и документирование текстового решения Рассматриваемые вопросы: - точность, полнота, F-мера, матрица ошибок и ручная проверка спорных случаев; - анализ смещения данных, неоднозначности текста и нестабильных классов; - состав технической документации по корпусу, обработке, моделям, ограничениям и результатам проверки.

4.2. Занятия семинарского типа.

Лабораторные работы

№ п/п	Наименование лабораторных работ / краткое содержание
1	Формирование текстового корпуса Студент собирает набор русскоязычных документов или сообщений и загружает его в структуру Pandas. Для каждого текста задаются источник, метаданные и целевая разметка при ее наличии. Выполняется проверка дублей, пропусков, длины документов и распределения классов.
2	Очистка и нормализация текста Студент реализует очистку текста от служебных символов, лишних пробелов, повторов и технического шума. Выполняются разбиение на предложения и токены, нормализация регистра и сохранение промежуточного корпуса. До и после обработки строятся сводные показатели длины текстов.
3	Морфологический и синтаксический анализ Студент выполняет лемматизацию, определение частей речи и выделение синтаксических связей средствами Natasha, pymorphy3 или spaCy. Для выбранных текстов формируются таблицы словоформ, лемм и грамматических признаков. Результаты используются для уточнения правил очистки и отбора признаков.
4	Классификация текстов по частотным признакам Студент строит признаки на основе n-грамм и TF-IDF средствами scikit-learn. Обучается модель классификации и проверяется качество на отложенной выборке. Ошибочные примеры сохраняются для последующего анализа.
5	Извлечение именованных сущностей Студент применяет Natasha, DeepPavlov или модель Transformer для выделения персон,

№ п/п	Наименование лабораторных работ / краткое содержание
	организаций, мест, дат, номеров и технических объектов. Извлеченные сущности приводятся к табличному виду с типом, позицией и исходным фрагментом. Для ошибочных случаев уточняются правила предобработки или постобработки.
6	Смысловое представление и поиск документов Студент строит векторные представления документов с использованием sentence-transformers. Документы загружаются в Qdrant, FAISS или локальный индекс для поиска ближайших фрагментов. Для контрольных запросов оценивается релевантность найденных документов.
7	Реализация RAG для ответа по корпусу документов Студент разбивает документы на фрагменты, сохраняет метаданные источников и подключает смысловой поиск к языковой модели. Ответ формируется только на основе найденных фрагментов с указанием использованных источников. Проверяются запросы с подтвержденным ответом, противоречивыми данными и отсутствием сведений в корпусе.
8	Оценка качества и оформление текстового решения Студент рассчитывает метрики классификации, извлечения сущностей и поиска по смысловой близости. По ошибочным примерам формируются выводы о качестве корпуса, признаков, модели и ограничениях применения. Техническая документация включает описание данных, обработки, моделей, параметров, метрик и результатов проверки.

4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Изучение рекомендованной литературы.
2	Подготовка к лабораторным работам.
3	Подготовка к промежуточной аттестации.
4	Подготовка к текущему контролю.

5. Перечень изданий, которые рекомендуется использовать при освоении дисциплины (модуля).

№ п/п	Библиографическое описание	Место доступа
1	Обработка естественного языка с использованием языка программирования Python : учебное пособие : в 2 частях / составитель А. Б. Мантусов. — Элиста : КГУ, 2022 — Часть 1 — 2022. — 56 с. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/360923 (дата обращения: 22.06.2026)
2	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: 22.06.2026)
3	Гольдберг, Й. Неи?росетевые методы в обработке естественного языка : руководство / Й. Гольдберг ;	Лань : электронно-библиотечная система. — URL:

	перевод с английского А. А. Слинкина. — Москва : ДМК Пресс, 2019. — 282 с. — ISBN 978-5-97060-754-1. — Текст : электронный	https://e.lanbook.com/book/131704 (дата обращения: 22.06.2026)
4	Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах. — Москва : ДМК Пресс, 2015. — 400 с. — ISBN 978-5-97060-273-7. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/69955 (дата обращения: 22.06.2026)
5	Антонов, И. В. Прикладное глубокое обучение : учебное пособие / И. В. Антонов, Ю. В. Бруттан. — Псков : ПсковГУ, 2024. — 138 с. — ISBN 978-5-00200-228-3. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/494504 (дата обращения: 22.06.2026)
6	Душкин, Р. В. RAG-системы. От теории к практике / Р. В. Душкин. — Москва : ДМК Пресс, 2026. — 286 с. — ISBN 978-5-93700-429-1. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/521215 (дата обращения: 22.06.2026)

6. Перечень современных профессиональных баз данных и информационных справочных систем, которые могут использоваться при освоении дисциплины (модуля).

ЭБС Лань – https://e.lanbook.com/.

Образовательная платформа Юрайт – https://urait.ru/.

Единый реестр российских программ для ЭВМ и баз данных – https://reestr.digital.gov.ru/reestr/.

Профессиональные стандарты и квалификации, справочная информация КонсультантПлюс – https://www.consultant.ru/document/cons_doc_LAW_157436/.

Национальная стратегия развития искусственного интеллекта на период до 2030 года – http://www.kremlin.ru/acts/bank/44731.

Документация Python – https://docs.python.org/3/.

Документация Pandas – https://pandas.pydata.org/docs/.

Документация scikit-learn – https://scikit-learn.org/stable/.

Документация PyTorch – https://docs.pytorch.org/.

Документация Hugging Face Transformers – https://huggingface.co/docs/transformers/.

Документация Natasha – https://natasha.github.io/.

Документация spaCy – https://spacy.io/usage.

Документация Qdrant – https://qdrant.tech/documentation/.

Документация LangChain – https://docs.langchain.com/.

7. Перечень лицензионного и свободно распространяемого программного обеспечения, в том числе отечественного производства, необходимого для освоения дисциплины (модуля).

Операционные системы – Astra Linux, ALT Linux, РЕД ОС, Debian GNU/Linux.

Среда разработки – Python, Jupyter Notebook, Visual Studio Code.

Обработка текста и машинное обучение – Pandas, scikit-learn, PyTorch, Hugging Face Transformers, Natasha, spaCy.

Смысловой поиск и RAG – Qdrant, FAISS, LangChain.

Сопровождение проекта – Git.

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

Учебные аудитории для проведения учебных занятий, оснащенные компьютерной техникой и наборами демонстрационного оборудования.

Для лабораторных занятий – наличие персональных компьютеров вычислительного класса.

9. Форма промежуточной аттестации:

Зачет в 7 семестре.

10. Оценочные материалы.

Оценочные материалы, применяемые при проведении промежуточной аттестации, разрабатываются в соответствии с локальным нормативным актом РУТ (МИИТ).

Авторы:

старший преподаватель кафедры
«Цифровые технологии управления
транспортными процессами» Е.А. Заманов

Согласовано:

Заведующий кафедрой ЦТУТП В.Е. Нутович

Председатель учебно-методической
комиссии Н.А. Андриянова