

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
09.03.01 Информатика и вычислительная техника,
утвержденной директором РУТ (МИИТ)
Париновым Д.В.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Анализ больших текстовых данных и текстовый поиск

Направление подготовки: 09.03.01 Информатика и вычислительная техника

Направленность (профиль): IT-сервисы и технологии обработки данных на транспорте

Форма обучения: Очная

Рабочая программа дисциплины (модуля) в виде
электронного документа выгружена из единой
корпоративной информационной системы управления
университетом и соответствует оригиналу

Простая электронная подпись, выданная РУТ (МИИТ)
ID подписи: 937226
Подписал: руководитель образовательной программы
Проневич Ольга Борисовна
Дата: 10.06.2026

1. Общие сведения о дисциплине (модуле).

Целями освоения учебной дисциплины являются формирование у обучающихся системных знаний и практических навыков в области обработки, анализа и поиска больших текстовых данных. В рамках курса рассматриваются современные методы и алгоритмы обработки естественного языка (NLP), извлечения информации из неструктурированных текстов, построения поисковых систем и рекомендательных сервисов. Особое внимание уделяется интеграции результатов текстового анализа в бизнес-процессы компании, формулировке практических рекомендаций и адаптации аналитических моделей под потребности заказчика.

Задачами освоения дисциплины являются:

1. формирование теоретических знаний о методах обработки естественного языка, алгоритмах токенизации, лемматизации, векторизации текстов и построении поисковых индексов;
2. освоение практических навыков предобработки и анализа больших текстовых массивов с использованием библиотек машинного обучения;
3. приобретение умений реализовывать алгоритмы машинного обучения для задач классификации, кластеризации, тематического моделирования и извлечения сущностей из текстов;
4. формирование компетенций по адаптации аналитических моделей под бизнес-потребности, преобразованию результатов анализа в бизнес-инсайты и практические рекомендации;
5. освоение инструментов создания аналитических отчетов и визуализации результатов текстового анализа для различных групп заинтересованных сторон.

Дисциплина предназначена для формирования компетенций, необходимых для решения профессиональных задач, связанных с обработкой больших текстовых данных, построением поисковых и рекомендательных систем, а также интеграцией аналитических решений в бизнес-процессы организации.

2. Планируемые результаты обучения по дисциплине (модулю).

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

ОПК-6 - Способен разрабатывать алгоритмы и компьютерные программы, пригодные для практического применения;

ПК-1 - Способен анализировать большие данные с использованием существующей в организации методологической и технологической инфраструктуры.

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

Знать:

- технологии, методы и инструменты развития компетенций в области анализа, хранения и обработки больших текстовых данных,
- технологии анализа больших текстовых данных и текстового, поиска,
- технологии обучение и дообучений больших текстовых моделей.

Уметь:

- использовать открытые источники информации (литература, интернет) для поиска актуальных средств статистической обработки и анализа больших текстовых данных,
- разрабатывать проектные решения, основанные на обработке текстовых данных,
- обучать модели формата text2text.

Владеть:

- навыками определения и соблюдения срока выполнения работ по разработке приложений, использующих большие текстовые данные,
- навыками программирования на языках, позволяющих анализировать текстовые данные,
- навыками суммаризации текста, определение тем текстов, поиска семантически близких текстов.

3. Объем дисциплины (модуля).

3.1. Общая трудоемкость дисциплины (модуля).

Общая трудоемкость дисциплины (модуля) составляет 7 з.е. (252 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

		№5	№6
Контактная работа при проведении учебных занятий (всего):	160	96	64
В том числе:			
Занятия лекционного типа	64	32	32
Занятия семинарского типа	96	64	32

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 92 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	Тема 1. Классификация текстов Рассматриваемые вопросы: - удаление нерелевантных символов - токенизация - нейронная сеть для классификации
2	Тема 2. Эмбединг в NLP Рассматриваемые вопросы: - векторное представление слов - история эмбединга - слой эмбединга в нейронной сети - борьба с переобучением в сетях с эмбедингом
3	Тема 3. Рекуррентные нейронные сети. Модели seq2seq Рассматриваемые вопросы: - рекуррентный нейрон - простая рекуррентная сеть - проблема затухающего градиента - LSTM-сети - модели, основанные на энкодере и декодере
4	Тема 4. Задача NER Рассматриваемые вопросы:

№ п/п	Тематика лекционных занятий / краткое содержание
	- постановка задачи NER - обученные NER-модели - подготовка данных для обучения решения задачи NER - лемматизация текста
5	Тема 5. Нейронные сети в текстовой аналитике Рассматриваемые вопросы: - слои свертки - слой эмбединга - архитектуры нейронных сетей
6	Тема 6. Семантические сети Рассматриваемые вопросы: - история семантических сетей и современное применения - граф знаний - векторное представление слова в задача анализа семантики - word2vector - glove - формирование графа знаний
7	Тема 7. Тематическое моделирование Рассматриваемые вопросы: - задача Topic Modeling - методы тематическое моделирования - LDA. Скрытое размещение Дирихле - Архитектуры нейронных сетей для больших лингвистических моделей
8	Тема 8. LSA-анализ и суммаризация Рассматриваемые вопросы: - LSA - Суммаризация и реферирование текстов
9	Тема 9. Seq2Seq модели с вниманием Рассматриваемые вопросы: - механизм внимания - архитектура энкодера с вниманием - архитектура с декодером с вниманием - особенности токенизации в моделях с вниманием
10	Тема 10. Трансформеры Рассматриваемые вопросы: - история появления - отличия от рекуррентных нейронных сетей - алгоритм работы трансформера - модель BERT
11	Тема 11. Большие лингвистические модели Рассматриваемые вопросы: - история развития - предобученные LLM на русском языке - дообучение обученных LLM
12	Тема 12. Настройка обученных больших лингвистических моделей Рассматриваемые вопросы: - подключение к обученной LLM - основные гиперпараметры, влияющие на длину и содержание ответа
13	Тема 13. Модели text2SQL Рассматриваемые вопросы:

№ п/п	Тематика лекционных занятий / краткое содержание
	- проблема задачи преобразования текста с естественного языка на искусственный язык - история моделей text2SQL - особенности обучения и использования text2SQL

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Тема 1. Токенизация Рассматриваемые вопросы: - библиотеки для токенизации - гиперпараметры для токенизации - проведения токенизации при различных значений гиперпараметров
2	Тема 2. Обучение модели классификации текста Рассматриваемые вопросы: - подготовка целевой метки - подготовка массива признаков - задание архитектуры нейронной сети - оценка качества обучения
3	Тема 3. Рекуррентные нейронные сети Рассматриваемые вопросы: - токенизация для рекуррентной сети - архитектуры нейронных сетей с рекуррентными слоями - сравнение результатов работы моделей - LSTM - упрощенные модели LSTM - исследование зависимости времени работы LSTM от гиперпараметров
4	Тема 4. Модель seq2seq Рассматриваемые вопросы: - программирование энкодера - программирование декодера - обучение модели seq2seq - формирование конечной архитектуры модели seq2seq - формирование ответа модели
5	Тема 5. Задача NER Рассматриваемые вопросы: - использование и анализ результатов обученных NER-моделей - обучение собственной NER-модели
6	Тема 6. Анализ результатов работы нейронной сети Рассматриваемые вопросы: - моделирование результатов работы сверточных слоев - моделирование результатов работы рекуррентных слоев в зависимости от гиперпараметров
7	Тема 7. Семантические сети Рассматриваемые вопросы: - векторное представление слов с помощью word2vector - векторное представление слов с glove - формирование графа знаний на примеры текстов о транспорте

№ п/п	Тематика практических занятий/краткое содержание
8	Тема 8. Тематическое моделирование Рассматриваемые вопросы: - пример выделения тем - обзор и применение библиотек для тематического моделирования
9	Тема 9. LSA-анализ и суммаризация Рассматриваемые вопросы: - проведения LSA анализа - Суммаризация и реферирование текстов
10	Тема 10. Трансформеры и большие лингвистические модели Рассматриваемые вопросы: - трансформеры open.ai - модель BERT - предобученные LLM на русском языке - дообучение обученных LLM - подключение к обученной LLM - основные гиперпараметры, влияющие на длину и содержание ответа
11	Тема 11. Модели text2SQL Рассматриваемые вопросы: - анализ моделей text2SQL - разработка алгоритмов использования обученных моделей для формирования запроса на русском языке
12	Тема 12. PolyAnalyst. Часть 1. Рассматриваемые вопросы: - словари, их место в текстовой аналитике; - структура приложения для анализа текстовых данных; - Узлы импорта, экспорта данных, узлы текстовой аналитики; - узлы анализа данных, создание словарей; - определение языка текстового массива; - токенизация текста; - выделение ключевых слов;
13	Тема 13. PolyAnalyst. Часть 2 Рассматриваемые вопросы: - шкалы тональности текста; - роль весовых значений в анализе тональности; - узлы импорта, экспорта данных, узлы текстовой аналитики; - узлы анализа данных, создание словарей; - определение языка текстового массива; - токенизация текста; - выделение ключевых слов;
14	Тема 14. Применение текстовой аналитики и бизнес-процессы организации Рассматриваемые вопросы: - описание бизнес-процессов с точки зрения анализа текстовых данных - методы анализа бизнес-процессов как объекта текстовых данных
15	Тема 15. Графовые методы анализа текста Рассматриваемые вопросы: - Применение узла термин слов - построение графа связей терминов - выделение закономерностей в графе - выделение связей в графе

№ п/п	Тематика практических занятий/краткое содержание
16	Тема 16. Создание словаря на основе графового анализа и применение его в графовом анализе с целью получения дополнительных данных и интерпретация результатов Рассматриваемые вопросы: - создание словаря на основе графа знаний - интерпретация результатов

4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Поиск алгоритмов обработки данных в открытых источниках
2	Развертывание предобученных LLM
3	Участие в онлайн мастер классах и конференциях
4	Подготовка к практическим занятиям
5	Выполнение курсовой работы.
6	Подготовка к промежуточной аттестации.
7	Подготовка к текущему контролю.

4.4. Примерный перечень тем курсовых работ

1. Источники текстовых данных как внутри организаций так и её за пределами
2. Компьютерная лингвистика и Text Mining
3. Частотный анализ терминов в коллекции документов
4. Выделение наиболее значимых слов
5. Автоматическое извлечение наиболее важных тем
6. Кластеризация документов на основе сходства их содержания
7. Построение текстовых правил для категоризации
8. Кодирование текстовой информации с помощью Python
9. Предварительная обработка данных
10. Модуль для анализа данных pandas
11. Модуль для анализа данных scikit-learn
12. Модуль для анализа данных rymorphy
13. Построение модели данных
14. Введение в анализ текстов, базовые методы предобработки и выделения признаков

15. Неглубокие векторные представления слов
16. Классификация текстов
17. Разметка последовательности
18. Предобученные языковые модели.
19. Синтаксис в рамках грамматики зависимостей
20. Тематическое моделирование
21. Суммаризация и симплификация текстов
22. QA-системы, чат-боты
23. Графы знаний

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

№ п/п	Библиографическое описание	Место доступа
1	Лабковская, Р. Я. Анализ больших данных : учебное пособие / Р. Я. Лабковская, П. В. Косов. — Санкт-Петербург : СПбГУТ им. М.А. Бонч-Бруевича, 2025. — 152 с. — ISBN 978-5-89160-366-0.	https://e.lanbook.com/book/508654
2	Цуканова, Н. И. Библиотека Pandas в задачах интеллектуального анализа данных и машинного обучения : учебное пособие / Н. И. Цуканова. — Рязань : РГРТУ, 2025. — 320 с. — ISBN 978-5-906923-13-4.	https://e.lanbook.com/book/494573

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

<https://habr.com/ru> - база знаний в виде статей, обзоров

<https://journal.tinkoff.ru/short/ai-for-all/> - база данных нейронных сетей

<https://vc.ru/services/916617-luchshie-neyroseti-bolshaya-podborka-iz-top-200-ii-generatorov-po-kategoriyam> - база данных нейронных сетей

<https://github.com/abalmumcu/bert-rest-api> - профессиональная платформа для командой работы над проектов (нейронная сеть bert)

<http://library.miit.ru/> - электронно-библиотечная система Научно-технической библиотеки МИИТ

<https://proglib.io/p/raspoznavanie-obektov-s-pomoshchyu-yolo-v3-na-tensorflow-2-0-2020-11-08> - профессиональная библиотека программистов
https://yandex.cloud/ru/blog/posts/2022/12/andrey-berger-and-yandex-cloud?utm_referrer=https%3A%2F%2Fyandex.ru%2F – библиотека профессиональных статей разработчиков Яндекс
<https://yandex.cloud/ru/blog> - библиотека профессиональных статей разработчиков Яндекс
<https://tproger.ru/translations/opencv-python-guide> - библиотека основных команд OpenCV

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

- 1 Операционная система семейства MicrosoftWindows
- 2 Пакет офисных программ MicrosoftOffice.

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

Компьютер преподавателя
Компьютеры студентов
экран для проектора, маркерная доска,
Проектор

9. Форма промежуточной аттестации:

Зачет в 5 семестре.
Курсовая работа в 5 семестре.
Экзамен в 6 семестре.

10. Оценочные материалы.

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

Авторы:

заместитель директора

Б.В. Игольников

руководитель образовательной
программы

О.Б. Проневич

Согласовано:

Руководитель образовательной
программы

О.Б. Проневич

Председатель учебно-методической
комиссии

Д.В. Паринов