

**МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ**  
**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ**  
**УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**  
**«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»**  
**(РУТ (МИИТ))**



Рабочая программа дисциплины (модуля),  
как компонент образовательной программы  
высшего образования - программы бакалавриата  
по направлению подготовки  
09.03.01 Информатика и вычислительная техника,  
утвержденной первым проректором РУТ (МИИТ)  
Тимониным В.С.

**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: 170737  
Подписал: заместитель директора академии Паринов Денис Владимирович  
Дата: 10.10.2024

## 1. Общие сведения о дисциплине (модуле).

Целью освоения учебной дисциплины «Анализ больших текстовых данных и текстовый поиск» является теоретическая и практическая подготовка студентов к работе с большими текстовыми данными и интеллектуальному анализу текста. Знания и компетенции, полученные в результате освоения дисциплины, помогут при автоматизированном интеллектуальном анализе больших объемов текстовой информации, что позволит успешно решать практические задачи обработки данных, возникающие в процессе профессиональной деятельности.

Задачи освоения дисциплины:

приобретение студентами знаний о моделях и методах интеллектуального анализа текстовых данных и машинного обучения;

развитие навыков программирования на языках, позволяющих анализировать текстовые данные;

формирование представления о сборе, обработке и анализе данных в интерактивных средах;

## 2. Планируемые результаты обучения по дисциплине (модулю).

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

**ОПК-3** - Способен решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности;

**ОПК-8** - Способен разрабатывать алгоритмы и 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 4 з.е. (144 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 №4
Контактная работа при проведении учебных занятий (всего):	66	66
В том числе:		
Занятия лекционного типа	16	16
Занятия семинарского типа	50	50

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 78 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	<p>Тема 1. Классификация текстов</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- удаление нерелевантных символов</li> <li>- токенизация</li> <li>- нейронная сеть для классификации</li> </ul>
2	<p>Тема 2. Эмбединг в NLP</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- векторное представление слов</li> <li>- история эмбединга</li> <li>- слой эмбединга в нейронной сети</li> <li>- борьба с переобучение в сетях с эмбедингом</li> </ul>
3	<p>Тема 3. Рекуррентные нейронные сети. Модели seq2seq</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- рекуррентный нейрон</li> <li>- простая рекуррентная сеть</li> <li>- проблема затухающего градиента</li> <li>- LSTM-сети</li> <li>- модели, основанные на энкодере и декодере</li> </ul>
4	<p>Тема 4. Задача NER</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- постановка задачи NER</li> <li>- обученные NER-модели</li> <li>- подготовка данных для обучения решения задачи NER</li> <li>- лемматизация текста</li> </ul>
5	<p>Тема 5. Нейронные сети в текстовой аналитике</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- слои свертки</li> <li>- слой эмбединга</li> <li>- архитектуры нейронных сетей</li> </ul>
6	<p>Тема 6. Семантические сети</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- история семантических сетей и современное применения</li> <li>- граф знаний</li> <li>- векторное представление слова в задача анализа семантики</li> <li>- word2vector</li> <li>- glove</li> <li>- формирование графа знаний</li> </ul>

№ п/п	Тематика лекционных занятий / краткое содержание
7	Тема 7. Тематическое моделирование Рассматриваемые вопросы: - задача Topic Modeling - методы тематического 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: - 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 значениях гиперпараметров
2	Тема 2. Обучение модели классификации текста Рассматриваемые вопросы:

№ п/п	Тематика практических занятий/краткое содержание
	<ul style="list-style-type: none"> <li>- подготовка целевой метки</li> <li>- подготовка массива признаков</li> <li>- задание архитектуры нейронной сети</li> <li>- оценка качества обучения</li> </ul>
3	<p>Тема 3. Обучении модели классификации текста с использование эмбединга</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- задание архитектуры нейронной сети с эмбединга-слоем</li> <li>- обучение модели</li> <li>- сравнение результатов работы моделей с эмбединга</li> </ul>
4	<p>Тема 4. Рекуррентные нейронные сети</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- токенизация для рекуррентной сети</li> <li>- архитектуры нейронных сетей с рекуррентными слоями</li> <li>- сравнение результатов работы моделей</li> </ul>
5	<p>Тема 5. Борьба с затухающим градиентом в рекуррентных нейронных сетях</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- LSTM</li> <li>- упрощенные модели LSTM</li> <li>- исследование зависимости времени работы LSTM от гиперпараметров</li> </ul>
6	<p>Тема 6. Модель seq2seq без внимания</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- программирование энкодера</li> <li>- программирование декодера</li> <li>- обучение модели seq2seq</li> <li>- формирование конечной архитектуры модели seq2seq</li> <li>- формирование ответа модели</li> </ul>
7	<p>Тема 7. Модель seq2seq с вниманием</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- программирование энкодера с вниманием</li> <li>- программирование декодера с вниманием</li> <li>- обучение модели seq2seq с вниманием</li> <li>- формирование конечной архитектуры модели seq2seq с вниманием</li> <li>- формирование ответа модели</li> </ul>
8	<p>Тема 8. Задача NER</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- использование и анализ результатов обученных NER-моделей</li> <li>- обучение собственной NER-модели</li> </ul>
9	<p>Тема 9. Анализ результатов работы нейронной сети</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- моделирование результатов работы сверточных слоев</li> <li>- моделирование результатов работы рекуррентных слоев в зависимости от гиперпараметров</li> </ul>
10	<p>Тема 10. Семантические сети</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- векторное представление слов с помощью word2vector</li> <li>- векторное представление слов с glove</li> <li>- формирование графа знаний на примеры текстов о транспорте</li> </ul>
11	<p>Тема 11. Тематическое моделирование</p> <p>Рассматриваемые вопросы:</p>

№ п/п	Тематика практических занятий/краткое содержание
	<ul style="list-style-type: none"> <li>- пример выделения тем</li> <li>- обзор и применение библиотек для тематического моделирования</li> </ul>
12	<p>Тема 12. LSA-анализ и суммаризация</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- проведения LSA анализа</li> <li>- Суммаризация и реферирование текстов</li> </ul>
13	<p>Тема 13. Трансформеры</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- трансформеры open.ai</li> <li>- модель BERT</li> </ul>
14	<p>Тема 14. Большие лингвистические модели</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- предобученные LLM на русском языке</li> <li>- дообучение обученных LLM</li> </ul>
15	<p>Тема 15. Настройка обученных больших лингвистических моделей</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- подключение к обученной LLM</li> <li>- основные гиперпараметры, влияющие на длину и содержание ответа</li> </ul>
16	<p>Тема 16. Модели text2SQL</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- анализ моделей text2SQL</li> <li>- разработка алгоритмов использования обученных моделей для формирование запроса на русском языке</li> </ul>
17	<p>Тема 17. PolyAnalyst. Основы</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>-словари, их место в текстовой аналитики;</li> <li>-структура приложения для анализа текстовых данных</li> </ul>
18	<p>Тема 18. Составление словарей для тематики РЖД</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- выделение тем</li> <li>- методы составления словарей</li> </ul>
19	<p>Тема 19. Анализ тональности</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>-шкалы тональности текста</li> <li>-Роль весовых значений в анализе тональности</li> </ul>
20	<p>Тема 20. PolyAnalyst. Узлы</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- Узлы импорта, экспорта данных, узлы текстовой аналитики;</li> <li>- узлы анализа данных, создание словарей</li> </ul>
21	<p>Тема 21. импорт пакета данных в аналитическую платформу Polyanalyst;</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- Определение языка текстового массива; токенизация текста;</li> <li>- выделение ключевых слов;</li> </ul>
22	<p>Тема 22. Применение текстовой аналитики и бизнес-процессы организации</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- описание бизнес-процессов с точки зрения анализа текстовых данных</li> <li>- методы анализа бизнес-процессов как объекта текстовых данных</li> </ul>
23	<p>Тема 23. Графовые методы анализа текста</p> <p>Рассматриваемые вопросы:</p>

№ п/п	Тематика практических занятий/краткое содержание
	- Применение узла термин слов - построение графа связей терминов
24	Тема 24. Процесса анализа графа с целью выделения закономерностей и выделения основных связей Рассматриваемые вопросы: - выделение закономерностей в графе - выделение связей в графе
25	Тема 25. 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
26	Тема 26. Предварительный анализ проекта "Условия эксплуатации тепловозов и их обслуживание" Рассматриваемые вопросы: - определение инструментов анализа - проведение анализа
27	Тема 27. Применение диаграммы с целью выделения закономерностей и выделения основных связей Рассматриваемые вопросы: - выявление закономерностей - выделение основных связей
28	Тема 28. Оценка результатов анализа текстов Рассматриваемые вопросы: - применение OLAP таблицы и агрегирования данных по метрикам; - интерпретация результата

#### 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	Язык программирования Python Г. Россум, Ф.Л.Дж. Дрейк, Д.С.	НТБ (ЭЭ)

	Откидач Однотомное издание 2001	
2	Рашка, С. Python и машинное обучение: крайне необходимое пособие по новейшей предсказатель ной аналитике, обязательное для более глубокого понимания методологии машинного обучения : руководство / С. Рашка ; перевод с английского А. В. Логунова. — Москва : ДМК Пресс, 2017. — 418 с. — ISBN 978-5-97060- 409-0	<a href="https://e.lanbook.com/book/100905?category_pk=1556&amp;publisher__fk=1028&amp;ysclid=1w60n980di451663507">https://e.lanbook.com/book/100905?category_pk=1556&amp;publisher__fk=1028&amp;ysclid=1w60n980di451663507</a>
3	Лю, Ю. Обучение с подкрепление м на PyTorch. Сборник рецептов : руководство / Ю. Лю ; перевод с английского А. А. Слинкина. —	<a href="https://e.lanbook.com/book/179493">https://e.lanbook.com/book/179493</a>

	Москва : ДМК Пресс, 2020. — 282 с. — ISBN 978-5-97060- 853-1	
4	Крошемор, М. Алгоритмы обработки текста. 125 задач с решениями / М. Крошемор, Т. Лекрок, В. Ритгер ; перевод с английского А. А. Слинкина. — Москва : ДМК Пресс, 2021. — 312 с. — ISBN 978-5-97060- 952-1	<a href="https://e.lanbook.com/book/241070">https://e.lanbook.com/book/241070</a>

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](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.

3 Язык программирования Python

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:

Курсовая работа в 4 семестре.

Экзамен в 4 семестре.

10. Оценочные материалы.

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

Авторы:

доцент, к.н. Академии "Высшая  
инженерная школа"

Б.В. Игольников

Согласовано:

Заместитель директора академии

Д.В. Паринов

Председатель учебно-методической  
комиссии

Д.В. Паринов