

**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 магистратуры  
по направлению подготовки  
09.04.03 Прикладная информатика,  
утвержденной первым проректором РУТ (МИИТ)  
Тимониным В.С.

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)**

**Анализ больших текстовых данных и текстовый поиск**

Направление подготовки: 09.04.03 Прикладная информатика

Направленность (профиль): Процессная аналитика

Форма обучения: Заочная

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: 30712  
Подписал: руководитель образовательной программы  
Моргунов Виталий Михайлович  
Дата: 03.06.2024

## 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 магистратуры 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-2** - Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач;

**ОПК-4** - Способен применять на практике новые научные принципы и методы исследований;

**ПК-4** - Способен разрабатывать информационные продукты, сервисы и инфраструктурные решения на основе аналитики больших данных.

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

### **Уметь:**

Работать с библиотеками Pandas, NLTK, textblob

Проводить токенизацию слов, работать со списками стоп-слов

Вычислять близость текстов и применять этот метод на реальных данных.

Применять метод LSTM для решения задачи NER

Реализовывать моноязычный и мультиязычный тематический поиск.

Создавать генераторы текста с помощью transformers.

Классифицировать тексты с использованием предобученной модели BERT

Решать практические задачи текстовой аналитики

### **Владеть:**

Навыками работы со следующими инструментами:

Pandas, NLTK, textblob, Scikit-learn, SpaCy

gensim — инструмент для решения различных задач NLP (тематическое моделирование, представление текстов

numpy — библиотека для работы с векторами.

scikit-learn — библиотека с многими реализованными алгоритмами машинного обучения для анализа данных.

pytorch – библиотека для работы с тензорами и обучения нейросетей.

bigartm, rumorphy2, nltk — инструменты для работы с естественными языками.

Навыками разработки алгоритмов анализа текста

### 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 №3
Контактная работа при проведении учебных занятий (всего):	8	8
В том числе:		
Занятия лекционного типа	4	4
Занятия семинарского типа	4	4

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 100 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>Текстовая аналитика</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- Кодирование текстовой информации с помощью Python</li> <li>- Предварительная обработка данных</li> <li>- Модуль для анализа данных pandas</li> <li>- Модуль для анализа данных scikit-learn</li> <li>- Модуль для анализа данных rymorphy</li> <li>- Построение модели данных</li> </ul>
2	<p>Анализ текстовой информации с помощью Python</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- Источники текстовых данных как внутри организаций (данные колл-центров, архив e-mail, онлайн-опросы, анкеты), за пределами (блоги и форумы, соцсети, поисковые запросы клиентов).</li> <li>- Компьютерная лингвистика и Text Mining</li> <li>- Частотный анализ терминов в коллекции документов</li> <li>- Выделение наиболее значимых слов</li> <li>- Автоматическое извлечение наиболее важных тем</li> <li>- Кластеризация документов на основе сходства их содержания</li> <li>- Построение текстовых правил для категоризации</li> </ul>
3	<p>Обработка текстов методами машинного обучения</p> <p>Рассматриваемые вопросы</p> <ul style="list-style-type: none"> <li>- Введение в анализ текстов, базовые методы предобработки и выделения признаков</li> <li>- Неглубокие векторные представления слов</li> <li>- Классификация текстов</li> <li>- Разметка последовательности</li> <li>- Предобученные языковые модели.</li> <li>- Синтаксис в рамках грамматики зависимостей</li> <li>- Тематическое моделирование</li> <li>- Суммаризация и симплификация текстов</li> <li>- QA-системы, чат-боты</li> <li>- Графы знаний</li> </ul>

## 4.2. Занятия семинарского типа.

### Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Библиотеки и модули анализа данных Python (Pandas, Scikit-learn, Pymorphy) 1. Использование Pandas 2. Использование Scikit-learn 3. Использование Pymorphy
2	Индивидуальные проекты на основе библиотек и модулей анализа данных Python (Pandas, Scikit-learn, Pymorphy) (аудиторный этап) Мультиязычный тематический поиск Генерация программного кода по заданному запросу с помощью transformers Классификация с использованием BERT и Transformers

## 4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	поиск алгоритмов обработки данных в открытых источниках
2	работа с учебной литературой
3	участие в онлайн мастер классах и конференциях
4	Индивидуальные проекты на основе библиотек и модулей анализа данных Python (Pandas, Scikit-learn, Pymorphy) (самостоятельный этап)
5	Решение задач
6	Подготовка к промежуточной аттестации.

## 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. Откидач Однотомное издание 2001	НТБ (ЭЭ)
2	Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения Рашка С. Издательство "ДМК Пресс" , 2017	<a href="https://reader.lanbook.com/book/100905#222">https://reader.lanbook.com/book/100905#222</a>
3	Васильев, Ю. Обработка естественного языка. Python и spaCy	<a href="https://www.labyrinth.ru/books/800781/">https://www.labyrinth.ru/books/800781/</a>

	на практике. – СПб.: Питер, 2021. – 256 с.: ил. – (Серия «Библиотека программиста»)	
4	Крошемор М., Лекрок Т., Ритгер В. Алгоритмы обработки текста: 125 задач с решениями / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2021. – 312 с.: ил.	<a href="https://www.labyrinth.ru/books/812710/">https://www.labyrinth.ru/books/812710/</a>
5	Лю Ю. Обучение с подкреплением на PyTorch: сборник рецептов / пер. с англ.	<a href="https://ibooks.ru/products/372101?category_id=12853">https://ibooks.ru/products/372101?category_id=12853</a>

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

[https://pyneng.readthedocs.io/ru/latest/book/Part\\_I.html](https://pyneng.readthedocs.io/ru/latest/book/Part_I.html)

<https://colab.research.google.com/>

<https://e.lanbook.com/>

<https://rusneb.ru/>

Основы Natural Language Processing для текста [Электронный ресурс]  
URL: <https://habr.com/ru/company/Voximplant/blog/446738/>

Основы Python [Электронный ресурс] URL:  
[https://pyneng.readthedocs.io/ru/latest/book/Part\\_I.html](https://pyneng.readthedocs.io/ru/latest/book/Part_I.html)

Готовые модели [Электронный ресурс] URL:  
<https://huggingface.co/models>

Vector Space Model для семантической классификации текстов [Электронный ресурс] URL: <https://habr.com/ru/sandbox/18635/>

Word2Vec: как работать с векторными представлениями слов [Электронный ресурс] URL: <https://neurohive.io/ru/osnovy-data-science/word2vec-vektornye-predstavlenija-slov-dlja-mashinnogo-obuchenija/>

Робот-помощник Stackoverflow [Электронный ресурс] URL:  
[https://translated.turbopages.org/proxy\\_u/en-ru.ru.48e81a0a-61b752eb-d4e9e5be-74722d776562/https/github.com/Vishwa22/StackOverflow-assistant-bot/blob/master/Stackoverflow%20assistant%20bot.md](https://translated.turbopages.org/proxy_u/en-ru.ru.48e81a0a-61b752eb-d4e9e5be-74722d776562/https/github.com/Vishwa22/StackOverflow-assistant-bot/blob/master/Stackoverflow%20assistant%20bot.md)

Математические методы анализа текстов [Электронный ресурс] URL:  
[http://www.machinelearning.ru/wiki/images/8/8b/Mel\\_lain\\_msu\\_nlp\\_sem\\_7.pdf](http://www.machinelearning.ru/wiki/images/8/8b/Mel_lain_msu_nlp_sem_7.pdf)

LSTM — нейронная сеть с долгой краткосрочной памятью [Электронный ресурс] URL: <https://neurohive.io/ru/osnovy-data-science/lstm-nejronnaja-set/>

Иллюстрированное руководство по LSTM и GRU: пошаговое объяснение [Электронный ресурс] URL: <https://www.machinelearningmastery.ru/illustrated->

guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21/

Тематический анализ больших данных [Электронный ресурс] URL: <http://www.machinelearning.ru/wiki/images/6/6d/BigARTM-short-intro.pdf>

Fast and Modular Regularized Topic Modelling [Электронный ресурс] URL: <https://fruct.org/publications/fruct21/files/Кос.pdf>

Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM [Электронный ресурс] URL: <http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>

Создание генератора поп-музыки с помощью Transformer [Электронный ресурс] URL: <https://www.machinelearningmastery.ru/creating-a-pop-music-generator-with-the-transformer-5867511b382a/>

Генераторы исходного кода [Электронный ресурс] URL: <https://docs.microsoft.com/ru-ru/dotnet/csharp/roslyn-sdk/source-generators-overview>

IntelliCode Compose: нейросеть дополняет код с помощью Transformer [Электронный ресурс] URL: <https://neurohive.io/ru/novosti/intellicode-compose-nejroset-dopolnyaet-kod-s-pomoshhju-transformer/>

BERT, ELMO и Ко в картинках (как в NLP пришло трансферное обучение) [Электронный ресурс] URL: <https://habr.com/ru/post/487358/>

Как использовать BERT для мультиклассовой классификации текста [Электронный ресурс] URL: <https://neurohive.io/ru/tutorial/bert-klassifikatsiya-teksta/>

Практическое руководство по классификации текста с использованием моделей трансформаторов (XLNet, BERT, XLM, RoBERTa) [Электронный ресурс] URL: <https://www.machinelearningmastery.ru/https-medium-com-chaturangarajakashe-text-classification-with-transformer-models-d370944b50ca/>

Классификация текста с помощью BERT Tokenizer и TF 2.0 в Python [Электронный ресурс] URL: <https://pythobyte.com/text-classification-with-bert-tokenizer-and-tf-2-0-in-python-44cafd87/>

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

Microsoft Office 2010

Google Colab

Jupyter Notebook

Google Drive

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

1 учебный класс (столы, стулья - по 25 ед)

Компьютер преподавателя

Intel Core i7-9700 / Asus PRIME H310M-R R2.0 / 2x8GB / SSD 250Gb /

DVDRW

Компьютеры студентов (24 ед)

Intel Core i9-9900 / B365M Pro4 / 2x16GB / SSD 512Gb

Проектор Optoma W340UST

Экран для проектора

Маркерная доска

9. Форма промежуточной аттестации:

Зачет в 3 семестре.

10. Оценочные материалы.

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



Авторы:

директор

Б.В. Игольников

Согласовано:

Руководитель образовательной  
программы

В.М. Моргунов

Председатель учебно-методической  
комиссии

Д.В. Паринов