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

1. Общие сведения о дисциплине (модуле).

Целью изучения дисциплины является теоретическая и практическая подготовка студентов к работе с большими текстовыми данными и интеллектуальному анализу текста.

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-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 анализа, хранения и обработки больших текстовых данных;
- методы анализа текстовых данных с использованием искусственного интеллекта;
- основные инструментальные средства анализа текстовых данных и текстового поиска

Уметь:

- классифицировать задачи текстовой аналитики;
- использовать стандартные библиотеки Python для решения задач анализа текстовых данных;
- использовать инструментальные средства для решения основных задач текстового поиска

Владеть:

- навыками использования инструментальных средств Pandas, NLTK, textblob, Scikit-learn, SpaCy для анализа больших текстовых данных;
- методами анализа больших текстовых данных с использованием алгоритмов машинного обучения;
- навыками выбора и обоснования алгоритмов анализа текста для решения задач в профессиональной сфере

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	Текстовая аналитика Рассматриваемые вопросы: - Кодирование текстовой информации с помощью Python - Предварительная обработка данных - Модуль для анализа данных pandas - Модуль для анализа данных scikit-learn - Модуль для анализа данных pymorphy - Построение модели данных
2	Анализ текстовой информации с помощью Python Рассматриваемые вопросы: - Источники текстовых данных как внутри организаций (данные колл-центров, архив e-mail, онлайн-опросы, анкеты), за пределами (блоги и форумы, соцсети, поисковые запросы клиентов). - Компьютерная лингвистика и Text Mining - Частотный анализ терминов в коллекции документов - Выделение наиболее значимых слов - Автоматическое извлечение наиболее важных тем - Кластеризация документов на основе сходства их содержания - Построение текстовых правил для категоризации
3	Обработка текстов методами машинного обучения Рассматриваемые вопросы - Введение в анализ текстов, базовые методы предобработки и выделения признаков - Неглубокие векторные представления слов - Классификация текстов - Разметка последовательности - Предобученные языковые модели. - Синтаксис в рамках грамматики зависимостей -Тематическое моделирование - Суммаризация и симплификация текстов - QA-системы, чат-боты - Графы знаний

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Библиотеки и модули анализа данных Python (Pandas, Scikit-learn, Pymorthy) 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	решение задач по темам дисциплины
4	Подготовка к промежуточной аттестации.

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

№ п/п	Библиографическое описание	Место доступа
1	Миркин, Б. Г. Базовые методы анализа данных : учебник и практикум для вузов / Б. Г. Миркин. — 3-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2025. — 297 с. — (Высшее образование). — ISBN 978-5-534-19709-9.	URL: https://urait.ru/bcode/560414 (дата обращения: 30.01.2025). — Текст : электронный.
2	Платонов, А. В. Машинное обучение : учебное пособие для вузов / А. В. Платонов. — 2-е изд. — Москва : Издательство Юрайт, 2025. — 89 с. — (Высшее образование). — ISBN 978-5-534-20732-3.	URL: https://urait.ru/bcode/558662 (дата обращения: 30.01.2025). — Текст : электронный.
3	Нугуманова, А. Б. Автоматизированная обработка текстовых массивов : учебник и практикум для вузов / А. Б. Нугуманова. — 2-е изд. — Москва : Издательство Юрайт, 2024. — 82 с. — (Высшее образование). — ISBN 978-5-534-20738-5.	URL: https://urait.ru/bcode/558668 (дата обращения: 30.01.2025). — Текст : электронный.

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

Научно-техническая библиотека РУТ (МИИТ): http://library.miit.ru

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

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

Пакет приложений Microsoft Office или аналог

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

Для проведения лекционных занятий необходима аудитория с мультимедиа аппаратурой. Для проведения практических занятий требуется аудитория, оснащенная мультимедиа аппаратурой и ПК с необходимым программным обеспечением и подключением к сети интернет.

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