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
23.03.02 Наземные транспортно-технологические комплексы,
утвержденной первым проректором РУТ (МИИТ)
Тимониным В.С.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Основы анализа данных

Направление подготовки: 23.03.02 Наземные транспортно-технологические комплексы

Направленность (профиль): Транспортный и промышленный дизайн

Форма обучения: Очная

1. Общие сведения о дисциплине (модуле).

Целью дисциплины является теоретическая и практическая подготовка студентов к работе с большими текстовыми данными, интеллектуальному анализу текста, а также освоение технологий текстового поиска и инструментов анализа данных, включая библиотеки и модули Python (например, Pandas, NumPy, NLTK, Scikit-learn). Полученные знания и компетенции позволят обучающимся эффективно решать задачи автоматизированной обработки и анализа текстовой информации, что необходимо для профессиональной деятельности в условиях цифровой трансформации.

Задачи освоения дисциплины:

Освоение теоретических основ:

– Изучение моделей, методов и алгоритмов интеллектуального анализа текстовых данных, включая техники машинного обучения.

Развитие практических навыков:

– Формирование умений программирования на Python с использованием специализированных библиотек для анализа текста (NLTK, SpaCy, TensorFlow и др.).

– Приобретение опыта работы с интерактивными средами разработки (Jupyter Notebook, Google Colab) для сбора, предобработки и визуализации данных.

Применение знаний в профессиональной деятельности:

– Обучение интерпретации результатов анализа данных и их использованию для решения прикладных задач, таких как классификация текстов, извлечение информации, анализ тональности и др.

Результат обучения:

Студенты смогут самостоятельно проектировать и реализовывать решения для анализа текстовых данных, применять современные инструменты и методы машинного обучения, а также эффективно представлять результаты исследований в рамках профессиональных задач.

2. Планируемые результаты обучения по дисциплине (модулю).

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

УК-1 - Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач.

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 задачу 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, pymorphy2, 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 №5
Контактная работа при проведении учебных занятий (всего):	32	32
В том числе:		
Занятия семинарского типа	32	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 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.

Не предусмотрено учебным планом

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Тема 1. Введение в предметную область аналитики больших данных: хайп или нераскрытый потенциал Рассматриваемые вопросы: - Данные, информация, знания – в чем отличия. - Что есть большие данные. Что есть аналитика. - Что за зверь такой: data-driven организация? - Ожидания рынка vs результаты. - Несколько основных ловушек и извлеченных из них опыт. - Пересмотр приоритетов за последние 10 лет. - Новые игроки на рынке и чем они характеризуются

№ п/п	Тематика практических занятий/краткое содержание
2	Тема 2. Стратегия управления данными: искусство видеть лес за деревьями Рассматриваемые вопросы: - Стратегия управления данными, как составная часть культуры современной организации. Почему это важный вопрос? - Что есть стратегия управления данными, кто за неё отвечает в организации. - Кто такой владелец данных, его сфера ответственности и полномочия. - Модель данных, что это такое и зачем она нужна. - Основные заинтересованные стороны. - Концепция Self-Service BI. - Техническая инфраструктура. - Революция open-source и доступность технологий. - Как измерить эффективность стратегии управления данными.
3	Тема 3. Хранилища данных: первый шаг к аналитике и зачем все так усложнять Рассматриваемые вопросы: - Разберемся в предмете: что есть хранилище данных, зачем оно нужно. - Что за зверь такой – ETL. - Определение источников и загрузка данных. - Какие виды преобразования и объединения данных существуют, в чём их принципиальные отличия и на что следует обращать внимание. - Что такое витрины данных.
4	Тема 4. Постановка задачи: как не ошибиться с выбором цели и инструментов Рассматриваемые вопросы: - С чего начать, с какой стороны подойти к задаче анализа данных. - Основные термины и понятия. - Как правильно ставить задачу и сформулировать цель анализа и почему это так важно. - Как оценить достаточно ли вам данных для анализа, где и как данные следует добывать, на что обращать внимание. - Понятие качества данных. - Методики и инструменты обеспечения и контроля качества данных. - Выработка гипотез и выбор методов.
5	Тема 5. Подготовка данных: искусство есть слона по частям Рассматриваемые вопросы: - Какие данные бывают, какие у них свойства, почему важно разобраться в их природе (рассмотрим на примере ошибок, допускаемых аналитиками при работе вслепую). - Что такое метаданные или бизнес-глоссарий и почему обязательно сопровождать аналитическую задачу созданием соответствующего репозитория метаданных. - Какие есть способы подготовки данных, на что следует обращать внимание, как не ошибиться.
6	Тема 6. Введение в статистический анализ: разбираемся в основах Рассматриваемые вопросы: - Основные понятия: генеральная совокупность и выборка. - Какие основные статистические показатели используются в аналитике и почему (рассмотрим на известных исторических примерах). - Что такое сводная таблица и как она используется в работе (рассмотрим на примере задачи финансового планирования).

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 практическим занятиям
7	Подготовка к промежуточной аттестации.
8	Подготовка к текущему контролю.

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

№ п/п	Библиографическое описание	Место доступа
1	Груздев, А. В. Прогнозное моделирование в IBM SPSS Statistics, R и Python: метод деревьев решений и случайный лес : руководство / А. В. Груздев. — Москва : ДМК Пресс, 2018. — 642 с. — ISBN 978-5-97060-539-4	https://e.lanbook.com/book/123700
2	Нестеров, С. А. Основы интеллектуального анализа данных. Лабораторный практикум : учебное пособие / С. А. Нестеров. — Санкт-Петербург : Лань, 2020. — 40 с. — ISBN 978-5-8114-4509-7	https://e.lanbook.com/book/130181

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

Электронно-библиотечная система «Лань» [Электронный ресурс]. – URL: https://e.lanbook.com/

Национальная электронная библиотека (НЭБ) [Электронный ресурс]. – URL: 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

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

Математические методы анализа текстов [Электронный ресурс] URL: 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/Koc.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-klassifikacya-teksta/

Практическое руководство по классификации текста с использованием моделей трансформаторов (XLNet, BERT, XLM, RoBERTa) [Электронный ресурс] URL: https://www.machinelearningmastery.ru/https-medium-com-chaturangarajapakshe-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).

Офисный пакет приложений – Microsoft Office

ПО для анализа данных Polymatica

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.

9. Форма промежуточной аттестации:

Зачет в 5 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:

Директор	Б.В. Игольников
Руководитель образовательной программы	Н.А. Любавин
Председатель учебно-методической комиссии	Д.В. Паринов