

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)

Управление данными и бизнес-аналитика

Направление подготовки: 09.04.03 Прикладная информатика

Направленность (профиль): Управление цифровыми активами 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: 02.09.2021

1. Общие сведения о дисциплине (модуле).

Краткая аннотация дисциплины (Целью освоения учебной дисциплины «Анализ больших текстовых данных и текстовый поиск» в соответствии с ФГОС ВО 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 анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями;

ПК-6 - Способен использовать в профессиональной деятельности методы, модели, алгоритмы, технологии и инструментальные средства работы с большими данными .

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

Знать:

Знает технологии, методы и инструменты развития компетенций в области анализа, хранения и обработки больших текстовых данных
технологии анализа больших текстовых данных и текстового поиска

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

Владеть:

навыками определения и соблюдения срока выполнения работ по разработке приложений, использующих большие текстовые данные

навыками программирования на языках, позволяющих анализировать текстовые данные.

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	Сем. №3
Контактная работа при проведении учебных занятий (всего):	26	26
В том числе:		
Занятия лекционного типа	8	8
Занятия семинарского типа	18	18

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 82 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 1. Кодирование текстовой информации с помощью Python 2. Предварительная обработка данных 3. Модуль для анализа данных pandas 4. Модуль для анализа данных scikit-learn 5. Модуль для анализа данных rymorphy 6. Построение модели данных
2	Анализ текстовой информации с помощью Python 1. Источники текстовых данных как внутри организаций (данные колл-центров, архив e-mail, онлайн-опросы, анкеты), за пределами (блоги и форумы, соцсети, поисковые запросы клиентов). 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
3	Обработка текстов методами машинного обучения 1. Введение в анализ текстов, базовые методы предобработки и выделения признаков 2. Неглубокие векторные представления слов 3. Классификация текстов 4. Разметка последовательности 5. Предобученные языковые модели. 6. Синтаксис в рамках грамматики зависимостей 7. Тематическое моделирование 8. Суммаризация и симплификация текстов 9. QA-системы, чат-боты 10. Графы знаний

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Текстовая аналитика Естественный язык и текст – введение в анализ Векторная модель текста и классификация длинных текстов Классификация новостных текстов моделью

№ п/п	Тематика практических занятий/краткое содержание
2	Тема 2. Анализ текстовой информации с помощью Python Распознавание текста моделью Word2Vec Определение частей речи при помощи нейросети Генерация имён и лозунгов с помощью рекуррентной нейронной сети Моделирование языка с помощью Transformer
3	Тема 3 Обработка текстов методами машинного обучения Распознавание структуры рецептов при помощи рекуррентной нейросети Аспектный сентимент-анализ как задача распознавания именованных сущностей Генерация программного кода по форуму решений Настройка языковой модели BERT Использование BERT для вопросно-ответного поиска

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	Выполнение курсовой работы.
5	Подготовка к промежуточной аттестации.

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. 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	https://reader.lanbook.com/book/100905#222

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

- 1 <http://www.google.ru/>
- 2 <http://ru.wikipedia.ru/>
3. <https://e.lanbook.com/>

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

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

Компьютеры студентов

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

экран для проектора, маркерная доска,

Проектор Optoma W340UST,

9. Форма промежуточной аттестации:

Курсовая работа в 3 семестре.

Экзамен в 3 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