Министерство науки и высшего образования РФ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Национальный исследовательский университет «МЭИ»

Направление подготовки/специальность: 27.03.04 Управление в технических системах

Наименование образовательной программы: Системы и средства автоматизации, интеллектуального управления и анализа данных

Уровень образования: высшее образование - бакалавриат

Форма обучения: Очная

Рабочая программа дисциплины
ОСНОВЫ АНАЛИЗА ТЕКСТОВЫХ ДАННЫХ

Блок:	**Блок 1 «Дисциплины (модули)»**
Часть образовательной программы:	**Часть, формируемая участниками образовательных отношений**
№ дисциплины по учебному плану:	**Б1.Ч.16**
Трудоемкость в зачетных единицах:	**8 семестр - 4;**
Часов (всего) по учебному плану:	**144 часа**
Лекции	**8 семестр - 24 часа;**
Практические занятия	**не предусмотрено учебным планом**
Лабораторные работы	**8 семестр - 24 часа;**
Консультации	**8 семестр - 2 часа;**
Самостоятельная работа	**8 семестр - 93,5 часа;**
в том числе на КП/КР	**не предусмотрено учебным планом**
Иная контактная работа	**проводится в рамках часов аудиторных занятий**
включая: **Лабораторная работа** **Тестирование**	
Промежуточная аттестация:	
Экзамен	**8 семестр - 0,5 часа;**

Москва 2024

ПРОГРАММУ СОСТАВИЛ:

Преподаватель	Подписано электронной подписью ФГБОУ ВО «НИУ «МЭИ» Сведения о владельце ЦЭП МЭИ Владелец: Мохов А.С. Идентификатор: R55ae9104-MokhovAS-2434a28b	А.С. Мохов

СОГЛАСОВАНО:

Руководитель образовательной программы	Подписано электронной подписью ФГБОУ ВО «НИУ «МЭИ» Сведения о владельце ЦЭП МЭИ Владелец: Шилин Д.В. Идентификатор: R495daf18-ShilinDV-59db3f0e	Д.В. Шилин
Заведующий выпускающей кафедрой	Подписано электронной подписью ФГБОУ ВО «НИУ «МЭИ» Сведения о владельце ЦЭП МЭИ Владелец: Бобряков А.В. Идентификатор: R2c90f415-BobriakovAV-70dec1fa	А.В. Бобряков

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Цель освоения дисциплины: формирование системы знаний и практических навыков студентов в области поиска, обработки, анализа текстовых данных с применением методов машинного обучения - моделей представления данных, особенностей обработки текстовых данных, методов классификации, кластеризации, выявления плагиата и других методов Text Mining

Задачи дисциплины

- освоение основных моделей и методов, используемых в интеллектуальных системах (ИС);
- овладение современными способами обработки и анализа текстовой информации;
- формирование у студентов способности самостоятельно решать задачи поиска, обработки и анализа текстовой информации.

Формируемые у обучающегося **компетенции** и запланированные **результаты обучения** по дисциплине, соотнесенные с **индикаторами достижения компетенций**:

Код и наименование компетенции	**Код и наименование индикатора достижения компетенции**	**Запланированные результаты обучения**
ПК-1 Способен разрабатывать системы и технические средства автоматизации и управления на основе современных программных и аппаратных средств	ИД-2$_{ПК-1}$ Формулирует критерии качества, разработки, настройки и тестирования алгоритмов анализа данных	знать: - методику проведения предварительной обработки и анализа текстовых данных; - методы интеллектуального анализа текстовых данных. уметь: - решать задачи интеллектуального анализа текстовых данных; - формировать выборки текстовых данных и приводить их к математическому виду.
ПК-1 Способен разрабатывать системы и технические средства автоматизации и управления на основе современных программных и аппаратных средств	ИД-4$_{ПК-1}$ Использует стандартное программное обеспечение и специализированные библиотеки для обработки и анализа данных	знать: - современные библиотеки и программные средства интеллектуального анализа данных. уметь: - использовать современные библиотеки и программные средства для разработки интеллектуальных систем.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ ВО

Дисциплина относится к основной профессиональной образовательной программе Системы и средства автоматизации, интеллектуального управления и анализа данных (далее – ОПОП), направления подготовки 27.03.04 Управление в технических системах, уровень образования: высшее образование - бакалавриат.

Базируется на уровне среднего общего образования.

Результаты обучения, полученные при освоении дисциплины, необходимы при выполнении выпускной квалификационной работы.

3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

3.1 Структура дисциплины

Общая трудоемкость дисциплины составляет 4 зачетных единицы, 144 часа.

№ п/п	Разделы/темы дисциплины/формы промежуточной аттестации	Всего часов на раздел	Семестр	Распределение трудоемкости раздела (в часах) по видам учебной работы										Содержание самостоятельной работы/ методические указания
				Контактная работа								СР		
				Лек	Лаб	Пр	Консультация		ИКР		ПА	Работа в семестре	Подготовка к аттестации /контроль	
							КПР	ГК	ИККП	ТК				
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Основы анализа данных	20	8	4	6	-	-	-	-	-	-	10	-	***Подготовка к лабораторной работе:*** Для выполнения заданий по лабораторной работе необходимо предварительно изучить тему и задачи выполнения лабораторной работы, а также изучить вопросы, связанные с особенностями задачи классификации данных и критериями качества классификации по изученному в разделе «Основы анализа данных» материалу. ***Подготовка к контрольной работе:*** Для выполнения контрольной работы необходимо изучить вопросы, связанные с задачей формирования выборок по изученному в разделе «Основы анализа данных» материалу. ***Изучение материалов литературных источников:*** [1], стр. 7-8 [2], стр. 24-25 [3], стр. 25-61
1.1	Введение в анализ данных и машинное обучение	6		1	2	-	-	-	-	-	-	3	-	
1.2	Введение в задачу классификации данных	7		2	2	-	-	-	-	-	-	3	-	
1.3	Особенности сформированных выборок	7		1	2	-	-	-	-	-	-	4	-	
2	Анализ текстовых данных (Text Mining)	18.0		2.0	6	-	-	-	-	-	-	10	-	***Подготовка к контрольной работе:*** Для выполнения контрольной работы необходимо изучить вопросы, связанные с задачей выявления информативных признаков по изученному в разделе «Основы анализа данных» материалу. ***Подготовка к лабораторной работе:*** Для выполнения заданий по лабораторной работе
2.1	Введение в Text Mining	5.5		0.5	2	-	-	-	-	-	-	3	-	
2.2	Особенности обработки текстовых данных	7		1	2	-	-	-	-	-	-	4	-	
2.3	Представление	5.5		0.5	2	-	-	-	-	-	-	3	-	

	текстовых данных в математическом виде													необходимо предварительно изучить тему и задачи выполнения лабораторной работы, а также изучить вопросы, связанные с особенностями обработки и анализа текстовых данных по изученному в разделе «Анализ текстовых данных (Text Mining)» материалу. ***Изучение материалов литературных источников:*** [1], стр. 15-16 [2], стр. 26-34
3	Задача классификации текстовых документов	29		8	6	-	-	-	-	-	-	15	-	***Подготовка к контрольной работе:*** Для выполнения контрольной работы необходимо изучить вопросы, связанные с методами классификации данных по изученному в разделе «Задача классификации текстовых документов» материалу. ***Подготовка к лабораторной работе:*** Для выполнения заданий по лабораторной работе необходимо предварительно изучить тему и задачи выполнения лабораторной работы, а также изучить вопросы, связанные с методами классификации текстовых документов по изученному в разделе «Задача классификации текстовых документов» материалу. ***Изучение материалов литературных источников:*** [1], стр. 16-23 [2], стр. 44-47, 52-61 [3], стр. 62-116
3.1	Методы классификации текстовых документов	18		4	4	-	-	-	-	-	-	10	-	
3.2	Создание коллективов решающих правил	7		2	2	-	-	-	-	-	-	3	-	
3.3	Нейросетевые подходы к классификации текстовых данных	4		2	-	-	-	-	-	-	-	2	-	
4	Другие задачи, решаемые в рамках Text Mining	41		10	6	-	-	-	-	-	-	25	-	***Подготовка к лабораторной работе:*** Для выполнения заданий по лабораторной работе необходимо предварительно изучить тему и задачи выполнения лабораторной работы, а также изучить вопросы, связанные с выявлением дубликатов и анализом тональностей текстов по изученному в
4.1	Кластеризация текстовых данных	12		3	3	-	-	-	-	-	-	6	-	
4.2	Выявление плагиата и авторства текста	11		2	3	-	-	-	-	-	-	6	-	

4.3	Выявление эмоциональной окраски текстов	8		2	-	-	-	-	-	-	-	6	-	разделе «Другие задачи, решаемые в рамках Text Mining.» материалу.
4.4	Применение нейронных сетей для решения задач Text Mining	10		3	-	-	-	-	-	-	-	7	-	***Подготовка к контрольной работе:*** Для выполнения контрольной работы необходимо изучить вопросы, связанные с методами выявления дубликатов текстов, анализом тональностей текстов и применением нейросетей в задачах анализа текстовых данных в разделе «Другие задачи, решаемые в рамках Text Mining» материалу. ***Подготовка к контрольной работе:*** Для выполнения контрольной работы необходимо изучить вопросы, связанные с методами кластеризации данных по изученному в разделе «Другие задачи, решаемые в рамках Text Mining» материалу. ***Подготовка к лабораторной работе:*** Для выполнения заданий по лабораторной работе необходимо предварительно изучить тему и задачи выполнения лабораторной работы, а также изучить вопросы, связанные с методами кластеризации данных по изученному в разделе «Другие задачи, решаемые в рамках Text Mining.» материалу. ***Изучение материалов литературных источников:*** [1], стр. 23-27 [3], стр. 256-270
	Экзамен	36.0		-	-	-	-	2	-	-	0.5	-	33.5	
	Всего за семестр	**144.0**		**24.0**	**24**	**-**	**-**	**2**	**-**	**-**	**0.5**	**60**	**33.5**	
	Итого за семестр	**144.0**		**24.0**	**24**	**-**	**2**		**-**		**0.5**	**93.5**		

Примечание: Лек – лекции; Лаб – лабораторные работы; Пр – практические занятия; КПР – аудиторные консультации по курсовым проектам/работам; ИККП – индивидуальные консультации по курсовым проектам/работам; ГК- групповые консультации по разделам дисциплины; СР – самостоятельная работа студента; ИКР – иная контактная работа; ТК – текущий контроль; ПА – промежуточная аттестация

3.2 Краткое содержание разделов

1. Основы анализа данных

1.1. Введение в анализ данных и машинное обучение

Типы задач, решаемых методами машинного обучения (МО). Формальная постановка задачи машинного обучения. Задачи анализа тестовой информации, решаемые методами МО.

1.2. Введение в задачу классификации данных

Постановка задачи классификации. Критерии качества: аккуратность, полнота, точность, F-мера, площадь под кривой ошибок, матрица неточностей. Способы формирование выборок. Обучающая, экзаменационная и тестовая выборки. Обучение моделей. Явление переобучения модели.

1.3. Особенности сформированных выборок

Несбалансированные классы. Методы борьбы с несбалансированностью: oversampling, undersampling. Специальные стратегии сэмплинга в условиях несбалансированных классов.

2. Анализ текстовых данных (Text Mining)

2.1. Введение в Text Mining

Text Mining и особенности задач, связанных с анализом текстов. Онтологии и тезаурусы. Статистический подход к анализу текстовой информации. Статистический подход к анализу текстовой информации.

2.2. Особенности обработки текстовых данных

Предварительная обработка данных: стемминг, лемматизация. Слова, не несущие информации. Выявление информативных признаков. Взвешивание как способ выявления информативных терминов. Статистический подход к выявлению информативных терминов. Теоретико-информационный подход к выявлению информативных терминов. Теоретико-информационный подход к выявлению информативных терминов.

2.3. Представление текстовых данных в математическом виде

Модель представления текстовых данных в математическом виде. Модель «Мешок слов» (Bag of words). Частично- и полностью структурированные модели.

3. Задача классификации текстовых документов

3.1. Методы классификации текстовых документов

Систематизация и обзор методов классификации. Метод ближайших соседей. Метод деревьев решений. Метод опорных векторов. Метод логистической регрессии. Наивный байесовский метод. Профильные методы классификации.

3.2. Создание коллективов решающих правил

Ансамблевые методы классификации. Оценка разнородности методов классификации. Бустинг, бэггинг. Метод случайного леса деревьев решений.

3.3. Нейросетевые подходы к классификации текстовых данных

Векторное представление модели Word2Vec. Алгоритм классификации Doc2Vec.

4. Другие задачи, решаемые в рамках Text Mining

4.1. Кластеризация текстовых данных

Задача кластеризации текстовых данных. EM-алгоритм кластеризации. Семейство алгоритмов k-means. Алгоритмы кластеризации FOREL, DBSCAN. Самоорганизующиеся карты Кохонена.

4.2. Выявление плагиата и авторства текста

Задача выявления плагиата. Виды дубликатов: полные дубликаты, явные дубликаты, нечеткие дубликаты. Коэффициент ассоциативности Жаккара. Семейство методов шинглов. Методы выявления дубликатов Winnowing, SpotSigs, I-Match, коэффициент Джаро-Винклера. Определение авторства текстов.

4.3. Выявление эмоциональной окраски текстов

Анализ тональности текста. Классификация по бинарной шкале, классификация по многополосной шкале. Особенности задачи анализа тональности – отрицание, сарказм. Подходы к определению тональности.

4.4. Применение нейронных сетей для решения задач Text Mining

Использование нейронных сетей в задачах анализа текстовых данных. Функции активации. Обучение нейронных сетей. Обучение нейронных сетей. Рекуррентные нейронные сети.

3.3. Темы практических занятий

не предусмотрено

3.4. Темы лабораторных работ

1. Лабораторная работа №4 «Решение задач анализа текстовых данных»;
2. Лабораторная работа №3 «Кластеризация текстовых документов»;
3. Лабораторная работа №2 «Классификация текстовых документов»;
4. Лабораторная работа №1 «Предварительная обработка текстовых данных».

3.5 Консультации

Текущий контроль (ТК)

1. Консультации направлены на обсуждение вопросов связанных с выполнением контрольных мероприятий по разделу «Основы анализа данных».
2. Консультации направлены на обсуждение вопросов связанных с выполнением контрольных мероприятий по разделу «Анализ текстовых данных (Text Mining)».
3. Консультации направлены на обсуждение вопросов связанных с выполнением контрольных мероприятий по разделу «Задача классификации текстовых документов».
4. Консультации направлены на обсуждение вопросов связанных с выполнением контрольных мероприятий по разделу «Другие задачи, решаемые в рамках Text Mining».

3.6 Тематика курсовых проектов/курсовых работ

Курсовой проект/ работа не предусмотрены

3.7. Соответствие разделов дисциплины и формируемых в них компетенций

Запланированные результаты обучения по дисциплине (в соответствии с разделом 1)	Коды индикаторов	Номер раздела дисциплины (в соответствии с п.3.1)				Оценочное средство (тип и наименование)
		1	2	3	4	
Знать:						
методы интеллектуального анализа текстовых данных	ИД-2$_{ПК-1}$			+	+	Лабораторная работа/Защита лабораторной работы №4
методику проведения предварительной обработки и анализа текстовых данных	ИД-2$_{ПК-1}$		+			Лабораторная работа/Защита лабораторной работы №1 Лабораторная работа/Защита лабораторной работы №2
современные библиотеки и программные средства интеллектуального анализа данных	ИД-4$_{ПК-1}$			+		Лабораторная работа/Защита лабораторной работы №3
Уметь:						
формировать выборки текстовых данных и приводить их к математическому виду	ИД-2$_{ПК-1}$	+	+			Тестирование/Контрольная работа №1
решать задачи интеллектуального анализа текстовых данных	ИД-2$_{ПК-1}$			+	+	Лабораторная работа/Защита лабораторной работы №3
использовать современные библиотеки и программные средства для разработки интеллектуальных систем	ИД-4$_{ПК-1}$			+	+	Лабораторная работа/Защита лабораторной работы №4

4. КОМПЕТЕНТНОСТНО-ОРИЕНТИРОВАННЫЕ ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ КОНТРОЛЯ ОСВОЕНИЯ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (ТЕКУЩИЙ КОНТРОЛЬ УСПЕВАЕМОСТИ, ПРОМЕЖУТОЧНАЯ АТТЕСТАЦИЯ ПО ДИСЦИПЛИНЕ)

4.1. Текущий контроль успеваемости

8 семестр

Форма реализации: Компьютерное задание

1. Защита лабораторной работы №1 (Лабораторная работа)
2. Защита лабораторной работы №2 (Лабораторная работа)
3. Защита лабораторной работы №3 (Лабораторная работа)
4. Защита лабораторной работы №4 (Лабораторная работа)
5. Контрольная работа №1 (Тестирование)

Балльно-рейтинговая структура дисциплины является приложением А.

4.2 Промежуточная аттестация по дисциплине

Экзамен (Семестр №8)

Оценка определяется в соответствии с Положением о балльно-рейтинговой системе для студентов НИУ «МЭИ» на основании семестровой и аттестационной составляющих.

В диплом выставляется оценка за 8 семестр.

Примечание: Оценочные материалы по дисциплине приведены в фонде оценочных материалов ОПОП.

5. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

5.1 Печатные и электронные издания:

1. Мохов, А. С. Анализ и обработка текстовых данных : практикум по курсу "Интеллектуальные информационные системы" по направлению подготовки магистров 27.04.04 "Управление в технических системах" / А. С. Мохов, В. О. Толчеев, А. А. Бородкин, Нац. исслед. ун-т "МЭИ" (НИУ"МЭИ") . – Москва : Изд-во МЭИ, 2020 . – 52 с. - ISBN 978-5-7046-2284-0 .
http://elib.mpei.ru/elib/view.php?id=11207;
2. Толчеев, В. О. Современные методы обработки и анализа текстовой информации : учебное пособие по курсу "Интеллектуальные информационные системы" по специальности "Управление и информатика в технических системах" / В. О. Толчеев, Моск. энерг. ин-т (МЭИ ТУ) . – М. : Изд-во МЭИ, 2006 . – 76 с. - ISBN 5-7046-1285-7 .;
3. Флах П.- "Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных", Издательство: "ДМК Пресс", Москва, 2015 - (400 с.) http://e.lanbook.com/books/element.php?pl1_id=69955.

5.2 Лицензионное и свободно распространяемое программное обеспечение:

1. Office / Российский пакет офисных программ;
2. Windows / Операционная система семейства Linux;
3. Python;
4. Jupiter Notebook.

5.3 Интернет-ресурсы, включая профессиональные базы данных и информационно-справочные системы:

1. **ЭБС Лань** - https://e.lanbook.com/
2. **Научная электронная библиотека** - https://elibrary.ru/
3. **База данных Web of Science** - http://webofscience.com/
4. **База данных Scopus** - http://www.scopus.com
5. **Национальная электронная библиотека** - https://rusneb.ru/
6. **Электронная библиотека МЭИ (ЭБ МЭИ)** - http://elib.mpei.ru/login.php
7. **Портал открытых данных Российской Федерации** - https://data.gov.ru
8. **База открытых данных Министерства труда и социальной защиты РФ** - https://rosmintrud.ru/opendata
9. **База открытых данных профессиональных стандартов Министерства труда и социальной защиты РФ** - http://profstandart.rosmintrud.ru/obshchiy-informatsionnyy-blok/natsionalnyy-reestr-professionalnykh-standartov/
10. **База открытых данных Министерства экономического развития РФ** - http://www.economy.gov.ru
11. **База открытых данных Росфинмониторинга** - http://www.fedsfm.ru/opendata
12. **Электронная открытая база данных "Polpred.com Обзор СМИ"** - https://www.polpred.com
13. **Национальный портал онлайн обучения «Открытое образование»** - https://openedu.ru
14. **Официальный сайт Федерального агентства по техническому регулированию и метрологии** - http://protect.gost.ru/
15. **Открытая университетская информационная система «РОССИЯ»** - https://uisrussia.msu.ru

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Тип помещения	Номер аудитории, наименование	Оснащение
Учебные аудитории для проведения лекционных занятий и текущего контроля	М-307, Учебная аудитория	стол преподавателя, стол учебный, стул, доска меловая, компьютерная сеть с выходом в Интернет, мультимедийный проектор, экран
	Ж-120, Машинный зал ИВЦ	сервер, кондиционер
Учебные аудитории для проведения практических занятий, КР и КП	М-307, Учебная аудитория	стол преподавателя, стол учебный, стул, доска меловая, компьютерная сеть с выходом в Интернет, мультимедийный проектор, экран
	Ж-120, Машинный зал ИВЦ	сервер, кондиционер
Учебные аудитории для проведения лабораторных занятий	М-304а/1, Учебная лаборатория моделирования систем и анализа данных	стол преподавателя, стол компьютерный, стул, компьютерная сеть с выходом в Интернет, доска маркерная, компьютер персональный
	Ж-120, Машинный зал ИВЦ	сервер, кондиционер
Учебные аудитории для проведения промежуточной аттестации	М-307, Учебная аудитория	стол преподавателя, стол учебный, стул, доска меловая, компьютерная сеть с выходом в Интернет, мультимедийный проектор, экран
	Ж-120, Машинный зал ИВЦ	сервер, кондиционер

Помещения для самостоятельной работы	НТБ-201, Компьютерный читальный зал	стол компьютерный, стул, стол письменный, вешалка для одежды, компьютерная сеть с выходом в Интернет, компьютер персональный, принтер, кондиционер
Помещения для консультирования	М-304а/2, Учебная лаборатория моделирования систем и анализа данных	кресло рабочее, стол преподавателя, стол учебный, стул, шкаф для документов, шкаф для одежды, компьютерная сеть с выходом в Интернет, компьютер персональный
Помещения для хранения оборудования и учебного инвентаря	М-309, Кладовая	стол, стул, шкаф для хранения инвентаря
	М-301/1, Кладовая	стул

Приложение А

БАЛЛЬНО-РЕЙТИНГОВАЯ СТРУКТУРА ДИСЦИПЛИНЫ

Основы анализа текстовых данных

(название дисциплины)

8 семестр

Перечень контрольных мероприятий <u>текущего контроля</u> успеваемости по дисциплине:
КМ-1 Контрольная работа №1 (Тестирование)
КМ-2 Защита лабораторной работы №1 (Лабораторная работа)
КМ-3 Защита лабораторной работы №2 (Лабораторная работа)
КМ-4 Защита лабораторной работы №3 (Лабораторная работа)
КМ-5 Защита лабораторной работы №4 (Лабораторная работа)

Вид промежуточной аттестации – Экзамен.

Номер раздела	Раздел дисциплины	Индекс КМ:	КМ-1	КМ-2	КМ-3	КМ-4	КМ-5
		Неделя КМ:	3	4	7	10	12
1	Основы анализа данных						
1.1	Введение в анализ данных и машинное обучение		+				
1.2	Введение в задачу классификации данных		+				
1.3	Особенности сформированных выборок		+				
2	Анализ текстовых данных (Text Mining)						
2.1	Введение в Text Mining		+	+	+		
2.2	Особенности обработки текстовых данных		+	+	+		
2.3	Представление текстовых данных в математическом виде		+	+	+		
3	Задача классификации текстовых документов						
3.1	Методы классификации текстовых документов					+	+
3.2	Создание коллективов решающих правил					+	+
3.3	Нейросетевые подходы к классификации текстовых данных					+	+
4	Другие задачи, решаемые в рамках Text Mining						
4.1	Кластеризация текстовых данных					+	+

4.2	Выявление плагиата и авторства текста				+	+
4.3	Выявление эмоциональной окраски текстов				+	+
4.4	Применение нейронных сетей для решения задач Text Mining				+	+
	Вес КМ, %:	10	20	20	20	30