

**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

**Оценочные материалы
по дисциплине
Основы анализа текстовых данных**

**Москва
2024**

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ РАЗРАБОТАЛ:

Разработчик

	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

ОБЩАЯ ЧАСТЬ

Оценочные материалы по дисциплине предназначены для оценки достижения обучающимися запланированных результатов обучения по дисциплине, этапа формирования запланированных компетенций и уровня освоения дисциплины.

Оценочные материалы по дисциплине включают оценочные средства для проведения мероприятий текущего контроля успеваемости и промежуточной аттестации.

Формируемые у обучающегося компетенции:

1. ПК-1 Способен разрабатывать системы и технические средства автоматизации и управления на основе современных программных и аппаратных средств
- ИД-2 Формулирует критерии качества, разработки, настройки и тестирования алгоритмов анализа данных
- ИД-4 Использует стандартное программное обеспечение и специализированные библиотеки для обработки и анализа данных

и включает:

для текущего контроля успеваемости:

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)

БРС дисциплины

8 семестр

Перечень контрольных мероприятий текущего контроля 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.

Раздел дисциплины	Веса контрольных мероприятий, %					
	Индекс КМ:	КМ-1	КМ-2	КМ-3	КМ-4	КМ-5
	Срок КМ:	3	4	7	10	12
Основы анализа данных						
Введение в анализ данных и машинное обучение		+				

Введение в задачу классификации данных	+				
Особенности сформированных выборок	+				
Анализ текстовых данных (Text Mining)					
Введение в Text Mining	+	+	+		
Особенности обработки текстовых данных	+	+	+		
Представление текстовых данных в математическом виде	+	+	+		
Задача классификации текстовых документов					
Методы классификации текстовых документов				+	+
Создание коллективов решающих правил				+	+
Нейросетевые подходы к классификации текстовых данных				+	+
Другие задачи, решаемые в рамках Text Mining					
Кластеризация текстовых данных				+	+
Выявление плагиата и авторства текста				+	+
Выявление эмоциональной окраски текстов				+	+
Применение нейронных сетей для решения задач Text Mining				+	+
Вес КМ:	10	20	20	20	30

СОДЕРЖАНИЕ ОЦЕНОЧНЫХ СРЕДСТВ ТЕКУЩЕГО КОНТРОЛЯ

I. Оценочные средства для оценки запланированных результатов обучения по дисциплине, соотнесенных с индикаторами достижения компетенций

Индекс компетенции	Индикатор	Запланированные результаты обучения по дисциплине	Контрольная точка
ПК-1	ИД-2 _{ПК-1} Формулирует критерии качества, разработки, настройки и тестирования алгоритмов анализа данных	Знать: методы интеллектуального анализа текстовых данных методику проведения предварительной обработки и анализа текстовых данных Уметь: формировать выборки текстовых данных и приводить их к математическому виду решать задачи интеллектуального анализа текстовых данных	КМ-1 Защита лабораторной работы №1 (Лабораторная работа) КМ-2 Защита лабораторной работы №2 (Лабораторная работа) КМ-3 Защита лабораторной работы №3 (Лабораторная работа) КМ-4 Защита лабораторной работы №4 (Лабораторная работа) КМ-5 Контрольная работа №1 (Тестирование)
ПК-1	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	КМ-3 Защита лабораторной работы №3 (Лабораторная работа) КМ-4 Защита лабораторной работы №4 (Лабораторная работа)

		интеллектуальных систем	
--	--	-------------------------	--

II. Содержание оценочных средств. Шкала и критерии оценивания

КМ-1. Контрольная работа №1

Формы реализации: Компьютерное задание

Тип контрольного мероприятия: Тестирование

Вес контрольного мероприятия в БРС: 10

Процедура проведения контрольного мероприятия: Выполнение тестовых заданий в СДО «Прометей».

Краткое содержание задания:

Дать правильные ответы на вопросы тестирования, связанные с формированием выборок, различиями разных типов выборок, постановкой задачи классификации данных и выявлением информативных признаков. Уметь применять методы взвешивания, рассчитывать показатели качества классификации.

Контрольные вопросы/задания:

Запланированные результаты обучения по дисциплине	Вопросы/задания для проверки
Уметь: формировать выборки текстовых данных и приводить их к математическому виду	1. Подсчитать полноту и точность классификации по тестовой выборке. 2. Выбрать информативные признаки для решения задачи классификации электронных писем.

Описание шкалы оценивания:

Оценка: 5 («отлично»)

Нижний порог выполнения задания в процентах: 85

Описание характеристики выполнения знания: Оценка "отлично" выставляется если задание выполнено в полном объеме или выполнено преимущественно верно.

Оценка: 4 («хорошо»)

Нижний порог выполнения задания в процентах: 65

Описание характеристики выполнения знания: Оценка "хорошо" выставляется если задание выполнено с небольшими ошибками.

Оценка: 3 («удовлетворительно»)

Нижний порог выполнения задания в процентах: 50

Описание характеристики выполнения знания: Оценка "удовлетворительно" выставляется если задание преимущественно выполнено, часть ответов дана с ошибками.

Оценка: 2 («неудовлетворительно»)

Описание характеристики выполнения знания: Оценка "неудовлетворительно" выставляется если задание выполнено неверно или преимущественно не выполнено, не даны ответы на дополнительные вопросы.

КМ-2. Защита лабораторной работы №1

Формы реализации: Компьютерное задание

Тип контрольного мероприятия: Лабораторная работа

Вес контрольного мероприятия в БРС: 20

Процедура проведения контрольного мероприятия: Выполнение индивидуального задания в программной среде Jupiter Notebook. Демонстрация выполнения работы программы с комментариями по реализации. Внесение изменений в программу в соответствии с индивидуальным дополнительным заданием. Демонстрация работы программы с внесенными изменениями.

Краткое содержание задания:

Защита лабораторной работы №1: «Предварительная обработка текстовых данных». Импортировать требуемые библиотеки. Загрузить выборку данных. Провести обработку данных и проанализировать влияние способов обработки на качество классификации.

Контрольные вопросы/задания:

Запланированные результаты обучения по дисциплине	Вопросы/задания для проверки
Знать: методику проведения предварительной обработки и анализа текстовых данных	1. Назовите основные этапы предварительной обработки данных. 2. Как влияет размер словаря терминов на точность классификации? 3. Какие способы выявления информативных терминов вам известны?

Описание шкалы оценивания:

Оценка: 5 («отлично»)

Нижний порог выполнения задания в процентах: 85

Описание характеристики выполнения знания: Оценка "отлично" выставляется если задание выполнено в полном объеме или выполнено преимущественно верно, даны правильные ответы на дополнительные вопросы.

Оценка: 4 («хорошо»)

Нижний порог выполнения задания в процентах: 65

Описание характеристики выполнения знания: Оценка "хорошо" выставляется если задание выполнено с небольшими ошибками, ответы на дополнительные вопросы преимущественно правильные.

Оценка: 3 («удовлетворительно»)

Нижний порог выполнения задания в процентах: 50

Описание характеристики выполнения знания: Оценка "удовлетворительно" выставляется если задание преимущественно выполнено, ответы на дополнительные вопросы неточные, неполные.

Оценка: 2 («неудовлетворительно»)

Описание характеристики выполнения знания: Оценка "неудовлетворительно" выставляется если задание выполнено неверно или преимущественно не выполнено, не даны ответы на дополнительные вопросы.

КМ-3. Защита лабораторной работы №2

Формы реализации: Компьютерное задание

Тип контрольного мероприятия: Лабораторная работа

Вес контрольного мероприятия в БРС: 20

Процедура проведения контрольного мероприятия: Выполнение индивидуального задания в программной среде Jupiter Notebook. Демонстрация выполнения работы программы с комментариями по реализации. Внесение изменений в программу в

соответствии с индивидуальным дополнительным заданием. Демонстрация работы программы с внесенными изменениями.

Краткое содержание задания:

Защита лабораторной работы №2: «Классификация текстовых документов».

Импортировать требуемые библиотеки. Загрузить выборку данных. Настроить и обучить методы классификации в соответствии с индивидуальным заданием. Проанализировать полученные результаты.

Контрольные вопросы/задания:

Запланированные результаты обучения по дисциплине	Вопросы/задания для проверки
Знать: методику проведения предварительной обработки и анализа текстовых данных	1. Алгоритм и особенности метода деревьев решений. 2. Что такое регуляризация?

Описание шкалы оценивания:

Оценка: 5 («отлично»)

Нижний порог выполнения задания в процентах: 85

Описание характеристики выполнения знания: Оценка "отлично" выставляется если задание выполнено в полном объеме или выполнено преимущественно верно, даны правильные ответы на дополнительные вопросы.

Оценка: 4 («хорошо»)

Нижний порог выполнения задания в процентах: 65

Описание характеристики выполнения знания: Оценка "хорошо" выставляется если задание выполнено с небольшими ошибками, ответы на дополнительные вопросы преимущественно правильные.

Оценка: 3 («удовлетворительно»)

Нижний порог выполнения задания в процентах: 50

Описание характеристики выполнения знания: Оценка "удовлетворительно" выставляется если задание преимущественно выполнено, ответы на дополнительные вопросы неточные, неполные.

Оценка: 2 («неудовлетворительно»)

Описание характеристики выполнения знания: Оценка "неудовлетворительно" выставляется если задание выполнено неверно или преимущественно не выполнено, не даны ответы на дополнительные вопросы.

КМ-4. Защита лабораторной работы №3

Формы реализации: Компьютерное задание

Тип контрольного мероприятия: Лабораторная работа

Вес контрольного мероприятия в БРС: 20

Процедура проведения контрольного мероприятия: Выполнение индивидуального задания в программной среде Jupiter Notebook. Демонстрация выполнения работы программы с комментариями по реализации. Внесение изменений в программу в соответствии с индивидуальным дополнительным заданием. Демонстрация работы программы с внесенными изменениями.

Краткое содержание задания:

Защита лабораторной работы №3: «Кластеризация текстовых документов».

Импортировать требуемые библиотеки. Загрузить выборку данных. Провести

кластеризацию данных иерархическими методами. Провести кластеризацию данных методом k-средних. Проанализировать полученные результаты.

Контрольные вопросы/задания:

Запланированные результаты обучения по дисциплине	Вопросы/задания для проверки
Знать: современные библиотеки и программные средства интеллектуального анализа данных	1.Что такое дендрограмма? 2.Алгоритм метода k-средних.
Уметь: решать задачи интеллектуального анализа текстовых данных	1.Как реализовать кластеризацию данных методом k-средних? 2.Как рассчитать среднее межкластерное расстояние?

Описание шкалы оценивания:

Оценка: 5 («отлично»)

Нижний порог выполнения задания в процентах: 85

Описание характеристики выполнения знания: Оценка "отлично" выставляется если задание выполнено в полном объеме или выполнено преимущественно верно, даны правильные ответы на дополнительные вопросы.

Оценка: 4 («хорошо»)

Нижний порог выполнения задания в процентах: 65

Описание характеристики выполнения знания: Оценка "хорошо" выставляется если задание выполнено с небольшими ошибками, ответы на дополнительные вопросы преимущественно правильные.

Оценка: 3 («удовлетворительно»)

Нижний порог выполнения задания в процентах: 50

Описание характеристики выполнения знания: Оценка "удовлетворительно" выставляется если задание преимущественно выполнено, ответы на дополнительные вопросы неточные, неполные.

Оценка: 2 («неудовлетворительно»)

Описание характеристики выполнения знания: Оценка "неудовлетворительно" выставляется если задание выполнено неверно или преимущественно не выполнено, не даны ответы на дополнительные вопросы.

КМ-5. Защита лабораторной работы №4

Формы реализации: Компьютерное задание

Тип контрольного мероприятия: Лабораторная работа

Вес контрольного мероприятия в БРС: 30

Процедура проведения контрольного мероприятия: Выполнение индивидуального задания в программной среде Jupiter Notebook. Демонстрация выполнения работы программы с комментариями по реализации. Внесение изменений в программу в соответствии с индивидуальным дополнительным заданием. Демонстрация работы программы с внесенными изменениями.

Краткое содержание задания:

Защита лабораторной работы №4: «Решение задач анализа текстовых данных».

Импортировать требуемые библиотеки. Загрузить выборку данных. Провести анализ уникальности текстов. Проанализировать полученные результаты.

Контрольные вопросы/задания:

Запланированные результаты обучения по дисциплине	Вопросы/задания для проверки
Знать: методы интеллектуального анализа текстовых данных	1.Что такое нечеткий дубликат? 2.Алгоритм метода шинглов.
Уметь: использовать современные библиотеки и программные средства для разработки интеллектуальных систем	1.Как реализовать проверку данных на уникальность? 2.Как рассчитать коэффициент Джаро-Винклера?

Описание шкалы оценивания:

Оценка: 5 («отлично»)

Нижний порог выполнения задания в процентах: 85

Описание характеристики выполнения знания: Оценка "отлично" выставляется если задание выполнено в полном объеме или выполнено преимущественно верно, даны правильные ответы на дополнительные вопросы.

Оценка: 4 («хорошо»)

Нижний порог выполнения задания в процентах: 65

Описание характеристики выполнения знания: Оценка "хорошо" выставляется если задание выполнено с небольшими ошибками, ответы на дополнительные вопросы преимущественно правильные.

Оценка: 3 («удовлетворительно»)

Нижний порог выполнения задания в процентах: 50

Описание характеристики выполнения знания: Оценка "удовлетворительно" выставляется если задание преимущественно выполнено, ответы на дополнительные вопросы неточные, неполные.

Оценка: 2 («неудовлетворительно»)

Описание характеристики выполнения знания: Оценка "неудовлетворительно" выставляется если задание выполнено неверно или преимущественно не выполнено, не даны ответы на дополнительные вопросы.

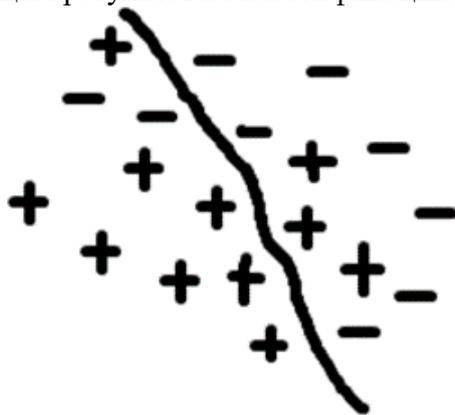
СОДЕРЖАНИЕ ОЦЕНОЧНЫХ СРЕДСТВ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

8 семестр

Форма промежуточной аттестации: Экзамен

Пример билета

1. Этапы предварительной обработки текстовых данных.
2. Профильные методы классификации: алгоритм, особенности.
3. Рассчитать показатель полноты, точности, F-меры относительно класса «+» по следующим результатам классификации:



Процедура проведения

Экзамен с оценкой проводится в устной форме по билетам. На подготовку ответа студенту отводится 45 минут.

I. Перечень компетенций/индикаторов и контрольных вопросов проверки результатов освоения дисциплины

1. Компетенция/Индикатор: ИД-2ПК-1 Формулирует критерии качества, разработки, настройки и тестирования алгоритмов анализа данных

Вопросы, задания

1. Особенности задачи классификации данных.
2. Особенности предварительной обработки текстовых данных.
3. Способы взвешивания терминов.

Материалы для проверки остаточных знаний

1. Этапы предварительной обработки текстовых данных:

Ответы:

Взвешивание терминов;

Стемминг\лемматизация;

Удаление стоп-слов (не несущих информации);

Удаление малоинформативных терминов;

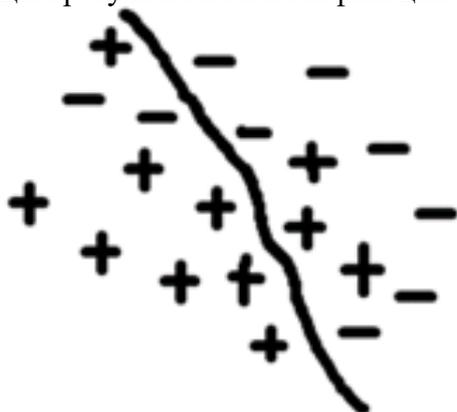
Токенизация текста (разбиение текста на слова\токены);

Верный ответ: 1)Токенизация текста (разбиение текста на слова\токены) 2)Удаление стоп-слов (не несущих информации) 3)Стемминг\лемматизация 4)Взвешивание терминов 5)Удаление малоинформативных терминов

2. Какой единственный настраиваемый параметр у профильных методов классификации?

Верный ответ: Длина профиля.

3. Рассчитать показатель полноты, точности, F-меры относительно класса «+» по следующим результатам классификации:



Верный ответ: $F1 = 2 \text{ Precision} * \text{Recall} / (\text{Precision} + \text{Recall}) = 2 * 0.8 * 0.73 / (0.8 + 0.73) = 0.76$

4. Даны два объекта, описываемые двумя признаками: $X1 = (2, 4)$; $X2 = (3, 1)$. Рассчитать евклидово расстояние и расстояния городских кварталов между ними.

Верный ответ: Воспользоваться формулой Евклидова расстояние. Ответ: $\sqrt{10} = 3.16$.

5. Даны два объекта, описываемые двумя признаками: $X1 = (2, 4)$; $X2 = (3, 1)$. Рассчитать расстояния городских кварталов между ними.

Верный ответ: Воспользоваться формулой Расстояние городских кварталов. Ответ: 4.

6. Когда вводится понятие «Отказ от классификации» при коллективной классификации?

Верный ответ: В случае, когда классификаторы не пришли к единому решению.

7. Значение какой меры близости увеличивается при увеличении степени схожести объектов?

Ответы:

Евклидово расстояние

Квадрат евклидова расстояния

Расстояние городских кварталов

Косинусная мера

Верный ответ: Косинусная мера

2. Компетенция/Индикатор: ИД-4_{ПК-1} Использует стандартное программное обеспечение и специализированные библиотеки для обработки и анализа данных

Вопросы, задания

1. Методы формирования выборок. Использование программных средств для формирования выборок.

2. Метод k-ближайших соседей: алгоритм, особенности.

3. Способы построения дендрограмм при решении задачи кластеризации. Особенности различных подходов.

Материалы для проверки остаточных знаний

1. Перечислите виды функций активации нейронов

Верный ответ: Ступенька; Сигмоида; Гиперболический тангенс; ReLU; Leaky ReLU.

2. Что подразумевается под дубликатом в широком смысле?

Ответы:

Дубликат – документ, идентичный по лексическому содержанию исходному

Дубликат – документ, идентичный по смысловому содержанию исходному

Дубликат – документ, содержащий текст из исходного

Верный ответ: Дубликат – документ, идентичный по смысловому содержанию исходному

3. Какие документы являются нечеткими дубликатами?

Ответы:

Один из документов пословно повторяет другой

В тексте одного из документов присутствуют блоки другого

Один из документов похож на другой с лексической точки зрения

Верный ответ: Один из документов похож на другой с лексической точки зрения

4. Способы выявления нечетких дубликатов:

Ответы:

Коэффициент Жаккара

Коэффициент корреляции

Метод К-Средних

Метод шинглов

Алгоритм FOREL

Хэш-функция

Метод Winnowing

Верный ответ: Коэффициент Жаккара Коэффициент корреляции Метод шинглов
Метод Winnowing

5. Для текста «Съешь еще этих мягких французских булок да выпей чаю» приведите 3 первых шингла длиной 4:

Верный ответ: Съешь еще этих еще этих мягких этих мягких французских

6. Даны два текста: «Съешь еще этих французских булок да выпей чаю» и «Съешь еще этих мягких французских булок да выпей чаю». Длина шингла = 3. Сколько совпадающих шинглов у этих текстов?

Верный ответ: 4

7. Какое расстояние при построении дендрограмм НЕ является монотонным?

Ответы:

Центроидов

Ближнего соседа

Дальнего соседа

Уорда

Верный ответ: Центроидов

8. Какое расстояние при построении дендрограмм является сжимающим?

Ответы:

Центроидов

Ближнего соседа

Дальнего соседа

Уорда

Верный ответ: Ближнего соседа

II. Описание шкалы оценивания

Оценка: 5 («отлично»)

Нижний порог выполнения задания в процентах: 85

Описание характеристики выполнения знания: Ответы на теоретические вопросы билета даны верно, задача из билета решена верно.

Оценка: 4 («хорошо»)

Нижний порог выполнения задания в процентах: 65

Описание характеристики выполнения знания: Ответы на теоретические вопросы билета даны в основном верно, задача из билета решена преимущественно верно.

Оценка: 3 («удовлетворительно»)

Нижний порог выполнения задания в процентах: 50

Описание характеристики выполнения знания: Ответы на теоретические вопросы билета даны с ошибками, неполные, задача из билета решена с грубыми ошибками.

Оценка: 2 («неудовлетворительно»)

Описание характеристики выполнения знания: Ответы на теоретические вопросы билета даны неверно, задача из билета решена неверно.

III. Правила выставления итоговой оценки по курсу

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