МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Утверждаю:
Зав. кафедрой компьютерных интеллектуальных технологий проектирования

_____ М.И. Чижов
«21» декабря 2021 г.

УЧЕБНО-МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ

«Системы хранения и обработки больших данных»

Направление подготовки: 09.04.01 Информатика и вычислительная техника

Направленность (профиль): Искусственный интеллект

Квалификация выпускника магистр

Нормативный период обучения 2 года / 2 года и 5 м.

Форма обучения очная / заочная

Год начала подготовки 2022

Составитель:
ПЕТРОВ Р.В., Д.Ф.-М.Н., ДОЦЕНТ, ЗАВЕДУЮЩИЙ КАФЕДРОЙ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И СИСТЕМ (НГУ)
ЖУРАВЛЕВА М.П., СТАРШИЙ ПРЕПОДАВАТЕЛЬ КАФ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И СИСТЕМ (НГУ)
Ершов Евгений Валентинович, д.т.н., профессор, директор института информационных технологий, зав. кафедрой МПО ЭВМ ЧГУ

г. Воронеж - 2021

Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля)

Основная литература:

1. Билл, Фрэнкс. Революция в аналитике [Электронный ресурс] : практическое руководство / Билл Фрэнкс ; Фрэнкс Билл; пер. И. Евстигнеева; ред. В. Мылов. - Революция в аналитике ; 2019-12-19. - Москва : Альпина Паблишер, 2017. - 320 с. - ISBN 978-5-9614-5302-7.

2. Воронова, Л.И. Big Data. Методы и средства анализа [Электронный ресурс] : учебное пособие / Л. И. Воронова, В. И. Воронов ; В.И. Воронов; Л.И. Воронова. - Big Data. Методы и средства анализа ; 2022-04-04. - Москва : Московский технический университет связи и информатики, 2016. - 33 с. URL: http://www.iprbookshop.ru/61463.html

3. Крутиков, В. Н. Анализ данных : учебное пособие / В. Н. Крутиков, В. В. Мешечкин ; В.Н. Крутиков; В.В. Мешечкин. - Кемерово : Кемеровский государственный университет, 2014. - 138 с. - ISBN 978-5-8353-1770-7. URL: http://biblioclub.ru/index.php?page=book&id=278426

Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины (модуля)

Ресурсы информационно-телекоммуникационной сети «Интернет»:
- http://www.edu.ru/
- Образовательный портал ВГТУ

Информационные справочные системы:
- http://window.edu.ru
- https://wiki.cchgeu.ru/

Учебно-методические указания и рекомендации к изучению тем лекционных и практических занятий, самостоятельной работе студентов

очная форма обучения

№ п/п	Наименование темы	Содержание раздела	Лекц	Прак зан.	Лаб. зан.	СРС	Всего, час
1	Введение в большие данные	Введение в большие данные	4	2	4	26	36
2	Жизненный цикл анализа больших данных	Жизненный цикл анализа больших данных	4	2	4	26	36

3	Корреляция и регрессия. Их роль в аналитике больших данных	Корреляция и регрессия. Их роль в аналитике больших данных	4	2	4	26	36
4	Задачи классификации и кластеризации. Ассоциативные правила	Задачи классификации и кластеризации. Ассоциативные правила	4	2	4	26	36
5	Языки Python и R, стек библиотек анализа данных. Готовые решения анализа данных (Weka и т.д.)	Языки Python и R, стек библиотек анализа данных. Готовые решения анализа данных (Weka и т.д.)	4	2	4	26	36
6	Переподготовка данных. Визуализация данных. Понимание данных	Переподготовка данных. Визуализация данных. Понимание данных	4	2	4	26	36
7	Парадигма Map Reduce. Ее реализация Hadoop	Парадигма Map Reduce. Ее реализация Hadoop	4	2	4	26	36
8	Проблема переобучения. Регуляризация. Нейронные сети. Машина опорных векторов	Проблема переобучения. Регуляризация. Нейронные сети. Машина опорных векторов	4	2	4	26	36
9	Научные проблемы в области больших данных	Научные проблемы в области больших данных	4	2	4	26	36
		Итого	**36**	**18**	**36**	**234**	**324**

заочная форма обучения

№ п/п	Наименование темы	Содержание раздела	Лекц	Прак зан.	Лаб. зан.	СРС	Всего, час
1	Введение в большие данные	Введение в большие данные	0,5	-	2	37	39,5
2	Жизненный цикл анализа больших данных	Жизненный цикл анализа больших данных	0,5	-	2	37	39,5
3	Корреляция и регрессия. Их роль в аналитике больших данных	Корреляция и регрессия. Их роль в аналитике больших данных	0,5	-	2	37	39,5
4	Задачи классификации и кластеризации. Ассоциативные правила	Задачи классификации и кластеризации. Ассоциативные правила	0,5	-	2	37	39,5
5	Языки Python и R, стек библиотек анализа данных. Готовые решения анализа данных (Weka и т.д.)	Языки Python и R, стек библиотек анализа данных. Готовые решения анализа данных (Weka и т.д.)	0,5	2	-	37	39,5
6	Переподготовка данных. Визуализация данных. Понимание данных	Переподготовка данных. Визуализация данных. Понимание данных	0,5	2	-	37	39,5
7	Парадигма Map Reduce. Ее реализация Hadoop	Парадигма Map Reduce. Ее реализация Hadoop	1			37	38
8	Проблема переобучения. Регуляризация. Нейронные сети. Машина опорных векторов	Проблема переобучения. Регуляризация. Нейронные сети. Машина опорных векторов	1			37	38
9	Научные проблемы в области больших данных	Научные проблемы в области больших данных	1			37	38
		Итого	**6**	**4**	**8**	**333**	**351**

Темы лабораторных работ

Лабораторная работа № 1. Основы работы с большими данными.

На основе открытых данных сформировать массив, отвечающий требованиям к большим данным. Структурировать информацию по типу. Предложить методы визуализации структурированных данных.

Контрольные вопросы:

1. Понятие «большие данные».

2. Характеристики больших данных.

3. Принципы работы с большими данными.

4. Как используются большие данные в научных исследованиях?

Лабораторная работа № 2. Основные стадии и этапы анализа больших данных. Разработать календарный план, детальный график и иерархическую структуру работ по анализу больших данных. Выполнить описание экосистемы больших данных, распределенной файловой системы, системы развертывания и интеграции данных. Предложить способы защиты данных. Обосновать необходимость применения методов машинного обучения.

Контрольные вопросы

1. Понятие экосистемы больших данных.

2. Назначение распределенной файловой системы.

3. Структура системы развертывания и интеграции данных.

4. Классификация методов машинного обучения.

Лабораторная работа № 3. Регрессионный анализ больших данных.
Выполнить регрессионный анализ сформированного в работе №1 массива больших данных. Предложить варианты визуального представления полученных результатов. Подготовить рекомендации по использованию результатов регрессионного анализа.

Контрольные вопросы 1. Задачи в области больших данных, решаемые методом

регрессионного анализа.

2. Методы сравнительного анализа. 3.
Сравнительный анализ больших данных.

Лабораторная работа № 4. Классификация и кластеризации больших данных.
Обосновать возможность классификации и кластеризации сформированного в работе №1 массива больших данных. Выполнить постановку задачи классификации и постановку задачи кластеризации.

Контрольные вопросы

1. Особенности построения ассоциативных правил.

2. Назначение классификации больших данных.

3. Назначение кластеризации больших данных.

4. Методы классификации и кластеризации больших данных.

Лабораторная работа № 5. Основы работы в Python и R.

Выполнить обзор функциональных возможностей библиотек и готовых решений анализа данных.

Контрольные вопросы

1. Роль готовых решений анализа данных в области больших данных.
2. Роль языков программирования Python и R в аналитике больших данных. 3. Отличительные особенности работы в Python и R.

Лабораторная работа № 6 Предварительная подготовка данных.

Выполнить предварительную подготовку данных на основе массива, сформированного в работе №1. Выполнить визуальное представление данных. Инструменты и методы визуализации данных

Контрольные вопросы 1. Методы предварительной подготовки данных.

2. Инструменты визуализации данных. 3. Оценка эффективности предварительной обработки данных.

Лабораторная работа № 7 Основы работы в Map Reduce.

Выполнить анализ функциональных возможностей Map Reduce, описание архитектуры Hadoop. Реализовать отдельные функции обработки сформированного в работе №1 массива данных. Сформировать отчет с применением встроенных средств визуализации.

Контрольные вопросы 1. Роль Map Reduce в аналитике больших данных.

2. Основные функциональные возможности Map Reduce. 3. Ограничения и особенности применения Map Reduce.

Лабораторная работа № 8 Принципы работы нейронных сетей.

Выполнить описание алгоритмов работы нейронных сетей. Обосновать целесообразность применения различных нейронных сетей для решения задач анализа больших данных.

Контрольные вопросы 1. Эффективность обучения нейронной сети.

2. Необходимость переобучения нейронной сети.
3. Регуляризация в нейронных сетях. 4. Оценка вычислительных мощностей при применении нейронных сетей.

Лабораторная работа № 9 Метод опорных векторов (SVM)

Выполнить разбор алгоритма SVM. Обосновать целесообразность применения для решения задач анализа больших данных.

Контрольные вопросы

1. Назначение алгоритма SVM.
2. Ограничения при применении алгоритма SVM.
3. Оценка вычислительных мощностей при применении алгоритма SVM.

Темы практических занятий

Практическое занятие № 1. Введение в большие данные.

Практическое занятие № 2. Жизненный цикл анализа больших данных.

Практическое занятие № 3. Корреляция и регрессия. Их роль в аналитике больших данных. Практическое занятие № 4. Задачи классификации и кластеризации. Ассоциативные правила.

Практическое занятие № 5. Языки Python и R, стек библиотек анализа данных. Готовые решения анализа данных (Weka и т.д.).

Практическое занятие № 6. Преподготовка данных. Визуализация данных. Понимание данных.

Практическое занятие № 7. Парадигма Map Reduce. Ее реализация Hadoop.

Практическое занятие № 8. Проблема переобучения. Регуляризация. Нейронные сети. Машина опорных векторов.

Практическое занятие № 9. Научные проблемы в области больших данных.

Темы самостоятельных работ

1 Соотношение аналоговой и цифровой информации
2 История больших данных
3 Определение больших данных

4 Характеристики Big Data

5 Принцип MapReduce

6 Технологии хранения

7 «Песочница» в аналитическом процессе

8 CRISP-DM

9 Hadoop

Средства контроля качества обучения

Вопросы к экзамену

1 Определение больших данных, ключевые характеристики. Примеры задач больших данных. Основные виды данных

2 Роль аналитика по данным (Data Scientist). Ключевые компетенции аналитика.

Отличия BI от Data Science

3 Корреляция и регрессионный анализ. Коэффициент корреляции. Графическое представление. Постановка задачи регрессионного анализа. Линейная регрессия. Метод наименьших квадратов. Привести примеры использования регрессионного анализа

4 Классификация. Признаковое описание объекта и таблица объект свойства. Постановка задачи. Отличия задачи классификации от задачи регрессии. Определение модели и алгоритма. Процесс обучения. Проблема переобучения. Регуляризация. Cross validation. Привести примеры использования алгоритмов классификации.

Дополнительный вопрос: привести модель в линейной регрессии

5 Кластеризация. Метрики. Матрица парных расстояний. Постановка задачи кластеризации. Отличие от задачи классификации. Привести примеры использования алгоритмов кластеризации

6 Ассоциативные правила. Определение. Достоверность и поддержка. Отличия построения ассоциативного правила от решающего правила задачи классификации.

Привести примеры использования ассоциативных правил

7 Парадигма Map Reduce. Описать принцип работы. Нарисовать диаграмму. Перечислить слабые и сильные стороны. Обозначить области применимости. Привести примеры использования

8 Визуализация. Дать определение визуализации. Показать важность визуализации в аналитике больших данных. Привести примеры использования визуализации

9 «Жизненный цикл» проекта по аналитике больших данных. Типовая архитектура проекта в области больших данных. Перечислить используемые технологии, указать степень вовлеченности каждой из технологий на каждом этапе работы над проектом. Перечислить основные роли исполнителей проекта

10 Научные проблемы больших данных. Показать значимость проблем,

актуальность, связь с областями математики и инженерии