Министерство науки и высшего образования РФ Ульяновский государственный университет	Форма	
Ф - Рабочая программа по дисциплине на основании ФГОС ВО		

УТВЕРЖДЕНО
решением Ученого совета факультета математики, информационных и авиационных технологий

от «16» мая 2023 г., протокол 4/23

Председатель _____М.А.Волков
подпись, расшифровка подписи

«16» __мая_____ 2023 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Дисциплина	Технологии хранения и обработки больших объемов информации
Факультет	Математики, информационных и авиационных технологий
Кафедра	Информационных технологий
Курс	1

Направление: **02.04.03 «**Математическое обеспечение и администрирование информационных систем (уровень магистратуры)»
код направления (специальности), полное наименование

Профиль: «Технология программирования»
полное наименование

Форма обучения:___очная
очная, заочная, очно-заочная (указать только те, которые реализуются)

Дата введения в учебный процесс УлГУ: « 01 » ___09____ 2023____ г.

Программа актуализирована на заседании кафедры: протокол №____ от _____ 20_____ г.

Сведения о разработчиках:

ФИО	Кафедра	Должность, ученая степень, звание
Шабалин Александр Станиславович	Информационных технологий	доцент, к.ф.-м.н

СОГЛАСОВАНО
Заведующий кафедрой информационных технологий, реализующей дисциплину/ Заведующий выпускающей кафедрой информационных технологий
_____/_____М.А.Волков_____/ *Подпись* *расшифровка подписи* «17»_____05___2023г.

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Целью освоения дисциплины «Технологии хранения и обработки больших объемов информации» является теоретическая и практическая подготовка студентов к работе с большими данными. Знания, полученные в результате освоения дисциплины, помогут при сборе и анализе огромных объемов структурированной или неструктурированной информации, при разработке моделей данных и получении новых знаний. Все это необходимо выпускнику, освоившему программу магистратуры, для решения различных задач практической и научно-исследовательской деятельности.

Задачи освоения дисциплины:

- приобретение студентами знаний о технологиях подготовки, хранения, обработки и анализа больших данных;
- применение статистических и математических методов для анализа больших объемов информации;
- приобретение практических навыков работы с методами Map Reduce.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП

Дисциплина «Технологии хранения и обработки больших объемов информации» относится к базовой части Блока Б1.В.02 «Дисциплины (модули)» Основной Образовательной Программы по направлению подготовки магистров 02.04.03 Математическое обеспечение и администрирование информационных систем.

3. ПЕРЕЧЕНЬ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ), СООТНЕСЕННЫХ С ПЛАНИРУЕМЫМИ РЕЗУЛЬТАТАМИ ОСВОЕНИЯ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Процесс изучения дисциплины направлен на формирование следующих компетенций:

Код и наименование реализуемой компетенции	Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с индикаторами достижения компетенций
способен использовать основные методы и средства автоматизации проектирования, реализации, испытаний и оценки качества при создании конкурентоспособного программного продукта и программных комплексов, а также способен использовать методы и средства автоматизации, связанные с сопровождением, администрированием и модернизацией программных продуктов и программных комплексов (ПК-5)	**Знать:** причины возникновения тренда больших данных; процессы анализа больших данных; основные подходы к обработке больших массивов данных; **Уметь:** формулировать алгоритмы; выбирать подходящий инструмент анализа больших данных; выбирать подходящую технологию хранения больших данных. **Владеть:** Современными инструментами работы с большими данными.
способен использовать знания направлений развития компьютеров с традиционной (нетрадиционной) архитектурой; современных системных программных средств, операционных систем, операционных и сетевых оболочек, сервисных программ; тенденции развития функций и	**Знать:** причины возникновения тренда больших данных; процессы анализа больших данных; основные подходы к обработке больших массивов данных; **Уметь:** формулировать алгоритмы; выбирать подходящий инструмент анализа больших данных; выбирать подходящую технологию хранения больших данных. **Владеть:** Современными инструментами работы с большими данными.

архитектур проблемно-ориентированных программных систем и комплексов в профессиональной деятельности (ПК-6)	
способен использовать основные концептуальные положения функционального, логического, объектно-ориентированного и визуального направлений программирования, методы, способы и средства разработки программ в рамках этих направлений (ПК-7)	**Знать:** причины возникновения тренда больших данных; процессы анализа больших данных; основные подходы к обработке больших массивов данных; **Уметь:** формулировать алгоритмы; выбирать подходящий инструмент анализа больших данных; выбирать подходящую технологию хранения больших данных. **Владеть:** Современными инструментами работы с большими данными.

4. ОБЪЕМ ДИСЦИПЛИНЫ

4.1. Объем дисциплины в зачетных единицах (всего) – 6 ЗЕ.

4.2. По видам учебной работы (в часах):

Вид учебной работы	Количество часов (форма обучения очная)			
	Всего по плану	В т.ч. по семестрам		
		1		
1	2	3	4	5
Аудиторные занятия:	36/36*	36/36*		
Лекции	18/18	18/18		
практические и семинарские занятия	-	-		
лабораторные работы (лабораторный практикум)	18/18	18/18		
Самостоятельная работа	144	144		
Текущий контроль (количество и вид: конт.работа, коллоквиум,реферат)	индивидуальное задание по л.р., задачи, опрос	индивидуальное задание по л.р., задачи, опрос		
Курсовая работа	-	-		
Виды промежуточной аттестации (экзамен, зачет)	Экзамен, 36	Экзамен, 36		
Всего часов по дисциплине	216	216		

**В случае необходимости использования в учебном процессе частично/исключительно дистанционных образовательных технологий в таблице через слеш указывается количество часов работы ППС с обучающимися для проведения занятий в дистанционном формате с применением электронного обучения*

4.3. Содержание дисциплины (модуля). Распределение часов по темам и видам учебной работы:

Форма обучения – *очная*

Название и разделов и тем	Всего	Виды учебных занятий					
		Аудиторные занятия				Самостоятельная работа	Форма текущего контроля знаний
		лекции	практические занятия, семинар	лабораторная работа*	в т.ч. занятия в интерактивной форме		
Тема 1. Вводный обзор: что такое Big Data и для чего нужен.	10	2				8	Устный опрос
Тема 2. Обзор реляционных баз данных.	68	2		6	6	54	Устный опрос, проверка решения задач
Тема 3. Предметно-ориентированные информационные базы данных.	10	2				8	Устный опрос
Тема 4. MapReduce: методология и технология распределенных вычислений.	10	2				8	Устный опрос
Тема 5. Введение в Hadoop.	10	2				8	Устный опрос
Тема 6. Обработка данных в реальном времени.	10	2		6	6	8	Устный опрос, проверка лабораторных работ
Тема 7. Массово-параллельная структура - Massive Parallel Processing.	10	2				8	Устный опрос
Тема 8. Вычисление дескриптивных статистик для больших объемов данных.	10	2				8	Устный опрос
Тема 9. Data Mining и Big Data.	42	2		6	6	34	Устный опрос, проверка лабораторных работ
Итого	216	18	-	18	18	144	

5. СОДЕРЖАНИЕ КУРСА

Тема 1. Вводный обзор: что такое Big Data и для чего нужен.
с каких объемов начинается Big Data
реляционные и нереляционные базы данных
потоки данных
Тема 2. Обзор реляционных баз данных.
SQL-сервер: основные принципы, примеры
NoSQL базы данных: обзор, примеры
Тема 3. Предметно-ориентированные информационные базы данных.
Data Warehausig
Тема 4. MapReduce: методология и технология распределенных вычислений.
Этап Map – предварительной обработки
Этап Reduce – свертки результатов
Примеры функций
Тема 5. Введение в Hadoop.
основные принципы Hadoop
компоненты Hadoop
работа с нереляционными данными
примеры использования
MapReduce в Hadoop
надстройки Hive и Pig
Тема 6. Обработка данных в реальном времени.
Storm,
Spark,
Impal
Тема 7. Массово-параллельная структура - Massive Parallel Processing
масштабирование реляционных баз данных
параллельное выполнение запросов к БД
архитектура Hub and Spoke
Тема 8. Вычисление дескриптивных статистик для больших объемов данных.
частоты,
средние,
стандартные отклонения,
медианы,
квартили
Тема 9. Data Mining и Big Data
кластеризация, сегментация, алгоритмы к-средних, EM - Expectation-maximization
иерархическая кластеризация
классификация данных
предиктивный анализ
регрессионные деревья
правила ассоциаций
machine learning.

6. ТЕМЫ ПРАКТИЧЕСКИХ И СЕМИНАРСКИХ ЗАНЯТИЙ

Не предусмотрены УП.

7. ЛАБОРАТОРНЫЕ РАБОТЫ (ЛАБОРАТОРНЫЙ ПРАКТИКУМ)

Лабораторная работа 1

Тема: работа с набором данных

Цель работы: получение практических навыков работы с библиотеками для работы с данными (предварительного анализа данных) на языке Python.

Задание: используя программу Jupiter Notebook или его альтернативу, язык программирования Python, библиотеки OS, IO, Pandas, Pandas-Profiling, AutoViz, нужно сделать следующее:

1) загрузить набор данных согласно варианту;
2) получить информацию о наборе данных и данные из набора;
3) обработать пустые значения и дубликаты (при отсутствии таких создать второй набор, в котором удалить и продублировать часть данных и выполнить эту часть задания с ним);
4) провести конструирование признаков набора данных, используя различные способы изменения состава столбцов (сделать не менее 3 разных наборов с разным составом признаков, вставить объяснения, почему появились или были удалены признаки);
5) сгенерировать новый набор данных, часть данных в котором будет из первого набора, и выполнить все возможные операции объединения данных и заполнения наборов данными с учётом другого набора;
6) выполнить не менее 5 различных операций группировки и агрегации(использовать разные методы);
7) придумать новые признаки в наборе (не менее 3);
8) в одном из новых наборов данных создать составной индекс;
9) найти категориальные признаки и произвести их кодирование (не менее чем 2 способами);
10) получить статистический данных о наборе;
11) построить по одному из полей исходного набора гистограмму, диаграммы рассеивания, диаграмму «ящиков с усиками», используя библиотеку Pandas;
12) построить интерактивный отчёт, используя библиотеку Pandas-Profiling;
13) построить графики по прогнозируемому параметру, используя библиотеку AutoViz.

Отчёт по лабораторной работе должен содержать:

1. Фамилию и номер группы учащегося, задание, вариант.
2. Описание набора данных.
3. Протокол выполнения работы со всеми задачам.
4. Выводы.
5. Код.

Варианты

1 https://www.kaggle.com/benroshan/factors-affecting-campus-placement
2 https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists
3 https://www.kaggle.com/sl6149/data-scientist-job-market-in-the-us
4 https://www.kaggle.com/malekzadeh/motionsense-dataset
5 https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018
6 https://www.kaggle.com/christianlillelund/passenger-list-for-the-estonia-ferry-disaster
7 https://www.kaggle.com/mosemet/south-african-powerball-results-lottery
8 https://www.kaggle.com/monogenea/birdsongs-from-europe
9 https://www.kaggle.com/olgabelitskaya/svhn-preproccessed-fragments
10 https://www.kaggle.com/yasserh/breast-cancer-dataset
11 https://www.kaggle.com/iabhishekofficial/mobile-price-classification
12 https://www.kaggle.com/yasserh/heart-disease-dataset
13 https://www.kaggle.com/kukuroo3/body-performance-data
14 https://www.kaggle.com/fedesoriano/heart-failure-prediction
15 https://www.kaggle.com/imakash3011/customer-personality-analysis
16 https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
17 https://www.kaggle.com/shivamb/machine-predictive-maintenance-classification
18 https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset
19 https://www.kaggle.com/affanazhar/covid19-daily-data-updated
20 https://www.kaggle.com/mitishaagarwal/patient

Выполнение лабораторной можно делать:

1) На своём компьютере Jupiter Notebook (необходимо скачать набор данных);
2) www.kaggle.com (набор доступен через Add data);
3) https://colab.research.google.com/ (необходимо загрузить данные в среду).

Лабораторная работа 2

Тема: регрессионные модели

Цель работы: получение практических навыков построения и использования

регрессионных моделей на языке Python с использованием библиотек Scikit-Learn и StatsModels.

Задание: используя программу Jupiter Notebook, язык программирования Python, библиотеки Scikit-Learn, StatsModels, NumPy, Matplotlib и др. выполнить следующие задания:

- **Парная регрессия**: построить две реализации парной линейной регрессионной модели на базе 2 библиотек Scikit-Learn, StatsModels, сравнить и интерпретировать полученные результаты, входные данные рассчитать согласно варианту в таблице.
- **Множественная регрессия**: для своего варианта провести регрессионное моделирование (построить множественную регрессионную модель, ссылка для скачки данных на странице в разделе Data tables, выбрать не менее 50 строк):
  - выбрать выходную прогнозируемую переменную,
  - построить регрессионную модель со значимыми параметрами (оценить корреляции между факторам, последовательно добавлять факторы и сравнивать качество получаемых моделей, подобрать вид функции (визуальный анализ), оценить адекватность модели по статистическим показателям, каждый из этапов прокомментировать в отчете),
  - интерпретируете результаты моделирования (что значит полученная формула, какие переменные вносят больший вклад, что будет при изменении независимых переменных с зависимой),
  - прогнозировать новые значения с помощью построенной модели.

Отчёт по лабораторной работе должен содержать:

1. Фамилию и номер группы учащегося, задание, вариант.
2. Описание полученных регрессионных моделей.
3. Протокол построения и использования регрессионных моделей.
4. Сравнительный анализ моделей по первой задаче.
5. Интерпретация результатов по второй задаче.
6. Код.

Варианты

Вариант	Парная	Множественная
1.	Y=3*x +random(5)	Набор данных https://dataportal.orr.gov.uk/statistics/passenger-experience/delay-compensation-claims/
2.	Y=5*x + random(10)	Набор данных https://dataportal.orr.gov.uk/statistics/usage/freight-rail-usage-and-performance/
3.	Y=4*x + random(13)	Набор данных https://dataportal.orr.gov.uk/statistics/passenger-experience/disabled-persons-railcards/
4.	Y=2*x + random(4)	Набор данных https://dataportal.orr.gov.uk/statistics/passenger-experience/passenger-rail-service-complaints/
5.	Y=6.8*x + random(7)	Набор данных https://dataportal.orr.gov.uk/statistics/passenger-experience/passenger-assistance/
6.	Y=3.3*x +7- random(9)	Набор данных https://dataportal.orr.gov.uk/statistics/passenger-experience/passenger-satisfaction-complaints-handling/
7.	Y=8.3*x +7 random(10)	Набор данных https://dataportal.orr.gov.uk/statistics/performance/passenger-rail-performance/
8.	Y=2*x +random(5)	Набор данных https://dataportal.orr.gov.uk/statistics/finance/rail-fares/
9.	Y=2.5*x + random(10)	Набор данных https://dataportal.orr.gov.uk/statistics/finance/rail-industry-finance/
10.	Y=1.4*x + random(13)	Набор данных https://dataportal.orr.gov.uk/statistics/health-and-safety/rail-safety/
11.	Y=2.3*x + random(4)	Набор данных https://dataportal.orr.gov.uk/statistics/health-and-safety/occupational-health/
12.	Y=1.8*x + random(7)	Набор данных https://dataportal.orr.gov.uk/statistics/infrastructure-and-emissions/rail-infrastructure-and-assets/
13.	Y=25.3*x -7- random(9)	Набор данных https://dataportal.orr.gov.uk/statistics/infrastructure-and-emissions/rail-emissions/
14.	Y=81.3*x -7 random(10)	Набор данных https://dataportal.orr.gov.uk/statistics/infrastructure-and-emissions/asset-condition/
15.	Y=2*x - random(4)	Набор данных https://dataportal.orr.gov.uk/statistics/finance/rail-industry-finance/
16.	Y=6.8*x - random(7)	Набор данных https://dataportal.orr.gov.uk/statistics/health-and-safety/rail-safety/
17.	Y=3.3*x -7- random(9)	Набор данных https://dataportal.orr.gov.uk/statistics/health-and-safety/occupational-health/
18.	Y=8.3*x - random(10)	Набор данных https://dataportal.orr.gov.uk/statistics/infrastructure-and-emissions/rail-infrastructure-and-assets/
19.	Y=2*x -random(5)	Набор данных https://dataportal.orr.gov.uk/statistics/infrastructure-and-emissions/rail-emissions/
20.	Y=2.5*x - random(10)	Набор данных https://dataportal.orr.gov.uk/statistics/infrastructure-and-emissions/asset-condition/

Лабораторная работа 3

Тема: обработка естественного языка (NLP)

Цель работы: получение практических навыков использования генетических алгоритмов на языке Python с использованием библиотеки Natasha.

Задание: используя программу Jupiter Notebook, язык программирования Python, библиотеку Natasha реализовать предварительную обработку текста на русском языке, выполнив задачи согласно варианту.

Отчёт по лабораторной работе должен содержать:

1. Фамилию и номер группы учащегося, задание, вариант.
2. Алгоритм решения задачи.
3. Результаты обработки текста.
4. Код.
5. Обрабатываемый текст на русском языке (найти подходящий или сгенерировать согласно заданию).

Варианты

№	Текст	Задание
1	Любой художественный рассказ	Извлечь все прилагательные из текста и для каждого прилагательного вывести список существительных, с которыми оно употреблялось (в нормализованном виде).
2	Любой художественный рассказ	Подсчитать количество предложений, слов, глаголов, существительных, сколько уникальных глаголов и существительных в тексте, вывести их списки.
3	https://histrf.ru/read/biographies/ivan-iv-groznyi	Извлечь все персоны и сопоставить им все глаголы, с которыми они были связаны (в нормальной форме), подсчитать частоту связанных глаголов.
4	https://ria.ru/20130304/925668903.html	Извлечь «тройки» дата – глагол – персона (упомянутые в одном предложении).
5	http://inmotion.live/notes/mysql-story/	Сопоставить организации и персоны.
6	Любой художественный рассказ	Подсчитать для каждой части речи, сколько уникальных слов было в тексте.
7	В тексте должно быть упоминание не менее 5 валют.	Подсчитать общую сумму денежных средств упомянутых в тексте с учётом курса валют.
8	https://www.rusempire.ru/istoriya-rossii-kratko.html	Для каждого века сделать список персон.
9	https://www.rusempire.ru/istoriya-rossii-kratko.html	Для каждого десятилетия сделать список локаций.
10	https://www.rusempire.ru/istoriya-rossii-kratko.html	Найти все уникальные имена и отчества.

Лабораторная работа 4

Тема: рекомендательные системы

Цель работы: получение практических навыков построения рекомендательных систем на языке Python с использованием библиотеки Surprise.

Задание: используя программу Jupiter Notebook, язык программирования Python, библиотеку Surprise и др.:

1) загрузить набор данных согласно варианту, преобразовать данные в случае необходимости в соответствующий вид,
2) использовать метод согласно варианту для получения рекомендаций (прогнозных рейтингов),
3) получить значения оценок модели прогноза и интерпретировать результат,
4) вывести запрашиваемый в варианте результат (написать функцию с соответствующими входными параметрами и выводом, привести в отчёте 3 результата вызова функции с разными параметрами).

Отчёт по лабораторной работе должен содержать:

1. Фамилию и номер группы учащегося, задание, вариант.
2. Описание полученного набора данных.
3. Полное описание метода из варианта (алгоритм/формулы, выдаваемые значения, их интерпретация).
4. Пример вычислений (в ручную пошагово) по данным и методу из варианта (обязательно).
5. Скрины выполнения программы.
6. Интерпретация результатов (объяснение на конкретных данных)
7. Код с комментариями.

Варианты

Вариант	**Набор данных (закачать любой подходящий набор данных с ресурса)**	**Метод прогноза**	**Вывод**
1	***MovieLens 25M Dataset*** – набор рейтинговых данных с веб-сайта MovieLens, который описывает 5-звездочные рейтинги и действия с произвольным тегированием по более 60 тысячам фильмов от 1,5 миллионов пользователем с 1995 по 2019 годы. https://grouplens.org/datasets/movielens/	random_pred.NormalPredictor	Топ-10 рекомендованных объектов (товаров) по пользователю
2	***Netflix Prize*** - многовариантный датасет временных рядов, который использовался в конкурсе Netflix Prize с рейтингами примерно 100 миллионов фильмов. В наборе данных более 480000 пользователей, каждый из которых промаркирован уникальным целочисленным идентификатором.	baseline_only.BaselineOnly	5 наиболее похожих пользователей для заданного пользователя

	http://academictorrents.com/details/9b13183dc4d60676b773c9e2cd6de5e5542cee9a		
3	***Book-Crossing*** – датасет с рейтингами около 300 тысяч миллионов книг и обезличенными демографическими данными о более 250 тысячах их читателей. http://www2.informatik.uni-freiburg.de/~cziegler/BX/	knns.KNNBasic	Топ-10 товаров (объектов), конкретного пользователя (по реальному рейтингу)
4	***Amazon Review Data*** – многомиллионный набор обзоров, рейтингов и метаданных продуктов (описание, категория, цена, бренд, характеристики, фото), а также данные о просмотре ссылок. https://nijianmo.github.io/amazon/index.html	knns.KNNWithMeans	Топ-10 рекомендованных объектов (товаров) по пользователю
5	***REKKO CHALLENGE*** – набор данных от онлайн-кинотеатра OKKO для конкурса по разработке рекомендательных систем 2019 года. https://boosters.pro/championship/rekko_challenge/data	knns.KNNBaseline	5 наиболее похожих пользователей для заданного пользователя
6	***LastFM*** – датасет содержит информацию о социальных сетях, тегах и прослушивании музыкальных исполнителей от 2 тысяч пользователей онлайн-музыки Last.fm. https://files.grouplens.org/datasets/hetrec2011/	matrix_factorization.SVD	Топ-10 товаров (объектов), конкретного пользователя (по реальному рейтингу)
7	***Social Network Influencer*** – датасет Peerindex, который включает стандартную задачу изучения парных предпочтений. Здесь каждая точка данных описывает двух человек и предварительно рассчитанные стандартизованные функции на основе активности в Twitter: объем взаимодействий, количество подписчиков и пр. для каждого человека. https://www.kaggle.com/c/predict-who-is-more-influential-in-a-social-network/data	matrix_factorization.SVDpp	Топ-10 рекомендованных объектов (товаров) по пользователю
8	***Million Song Dataset*** – набор звуковых фич и метаданных для миллиона современных музыкальных треков от Echo Nest. http://millionsongdataset.com/	matrix_factorization.NMF	5 наиболее похожих пользователей для заданного пользователя
9	***Free Music Archive (FMA)*** – набор легальных аудиозаписей для задач анализа музыки - просмотр, поиск и организация коллекций. https://github.com/mdeff/fma	slope_one.SlopeOne	Топ-10 товаров (объектов), конкретного пользователя (по реальному рейтингу)
10	***Steam Video Games*** - набор данных о действиях пользователей и их характеристиках от самого популярного хаба видеоигр, PC Gaming Steam https://www.kaggle.com/tamber/steam-video-games/data	co_clustering.CoClustering	Топ-10 рекомендованных объектов (товаров) по пользователю
11	***Ta-Feng*** – набор данных о покупках от ACM RecSys по 23+ тысяч товаров, от продуктов питания и канцелярских товаров до мебели. http://www.bigdatalab.ac.cn/benchmark/bm/dd?data=Ta-Feng	random_pred.NormalPredictor	5 наиболее похожих пользователей для заданного пользователя
12	***Beiren*** – данные о реальных покупках более миллиона человек в супермаркетах Китая за период с 2012 по 2013 год. http://www.bigdatalab.ac.cn/benchmark/bm/dd?data=Beiren	baseline_only.BaselineOnly	Топ-10 товаров (объектов), конкретного пользователя (по реальному рейтингу)

13	***MovieLens 25M Dataset*** – набор рейтинговых данных с веб-сайта MovieLens, который описывает 5-звездочные рейтинги и действия с произвольным тегированием по более 60 тысячам фильмов от 1,5 миллионов пользователем с 1995 по 2019 годы. https://grouplens.org/datasets/movielens/	knns.KNNBasic	Топ-10 рекомендованных объектов (товаров) по пользователю
14	***Jester*** - Анонимные данные о рейтингах шуток (анекдотов) из системы Jester. https://goldberg.berkeley.edu/jester-data/	knns.KNNWithMeans	5 наиболее похожих пользователей для заданного пользователя
15	***REKKO CHALLENGE*** – набор данных от онлайн-кинотеатра OKKO для конкурса по разработке рекомендательных систем 2019 года. https://boosters.pro/championship/rekko_challenge/data	knns.KNNBaseline	Топ-10 товаров (объектов), конкретного пользователя (по реальному рейтингу)
16	***MovieLens 25M Dataset*** – набор рейтинговых данных с веб-сайта MovieLens, который описывает 5-звездочные рейтинги и действия с произвольным тегированием по более 60 тысячам фильмов от 1,5 миллионов пользователем с 1995 по 2019 годы. https://grouplens.org/datasets/movielens/	matrix_factorization.SVD	Топ-10 рекомендованных объектов (товаров) по пользователю
17	***Netflix Prize*** - многовариантный датасет временных рядов, который использовался в конкурсе Netflix Prize с рейтингами примерно 100 миллионов фильмов. В наборе данных более 480000 пользователей, каждый из которых промаркирован уникальным целочисленным идентификатором. http://academictorrents.com/details/9b13183dc4d60676b773c9e2cd6de5e5542cee9a	matrix_factorization.SVDpp	5 наиболее похожих пользователей для заданного пользователя
18	***Book-Crossing*** – датасет с рейтингами около 300 тысяч миллионов книг и обезличенными демографическими данными о более 250 тысячах их читателей. http://www2.informatik.uni-freiburg.de/~cziegler/BX/	matrix_factorization.NMF	Топ-10 товаров (объектов), конкретного пользователя (по реальному рейтингу)
19	***MovieLens 25M Dataset*** – набор рейтинговых данных с веб-сайта MovieLens, который описывает 5-звездочные рейтинги и действия с произвольным тегированием по более 60 тысячам фильмов от 1,5 миллионов пользователем с 1995 по 2019 годы. https://grouplens.org/datasets/movielens/	slope_one.SlopeOne	Топ-10 рекомендованных объектов (товаров) по пользователю
20	***Netflix Prize*** - многовариантный датасет временных рядов, который использовался в конкурсе Netflix Prize с рейтингами примерно 100 миллионов фильмов. В наборе данных более 480000 пользователей, каждый из которых промаркирован уникальным целочисленным идентификатором. http://academictorrents.com/details/9b13183dc4d60676b773c9e2cd6de5e5542cee9a	co_clustering.CoClustering	5 наиболее похожих пользователей для заданного пользователя

* При несогласованности указаний в варианте, сделать соответствующие пояснения и изменения условий.

8. ТЕМАТИКА КУРСОВЫХ, КОНТРОЛЬНЫХ РАБОТ, РЕФЕРАТОВ

Курсовые и контрольные работы, рефераты не предусмотрены учебным планом.

9. ПЕРЕЧЕНЬ ВОПРОСОВ К ЭКЗАМЕНУ

1.	Технологии BigData: дать определение для BigData, назначение BigData, история появления и основные принципы BigData. Достоинства и недостатки BigData.
2.	Технологии управления знаниями, визуализации знаний и интеллектуальные карты. Дать определение понятиям, назначение технологии, привести примеры программного обеспечения для визуализации знаний и построения интеллектуальных карт.
3.	Данные, информация, знания, модели. Наука о данных.
4.	Эволюционное развитие архитектур и данных.
5.	Критерии больших данных. Источники больших данных. Интернет вещей. Робототехника.
6.	Возможные этапы работы с большими данными.
7.	Примеры и истории успеха работы с большими данными: торговля, финансы, кадры.
8.	Обзор подходов к работе с данными: от языка простых запросов до методов анализа больших данных.
9.	Интеллектуальный анализ данных: краткий обзор подходов.
10.	Генетические алгоритмы.
11.	Деревья принятия решений.
12.	Визуализация больших данных.
13.	Специфика хранения и обработки больших данных.
14.	Парадигма MapReduce
15.	Файловая система HDFS.
16.	Особенности хранилищ данных NoSQL.
17.	Архитектура высоконагруженных систем.

10. САМОСТОЯТЕЛЬНАЯ РАБОТА СТУДЕНТОВ

Название разделов и тем	Вид самостоятельной работы	Объем в часах	Форма контроля
Тема1.	Проработка учебного материала, лабораторные работы, подготовка к сдаче зачета	8	Проверка домашних и лабораторных работ, заданий, сообщений и др.
Тема 2.	Проработка учебного материала, лабораторные работы, подготовка к сдаче зачета	54	Проверка домашних и лабораторных работ, заданий, сообщений и др.
Тема3.	Проработка учебного материала, лабораторные работы, подготовка к сдаче зачета	8	Проверка домашних и лабораторных работ, заданий, сообщений и др.
Тема 4.	Проработка учебного материала, лабораторные работы, подготовка к сдаче зачета	8	Проверка домашних и лабораторных работ, заданий, сообщений и др.
Тема 5.	Проработка учебного материала, лабораторные работы, подготовка к сдаче зачета	8	Проверка домашних и лабораторных работ, заданий, сообщений и др.

Тема 6.	Проработка учебного материала, лабораторные работы, подготовка к сдаче зачета	8	Проверка домашних и лабораторных работ, заданий, сообщений и др.
Тема 7.	Проработка учебного материала, лабораторные работы, подготовка к сдаче зачета	8	Проверка домашних и лабораторных работ, заданий, сообщений и др.
Тема 8.	Проработка учебного материала, лабораторные работы, подготовка к сдаче экзамена	8	Проверка домашних и лабораторных работ, заданий, сообщений и др.
Тема 9.	Проработка учебного материала, лабораторные работы, подготовка к сдаче зачета	34	Проверка домашних и лабораторных работ, заданий, сообщений и др.

По данной дисциплине организуется и проводится внеаудиторная самостоятельная работа.

Самостоятельная работа студентов, предусмотренная учебным планом в объеме не менее 50-70% общего количества часов, должна соответствовать более глубокому усвоению изучаемого курса, формировать навыки исследовательской работы и ориентировать студентов на умение применять теоретические знания на практике.

Самостоятельная работа по данной дисциплине состоит из следующих модулей:

- подготовка к лабораторным занятиям;
- подготовка к экзамену.

При подготовке к лабораторным занятиям и контрольным мероприятиям рекомендуется руководствоваться учебниками и учебными пособиями, в том числе и информацией, полученной в INTERNET.

Задания для самостоятельной работы требует дополнительной проработки и анализа рассматриваемого преподавателем материала в объеме запланированных часов.

Задания по самостоятельной работе оформлены в виде таблицы с указанием конкретного вида самостоятельной работы:

- проработка учебного материала (по конспектам лекций учебной и научной литературе) и подготовка лабораторным занятиям;
- поиск и обзор научных публикаций и электронных источников информации;

Студентам рекомендуется следующий порядок организации самостоятельной работы над темами и подготовки к практическим занятиям:

- ознакомиться с содержанием темы;
- прочитать материал лекций, при этом нужно составить себе общее представление об излагаемых вопросах;
- прочитать параграфы учебника, относящиеся к данной теме;
- перейти к тщательному изучению материала, усвоить теоретические положения и выводы, при этом нужно записывать основные положения темы (формулировки, определения, термины, воспроизводить отдельные схемы и чертежи из учебника и конспекта лекций);

РЕЗУЛЬТАТЫ САМОСТОЯТЕЛЬНОЙ РАБОТЫ КОНТРОЛИРУЮТСЯ ПРЕПОДАВАТЕЛЕМ И УЧИТЫВАЮТСЯ ПРИ АТТЕСТАЦИИ СТУДЕНТА (ЭКЗАМЕН).

11. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Список рекомендуемой литературы

а) *Основная литература*

1.Жуковский, О. И. Информационные технологии и анализ данных [Электронный ресурс] : учебное пособие / О. И. Жуковский. — Электрон. текстовые данные. — Томск : Томский государственный университет систем управления и радиоэлектроники, Эль Контент, 2014. — 130 c. — 978-5-4332-0158-3. — Режим доступа: http://www.iprbookshop.ru/72106.html

2. Воронова, Л. И. Machine Learning: регрессионные методы интеллектуального анализа данных [Электронный ресурс] : учебное пособие / Л. И. Воронова, В. И. Воронов. — Электрон. текстовые данные. — М. : Московский технический университет связи и информатики, 2018. — 82 c. — 2227-8397. — Режим доступа: http://www.iprbookshop.ru/81325.html

Дополнительная литература:

1.1 Федин, Ф. О. Анализ данных. Часть 1. Подготовка данных к анализу [Электронный ресурс] : учебное пособие / Ф. О. Федин, Ф. Ф. Федин. — Электрон. текстовые данные. — М. : Московский городской педагогический университет, 2012. — 204 c. — 2227-8397. — Режим доступа: http://www.iprbookshop.ru/26444.html

1.2 Федин, Ф. О. Анализ данных. Часть 2. Инструменты Data Mining [Электронный ресурс] : учебное пособие / Ф. О. Федин, Ф. Ф. Федин. — Электрон. текстовые данные. — М. : Московский городской педагогический университет, 2012. — 308 c. — 2227-8397. — Режим доступа: http://www.iprbookshop.ru/26445.html

2. Лесковец, Ю. Анализ больших наборов данных / Лесковец Ю. , Раджараман А. , Джеффри Д. Ульман - Москва : ДМК Пресс, 2016. - 498 с. - ISBN 978-5-97060-190-7. - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785970601907.html

3. Вольфсон, М. Б. Анализ данных : учебное пособие / М. Б. Вольфсон. — Санкт-Петербург : СПбГУТ им. М.А. Бонч-Бруевича, 2015. — 81 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/180254

Учебно-методическая

1. Головин В.А. Методические указания для самостоятельной работы студентов по дисциплине «Технологии хранения и обработки больших объемов данных» / В.А. Головин. Ульяновск: УлГУ, 2019/ В. А. Головин; УлГУ, Фак. математики, информ. и авиац. технологий. - Ульяновск : УлГУ, 2019 - Загл. с экрана; Неопубликованный ресурс. - Электрон. текстовые дан. (1 файл : 752 КБ). - Текст : электронный.— URL: http://lib.ulsu.ru/MegaPro/Download/MObject/7120.

Согласовано:
Специалист ведущий НБ УлГУ Должность сотрудника научной библиотеки | Боброва Н.А. ФИО | _____/_____ подпись / дата | 2023

б) программное обеспечение:

Для образовательного процесса по данной дисциплине необходим стационарный класс ПК с установленным следующим программным обеспечением:

- операционная среда MS Windows;
- пакет приложений MS Office
- СУБД MS SQL и Eclipse;

в) Профессиональные базы данных, информационно-справочные системы:

1. Электронно-библиотечные системы:

1.1. Цифровой образовательный ресурс IPRsmart : электронно-библиотечная система : сайт / ООО Компания «Ай Пи Ар Медиа». - Саратов, [2023]. – URL: http://www.iprbookshop.ru. – Режим доступа: для зарегистрир. пользователей. - Текст : электронный.

1.2. Образовательная платформа ЮРАЙТ : образовательный ресурс, электронная библиотека : сайт / ООО Электронное издательство «ЮРАЙТ». – Москва, [2023]. - URL: https://urait.ru. – Режим доступа: для зарегистрир. пользователей. - Текст : электронный.

1.3. База данных «Электронная библиотека технического ВУЗа (ЭБС «Консультант студента») : электронно-библиотечная система : сайт / ООО «Политехресурс». – Москва, [2023]. – URL: https://www.studentlibrary.ru/cgi-bin/mb4x. – Режим доступа: для зарегистрир. пользователей. – Текст : электронный.

1.4. Консультант врача. Электронная медицинская библиотека : база данных : сайт / ООО «Высшая школа организации и управления здравоохранением-Комплексный медицинский консалтинг». – Москва, [2023]. – URL: https://www.rosmedlib.ru. – Режим доступа: для зарегистрир. пользователей. – Текст : электронный.

1.5. Большая медицинская библиотека : электронно-библиотечная система : сайт / ООО «Букап». – Томск, [2023]. – URL: https://www.books-up.ru/ru/library/ . – Режим доступа: для зарегистрир. пользователей. – Текст : электронный.

1.6. ЭБС Лань : электронно-библиотечная система : сайт / ООО ЭБС «Лань». – Санкт-Петербург, [2023]. – URL: https://e.lanbook.com. – Режим доступа: для зарегистрир. пользователей. – Текст : электронный.

1.7. ЭБС **Znanium.com :** электронно-библиотечная система : сайт / ООО «Знаниум». - Москва, [2023]. - URL: http://znanium.com . – Режим доступа : для зарегистрир. пользователей. - Текст : электронный.

2. КонсультантПлюс [Электронный ресурс]: справочная правовая система. / ООО «Консультант Плюс» - Электрон. дан. - Москва : КонсультантПлюс, [2023].

3. Базы данных периодических изданий:

3.1. eLIBRARY.RU: научная электронная библиотека : сайт / ООО «Научная Электронная Библиотека». – Москва, [2023]. – URL: http://elibrary.ru. – Режим доступа : для авториз. пользователей. – Текст : электронный

3.2. Электронная библиотека «Издательского дома «Гребенников» (Grebinnikon) : электронная библиотека / ООО ИД «Гребенников». – Москва, [2023]. – URL: https://id2.action-media.ru/Personal/Products. – Режим доступа : для авториз. пользователей. – Текст : электронный.

4. Федеральная государственная информационная система «Национальная электронная библиотека» : электронная библиотека : сайт / ФГБУ РГБ. – Москва, [2023]. – URL: https://нэб.рф. – Режим доступа : для пользователей научной библиотеки. – Текст : электронный.

5. Российское образование : федеральный портал / учредитель ФГАУ «ФИЦТО». – URL: http://www.edu.ru. – Текст : электронный.

6. Электронная библиотечная система УлГУ : модуль «Электронная библиотека» АБИС Мега-ПРО / ООО «Дата Экспресс». – URL: http://lib.ulsu.ru/MegaPro/Web. – Режим доступа : для пользователей научной библиотеки. – Текст : электронный.

Согласовано:

начальник УИТТ / Бурдин О.П. / /

Должность сотрудника УИТиТ | ФИО | подпись | дата

12. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ

ДИСЦИПЛИНЫ

Аудитории для проведения лекций, семинарских занятий, для проведения лабораторных работ, для проведения текущего контроля и промежуточной аттестации.

Помещение 3/321. Аудитории укомплектованы специализированной мебелью, учебной доской. Аудитории для проведения лекций оборудованы мультимедийным оборудованием для представления информации большой аудитории. 432017, Ульяновская область, г. Ульяновск, ул. Набережная реки Свияги, д. 106 (3 корпус).

Помещения для самостоятельной работы оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа к электронной информационно-образовательной среде, электронно-библиотечной системе.

13. СПЕЦИАЛЬНЫЕ УСЛОВИЯ ДЛЯ ОБУЧАЮЩИХСЯ С ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ЗДОРОВЬЯ

В случае необходимости, обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающимся) могут предлагаться одни из следующих вариантов восприятия информации с учетом их индивидуальных психофизических возможностей:

- для лиц с нарушением зрения: в форме электронного документа, индивидуальные консультации с привлечением тифлосурдопереводчика, индивидуальные задания и консультация;
- для лиц с нарушением слуха: в форме электронного документа, индивидуальные консультации с привлечением сурдопереводчика, индивидуальные задания и консультация;
- для лиц с нарушением опорно-двигательного аппарата: в форме электронного документа, индивидуальные задания и консультация.

В случае необходимости использования в учебном процессе частично/исключительно дистанционных образовательных технологий, организация работы ППС с обучающимися с ОВЗ и инвалидами предусматривается в электронной информационно-образовательной среде с учетом их индивидуальных психофизических особенностей.

Разработчик _ _____ __/_____Шабалин А.С.__/
(Подпись) ФИО