МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Воронежский государственный технический университет»

УТВЕРЖДАЮ
Декан факультета информационных технологий и компьютерной безопасности

_____ / П.Ю. Гусев
«21» февраля 2023 г.

РАБОЧАЯ ПРОГРАММА
дисциплины (модуля)
«Технологии работы с большими данными Big Data»

Направление подготовки 09.04.01 Информатика и вычислительная техника

Профиль Управление программным инжинирингом

Квалификация выпускника магистр

Нормативный период обучения 2 года

Форма обучения Очная

Год начала подготовки 2023

Автор программы _____ В.В. Сафронов

Заведующий кафедрой автоматизированных и вычислительных систем _____ В.Ф. Барабанов

Руководитель ОПОП _____ С.А. Олейникова

Воронеж 2023

1. ЦЕЛИ И ЗАДАЧИ ДИСЦИПЛИНЫ

1.1. Цели дисциплины

Целями изучения дисциплины являются:

- формирование теоретических знаний по основам применению технологий работы с большими данными (в части машинного обучения) для построения формальных математических моделей и интерпретации результатов моделирования;

- приобретение практических навыков применения методов машинного обучения при построении формальных математических моделей и интерпретации результатов моделирования при решении практических задач в прикладных областях;

- приобретение умений и навыков использования различных программных инструментов для осуществления анализа баз данных и применения систем машинного обучения.

1.2. Задачи освоения дисциплины

Задачи освоения дисциплины: изучить технологии хранения, обработки и анализа больших данных, изучить методы построения информационных систем на основе нереляционных баз данных и распределенных систем хранения.

В теоретическом плане рассматриваются вопросы группировки данных, обнаружения значимых корреляций, В теоретическом плане рассматриваются вопросы группировки данных, обнаружения значимых корреляций, зависимостей тенденций на основе анализа имеющейся информации, определения отношений между данными различного типа, выявления систематизированных структур данных и вывода из них правила для принятия решений и прогнозирования их последствий (регрессионный, дисперсионные, кластерные, дискриминантный, факторный анализы).

В практическом плане рассматриваются: модели распределенных файловых систем и вычисления на основе баз данных; поиск подобий в данных; анализ потоковых данных, связей, социально-сетевых графов и частых наборов данных; методы кластеризации и их приложения, способы применения нейронных сетей и их приложений, сетевые аналитические модели; модели снижения размерности данных; методы машинного обучения большими данными.

Основные задачи дисциплины:

- приобретение студентами знаний о технологиях подготовки, хранения, обработки и анализа больших данных;

- изучение информационных технологий обработки и анализа больших данных;

- изучение современных технологий сбора и хранения данных;

- изучение жизненного цикла анализа данных;

- изучение процессов управления качеством обработки больших данных.

- формирование представлений о целях, способах реализации и инструментах многомерного анализа данных;

- изучение сфер применения, методов и средств Data Mining**;**
- формирование практических навыков анализа данных;
- применение статистических и математических методов для анализа больших объемов информации.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП

Дисциплина (модуль) «Технологии работы с большими данными Big Data» относится к дисциплинам части, формируемой участниками образовательных отношений блока Б.1 учебного плана.

3. ПЕРЕЧЕНЬ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Процесс изучения дисциплины «Технологии работы с большими данными Big Data» направлен на формирование следующих компетенций:

ПК-1 – Способен осуществлять администрирование и управление информационно-коммуникационными системами и сетями

ПК-5 – Способен осуществлять управление программным инжинирингом на всех этапах жизненного цикла программного обеспечения

Компетенция	**Результаты обучения, характеризующие сформированность компетенции**
ПК-1	знать оценку производительности критических приложений, влияющих на производительность сетевых устройств и программного обеспечения в целом, а также методики планирования заданной производительности сетевых устройств и программного обеспечения администрируемой сети при работе с большими данными.
	уметь оценивать производительность сетевых устройств и программного обеспечения администрируемой сети;
	владеть навыками применения административных утилит операционных систем в том числе для установки дополнительных программных продуктов и их параметризации в администрируемой сети.
ПК-5	знать информационные технологии развертывания сервисов и инструментов управления процессом обработки данных, в том числе этапы жизненного цикла обработки больших данных, языки обработки и аналитики больших данных, способы организации хранения и доступа к большим данным, методы машинного обучения и средства анализа больших данных.
	уметь решать задачи машинного обучения из различных, в том числе незнакомых, предметных областей путем подбора и комбинирования в технологические цепочки готового математического и программного обеспечения для решения задач машинного обучения и анализа больших данных.
	владеть актуальными методами исследования и обработки данных и их применению в самостоятельной научно-исследовательской деятельности в области профессиональной деятельности.

4. ОБЪЕМ ДИСЦИПЛИНЫ (МОДУЛЯ)

Общая трудоемкость дисциплины «Технологии работы с большими данными big data» составляет 4 зачетных единиц.

Распределение трудоемкости дисциплины по видам занятий

Очная форма обучения

Вид учебной работы	Всего часов	Семестры			
		3			
Аудиторные занятия (всего)	72	72			
В том числе:					
Лекции	36	36			
Практические занятия (ПЗ)					
Лабораторные работы (ЛР), в том числе в форме практической подготовки (*при наличии*)	36	36			
Самостоятельная работа	72	72			
Курсовой проект (работа) (есть, нет)	нет	нет			
Контрольная работа (есть, нет)	нет	нет			
Вид промежуточной аттестации (зачет, зачет с оценкой, экзамен)	Зачет с оценкой	Зачет с оценкой			
Общая трудоемкость час	144	144			
зач. ед.	4	4			

5. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

5.1 Содержание разделов дисциплины и распределение трудоемкости по видам занятий

очная форма обучения

№ п/п	Наименование темы	Содержание раздела	Лекц	Прак зан.	Лаб. зан.	СРС	Всего, час
1	**Введение в машинное обучение**	Основные понятия. Определение предмета машинного обучения. Примеры задач и областей приложения. Классификация. Общие принципы. Этапы классификации. Алгоритмы обучения классификаторов.	4	-	4	12	20
2	**Основные методы машинного обучения**	Нейронные сети и искусственный интеллект. Понятие и принцип работы нейронных сетей для обработки данных.	8	-	8	14	30
3	**Введение в большие данные**	Предпосылки формирования тренда больших данных. Понятие Data Minig. Прикладные инструменты для работы с Big Data. Инструменты для обработки больших данных. Технология MapRaduce. Hadoop.	8	-	8	14	30
4	**Технологии анализа данных**	Жизненный цикл анализа больших данных, стандарты. Когнитивный анализ данных Визуали-	8	-	8	16	32

		зация больших данных. Аналитика больших данных. Когнитивный анализ данных. Методы DATA MINING. Методы анализа на графах. Прикладные инструменты анализа данных. Фреймворки.					
5	**Технологии хранения больших данных.**	Распределенные хранилища, NoSql хранилища, классификация и примеры. Хранилища данных. Распределенные базы данных. Решение задач Data Mining. Задачи классификации, кластеризации. Распределенные файловые системы.	8	-	8	16	32
		Итого	**36**	**-**	**36**	**72**	**144**

5.2 Перечень лабораторных работ

Лабораторная работа 1. Основы построения и использования систем больших данных.

Лабораторная работа 2. Разработка и использование приложений на основе распределенных баз данных.

Лабораторная работа 3. Методы анализа данных Data Mining. Визуализация данных. Введение в визуализацию данных. Визуализаторы общего назначения.

Лабораторная работа 4. Реализация алгоритма извлечения данных.

Лабораторная работа 5. Настройка системы хранения данных

Лабораторная работа 6. Поиск закономерностей в данных. Нейросети.

Лабораторная работа 7. Решение задач Data Mining.

Лабораторная работа 8. Развертывание локального кластера Hadoop.

Лабораторная работа 9. Подсчет слов в тексте, с помощью MapReduce.

6. ПРИМЕРНАЯ ТЕМАТИКА КУРСОВЫХ ПРОЕКТОВ (РАБОТ) И КОНТРОЛЬНЫХ РАБОТ

В соответствии с учебным планом освоение дисциплины не предусматривает выполнение курсового проекта (работы) в 3 семестре.

Учебным планом по дисциплине «Технологии работы с большими данными Big Data» не предусмотрено выполнение контрольной работы (контрольных работ) в 3 семестре.

7. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕДЕНИЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ

7.1. Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания

7.1.1 Этап текущего контроля

Результаты текущего контроля знаний и межсессионной аттестации оцениваются по следующей системе:

«аттестован»;

«не аттестован».

Компетенция	Результаты обучения, характеризующие сформированность компетенции	Критерии оценивания	Аттестован	Не аттестован
ПК-1	знать оценку производительности критических приложений, влияющих на производительность сетевых устройств и программного обеспечения в целом, а также методики планирования заданной производительности сетевых устройств и программного обеспечения администрируемой сети при работе с большими данными.	Активная работа на лабораторных занятиях, отвечает на теоретические вопросы при защите лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	уметь оценивать производительность сетевых устройств и программного обеспечения администрируемой сети;	Решение стандартных практических задач, написание курсового проекта	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	владеть навыками применения административных утилит операционных систем в том числе для установки дополнительных программных продуктов и их параметризации в администрируемой сети.	Решение прикладных задач в конкретной предметной области, выполнение плана работ по разработке курсового проекта	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
ПК-5	знать информационные технологии развертывания сервисов и инструментов управления процессом обработки данных в том числе этапы жизненного цикла обработки больших данных, языки обработки и аналитики больших данных, способы организации хранения и доступа к большим данным, методы машинного обучения и средства анализа больших данных.	Активная работа на лабораторных занятиях, отвечает на теоретические вопросы при защите лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	уметь решать задачи машинного обучения из различных, в том числе незнакомых, предметных областей путем подбора и комбинирования в технологические цепочки готового математического и программного обеспечения для решения задач машинного обучения и	Решение стандартных практических задач, написание курсового проекта	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах

	анализа больших данных.			
	владеть актуальными методами исследования и обработки данных и их применению в самостоятельной научно-исследовательской деятельности в области профессиональной деятельности.	Решение прикладных задач в конкретной предметной области, выполнение плана работ по разработке курсового проекта	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах

7.1.2 Этап промежуточного контроля знаний

Результаты промежуточного контроля знаний оцениваются в 3 семестре для очной формы обучения по системе:

«отлично»;

«хорошо»;

«удовлетворительно»;

«неудовлетворительно».

Компетенция	Результаты обучения, характеризующие сформированность компетенции	Критерии оценивания	Отлично	Хорошо	Удовл	Неудовл
ПК-1	знать оценку производительности критических приложений, влияющих на производительность сетевых устройств и программного обеспечения в целом, а также методики планирования заданной производительности сетевых устройств и программного обеспечения администрируемой сети при работе с большими данными.	Тест	Выполнение теста на 90-100%	Выполнение теста на 80-90%	Выполнение теста на 70-80%	В тесте менее 70% правильных ответов
	уметь оценивать производительность сетевых устройств и программного обеспечения администрируемой сети;	Тест	Выполнение теста на 90-100%	Выполнение теста на 80-90%	Выполнение теста на 70-80%	В тесте менее 70% правильных ответов
	владеть навыками применения административных утилит операционных систем в том числе для установки дополнительных программных продуктов и их параметризации в администрируемой сети.	Тест	Выполнение теста на 90-100%	Выполнение теста на 80-90%	Выполнение теста на 70-80%	В тесте менее 70% правильных ответов
ПК-5	знать информационные технологии	Тест	Выполнение теста на 90-	Выполнение теста на 80-90%	Выполнение теста на 70-	В тесте менее 70% пра-

	развертывания сервисов и инструментов управления процессом обработки данных в том числе этапы жизненного цикла обработки больших данных, языки обработки и аналитики больших данных, способы организации хранения и доступа к большим данным, методы машинного обучения и средства анализа больших данных.		100%		80%	вильных ответов
	уметь решать задачи машинного обучения из различных, в том числе незнакомых, предметных областей путем подбора и комбинирования в технологические цепочки готового математического и программного обеспечения для решения задач машинного обучения и анализа больших данных.	Тест	Выполнение теста на 90-100%	Выполнение теста на 80-90%	Выполнение теста на 70-80%	В тесте менее 70% правильных ответов
	владеть актуальными методами исследования и обработки данных и их применению в самостоятельной научно-исследовательской деятельности в области профессиональной деятельности.	Тест	Выполнение теста на 90-100%	Выполнение теста на 80-90%	Выполнение теста на 70-80%	В тесте менее 70% правильных ответов

7.2 Примерный перечень оценочных средств (типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности)

7.2.1 Примерный перечень заданий для подготовки к тестированию

1. В какой последовательности технология MapReduce использует в рабочем процессе задачи-распределители и задачи-редукторы?

а) последовательно, сначала одни, а затем другие.

б) параллельно или обе одновременно.

в) поочередно, одну за другой.

2. Какие функции в MapReduce запускает главный контроллер-Мастер (найти неверный ответ)?

а) создание распределителей и редукторов.

б) назначает задачам рабочие процессы.

в) обрабатывает отказ узла редуктора.

3. В чем состоит рекурсивное обобщение MapReduce?

а) путем повторения заданий.

б) выход каждой рекурсивной задачи подается на вход следующей.

в) следует сохранять контрольную точку всего вычисления, для перезаписи после сбоя.

4. Как осуществляется разбиение документов на шинглы?

а) разбиение на слова, принадлежащие одной части речи (существительные, глаголы и т.п.).

б) разбиение на слова одинаковой длины.

в) разбиение на последовательность подстрок некоторой длины *к.*

5. Как часто в документах встречаются стоп-слова?

а) редко.

б) часто.

в) никогда.

6. В какой степени сигнатуры отражают свойства множеств?

а) сигнатуры состоят из усредненных величин множества.

б) чем длиннее сигнатура, тем в большей степени она отражает свойства множества.

в) сигнатура содержит все элементы множества за исключением повторяющихся.

7. Как вычислить minхэш характеристической матрицы?

а) для *j*-го столбца. номер *i* – строки в которой встречается первая единица.

б) для *j*-го столбца, номер строки состоящей из одних единиц.

в) для *j*-го столбца, номер строки состоящей из одних нулей.

8. Какая связь существует между minхэш и коэффициентом Жаккара?

а) значение *minхэш* -функции для двух множеств равно двоичному логарифму коэффициента Жаккара.

б) *minхэш* -функция случайных перестановок строк двух множеств порождает одно значение равное коэффициенту Жаккара для этих множеств.

в) *minхэш* -функция для двух множеств равна произведению коэффициента Жаккара на его алгебраическое дополнение.

9. В чем состоит основная проблема обработки потоков данных?

а) потоки имеют нормальное распределение и изменяются вблизи среднего значения.

б) потоки обрабатываются как в оперативной памяти, так и на внешних носителях.

в) чтобы не потерять данные их надо обрабатывать в масштабе реального времени.

10. Как устроен фильтр Блюма?

а) состоит из массива *n* бит, принимающих случайные значения (0;1) и массива «ключей».

б) состоит из набора хэш-функций и множества содержащие *m* ключей.

в) массив *n* – бит, первоначально равных 0, набор хэш-функций отображающих значения ключей и множество содержащее *m* ключей.

7.2.2 Примерный перечень заданий для решения стандартных задач

1. В чем состоит стратегия кластеризации?

а) в объединении близких точек многомерного пространства в один объект (кластер) с усредненными характеристиками.

б) разделение множества на части с помощью плоскостей.

в) разделение множества на внутренние точки или «свои» и внешние точки или «чужие».

2. Как ускорить алгоритм иерархической кластеризации для евклидовой метрики?

а) следует матрицу расстояний между элементами рассчитать только один раз.

б) следует элементы объединенные в кластер вычеркивать из матрицы расстояний.

в) пересчитывать среднее значение для кластера без привлечения исходных координат.

3. В чем состоит алгоритм к-средних?

а) из исходного множества случайным образом выбираются *к*-центров кластеров.

б) рассчитывается диаметр множества, делится на *к* и кругами равными полученному значению покрывается все множество; внутри каждого круга находится центр кластера.

в) делим множество на части с помощью *к*-плоскостей.

4.Как реализуется алгоритм кластеризации потока?

а) точки потока разбиваются на одинаковые интервалы, в которых хранится информация о кластере.

б) точки потока разбиваются на интервалы, размеры которых являются степенями двойки.

в) точки потока разбиваются на интервалы размеры, которых уменьшаются в два раза.

5. Как проводится кластеризация с помощью алгоритма CURE?

а) предварительно провести кластеризацию на части данных, провести сдвиг данных относительно центра кластера, объединить если имеется близкая пара.

б) это метод рекурсивной кластеризации.

в) кластеризация выборки, не принадлежащей нормальному распределению.

6. Эксперт это …

1. специалист в области анализа и моделирование;

2. специалист в предметной области;

3. человек, решающий определенные задачи;

4. человек, который имеет опыт в программировании.

7. Задача классификации сводится к …

1. нахождения частых зависимостей между объектами или событиями;

2. определения класса объекта по его характеристиками;

3. определение по известным характеристикам объекта значение некоторого его параметра;

4. поиска независимых групп и их характеристик в всем множестве анализируемых данных.

8. Какие требования предъявляются к вычислительным методам?

1. Адекватность дискретной модели задачи

2. Точность, простота

3. Устойчивость алгоритма

4. Корректность, приемлемое время

9. Статистической вероятностью события А называется относительная частота появления события А в n произведенных испытаниях, Выберите правильную формулу статистической вероятности, если P(A)- статистическая вероятность события А, m(A) - число испытаний, в которых появилось событие А, а n-число независимых испытаний:

1. P(A)=m(A)=m(A)/1

2. P(A)=m(A)= n*m(A)

3. P(A)=m(A)=1/n

4. P(A)=m(A)=m(A)/n

10. Коэффициент корреляции может принимать значение:

1. от -1 до +1

2. от 0 до +1

3. от -1 до 0

4. от +1 до + 2

7.2.3 Примерный перечень заданий для решения прикладных задач

1. Случайная величина X называется непрерывной, если:

1. если ее функция распределения непрерывна в любой точке.

2. если ее функция распределения непрерывна в любой точке и дифференцируема всюду, кроме,

быть может, отдельных точек.

3. если ее функция распределения непрерывна в любой точке и дифференцируема всюду.

4. если ее функция распределения непрерывна.

2. Корреляционный метод может быть применен, если число наблюдений:

1. >5

2. равно 2

3. равно 5

4. равно числу наблюдаемых значений

3. К описательным моделям относятся следующие модели данных:

1. модели классификации и последовательностей;

2. регрессионные, кластеризации, исключений, итоговые и ассоциации;

3. классификации, кластеризации, исключений, итоговые и ассоциации;

4. модели классификации, последовательностей и исключений.

4. Модели классификации описывают …

1. правила или набор правил в соответствии с которыми можно отнести описание любого нового

объекта к одному из классов;

2. функции, которые позволяют прогнозировать изменения непрерывных числовых параметров;

3. функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме;

4. группы, на которые можно разделить объекты, данные о которых подвергаются анализа.

5. В случае линейного уравнения регрессии связь между факторным и результативным признаками является тесной, если:

1. r < -1

2. r = 0

3. r = -1

4. r = 1

6. Корреляционный анализ определяет:

1. интеграл(xdx)+интеграл(ydy)

2. форму связи между X и Y

3. тесноту связи между X и Y

4. производную Y'x

7. Сколько трехзначных чисел можно составить из цифр 1, 2, 3, 4, 5, если все цифры в числе различны?

1. 20

2. 60

3. 10

4. 125

8. Бросают два кубика. Какие из следующих событий случайные? (Тип ответа: Многие из многих)

1. A={на кубиках выпало одинаковое число очков}

2. B={сумма очков на кубиках не превосходит 12}

3. C={сумма очков на кубиках равна 11}

4. D={произведение очков на кубиках равно 11}

9. В коробке 3 красных, 3 желтых, 3 зеленых шара. Вытащили наугад 4 шара. Какие из следующих событий невозможные? (Тип ответа: Многие из многих)

1. Все вынутые шары одного цвета.

2. Все вынутые шары разных цветов.

3. Среди вынутых шаров есть шары разных цветов.

4. Среди вынутых есть шары всех трех цветов.

10. В партии из 10 деталей имеются 4 бракованных. Какова вероятность того, что среди наудачу отобранных 5 деталей окажутся 2 бракованные?

1. 0,25
2. 0,476
3. 0,5
4. 0,235

11. Стрелок попадает в десятку с вероятностью 0,05, в девятку – с вероятностью 0,2, в восьмерку – с вероятностью 0,5. Сделан один выстрел. Какова вероятность того, что будет выбито менее 8 очков?

1. 0,1
2. 0,75
3. 0,25
4. 0,9

7.2.4 Примерный перечень вопросов для подготовки к зачету

1. Понятие Больших данных.
2. Особенности сбора, хранения, обработки и анализа больших массивов данных.
3. Источники больших данных.
4. Варианты построения распределённых баз данных, репликация, фрагментация.
5. Согласованность. CAP-теорема.
6. Классы NoSQL баз данных.
7. Примеры СУБД NoSQL.
8. Графовые СУБД.
9. Задачи консолидации данных.
10. Многомерные хранилища данных.
11. Реляционные хранилища данных.
12. Виртуальные хранилища.
13. Нечеткие среды.
14. Преобразование данных в ETL.
15. Обогащение данных.
16. Технология Map- Reduce, GOOGLE BIGTABLE.
17. Полнотекстовый поиск.
18. Параллельные запросы.
19. Принципы анализа данных.
20. Структурированные данные.
21. Подготовка данных к анализу.
22. Технологии KDD и Data Mining.
23. Трансформация данных. Трансформация упорядоченных данных.
24. Группировка данных.
25. Слияние данных.
26. Квантование.
27. Нормализация и кодирование данных.
28. OLAP-анализ.
29. Визуализация данных. Визуализаторы общего назначения.
30. Выявление закономерностей в виде деревьев решения,

31. Выявление закономерностей в виде логических правил,

32. Выявление закономерностей в виде нейронных сетей.

33. Процесс аналитики анализа больших данных.

34. Определите понятие Data Mining.

35. В чем состоит когнитивный анализ данных.

36. Определите различия между параметрическими, непараметрическими и номинальными методами.

37. Опишите основную идею корреляционного анализа.

38. Регрессионный анализ.

39. Основная идея дисперсионного анализа.

40. Сущность кластерного анализа.

41. Дискриминантный анализ: модель и общая процедура выполнения.

42. Цели факторного анализа.

43. Аналитический и информационный походы к работе с данными.

44. Большие данные в региональном управлении.

45. Введение в визуализацию данных.

46. Введение в оценку качества данных.

47. Визуализаторы OLAP-анализа.

48. Визуализаторы общего назначения.

49. Визуализаторы, применяемые для оценки качества моделей.

50. Восстановление пропущенных значений.

51. Выявление аномальных значений.

52. Математические модели выявления закономерностей и взаимосвязей в данных.

53. Математические основы нейронных сетей.

54. Нейронные сети для обработки временных рядов.

55. Нейронные сети для обработки изображений.

56. Нейронные сети для обработки текстов.

57. Нейронные сети и большие данные в юриспруденции.

58. Нейронные сети и искусственный интеллект.

59. Обработка дубликатов и противоречий.

60. Открытые государственные данные для граждан и бизнеса: инструменты для анализа и валидации, повышение качества и достижение социально-экономического эффекта от их применения в различных областях жизнедеятельности общества

61. Оценка качества, очистка и предобработка данных.

62. Очистка и предобработка данных.

63. Построение графиков и диаграмм.

64. Технологии и методы оценки качества данных.

65. Фильтрация данных.

66. Цифровая трансформация.

67. Технологии обработки больших данных: NoSQL, MapReduce, Hadoop , R.

68. Технологии Business Intelligence и реляционные системы управления базами данных.

69. Прогнозирование и предвидение: общее и особенное.
70. Виды прогнозов
71. Опишите методики анализа больших данных.
72. Процесс аналитики анализа больших данных.
73. Вопросы безопасности больших данных.

7.2.5 Примерный перечень вопросов для подготовки к экзамену

Учебным планом не предусмотрено

7.2.6 Методика выставления оценки при проведении промежуточной аттестации

Зачет проводится по билетам, включающим по два вопроса. Допуском к зачету является выполнение всех лабораторных работ и положительное текущее тестирование.

Зачет ставится, если студент выполнил все лабораторные работы, прошел тестирование по темам теоретического материала и ответил на один или два вопроса.

Зачет не ставится, если студент не выполнил лабораторные работы и не ответил ни на один вопрос на зачете.

7.2.7 Паспорт оценочных материалов

№ п/п	Контролируемые разделы (темы) дисциплины	Код контролируемой компетенции (или ее части)	Наименование оценочного средства
1	Введение в машинное обучение	ПК-1, ПК-5	Тест, защита лабораторных работ, зачет, устный опрос
2	Основные методы машинного обучения	ПК-1, ПК-5	Тест, защита лабораторных работ, зачет, устный опрос
3	Введение в большие данные	ПК-1, ПК-5	Тест, защита лабораторных работ, зачет, устный опрос
4	Технологии анализа данных	ПК-1, ПК-5	Тест, защита лабораторных работ, зачет, устный опрос
5	Технологии хранения больших данных	ПК-1, ПК-5	Тест, защита лабораторных работ, зачет, устный опрос

7.3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Тестирование осуществляется, либо при помощи компьютерной системы тестирования, либо с использованием выданных тест-заданий на бумажном носителе. Время тестирования 30 мин. Затем осуществляется проверка теста экза-

менатором и выставляется оценка согласно методике выставления оценки при проведении промежуточной аттестации.

Решение стандартных задач осуществляется, либо при помощи компьютерной системы тестирования, либо с использованием выданных задач на бумажном носителе. Время решения задач 30 мин. Затем осуществляется проверка решения задач экзаменатором и выставляется оценка, согласно методике выставления оценки при проведении промежуточной аттестации.

Решение прикладных задач осуществляется, либо при помощи компьютерной системы тестирования, либо с использованием выданных задач на бумажном носителе. Время решения задач 30 мин. Затем осуществляется проверка решения задач экзаменатором и выставляется оценка, согласно методике выставления оценки при проведении промежуточной аттестации.

8 УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

8.1 Перечень учебной литературы, необходимой для освоения дисциплины

1. Воронов В.И. Data Mining - технологии обработки больших данных: учебное пособие / Воронов В.И., Воронова Л.И., Усачев В.А.. — Москва : Московский технический университет связи и информатики, 2018. — 47 с. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: https://www.iprbookshop.ru/81324.html

2. Железнов М.М. Методы и технологии обработки больших данных: учебно-методическое пособие / Железнов М.М.. — Москва : МИСИ-МГСУ, ЭБС АСВ, 2020. — 46 с. — ISBN 978-5-7264-2193-3. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: https://www.iprbookshop.ru/101802.html

3. Адлер Ю.П. Статистическое управление процессами. «Большие данные»: учебное пособие / Адлер Ю.П., Черных Е.А.. — Москва : Издательский Дом МИСиС, 2016. — 52 с. — ISBN 978-5-87623-969-3. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: https://www.iprbookshop.ru/64199.html

4. Организация самостоятельной работы обучающихся: методические указания для студентов, осваивающих основные образовательные программы высшего образования – бакалавриата, специалитета, магистратуры: методические указания / сост. В.Н. Почечихина, И.Н. Крючкова, Е.И. Головина, В.Р. Демидов; ФГБОУ ВО «Воронежский государственный технический университет». – Воронеж, 2020. – 14 с.

5. Разработка мультимедийных приложений с использованием библиотек OpenCV и IPP / А.В. Бовырин [и др.].— М. : Интернет-Университет Информационных Технологий (ИНТУИТ), 2016. — 515 с. — 2227-8397. — Текст: электронный // Электронно-библиотечная система IPR BOOKS: [сайт]. — URL: http://www.iprbookshop.ru/39564.html

6. Неделько В.М. Основы статистических методов машинного обучения: учебное пособие / В.М. Неделько. — Новосибирск: Новосибирский государственный технический университет, 2010. — 72 с. — 978-5-7782-1385-2. — Текст: электронный // Электронно-библиотечная система IPR BOOKS: [сайт]. — URL: http://www.iprbookshop.ru/45418.html

8.2 Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень лицензионного программного обеспечения, ресурсов информационно-телекоммуникационной сети «Интернет», современных профессиональных баз данных и информационных справочных систем

Лицензионное ПО:

- Windows Professional 7 Single Upgrade MVL A Each Academic
- Microsoft Office Word 2007
- Microsoft Office Power Point 2007

Свободно распространяемое ПО:

- Microsoft Visual Studio Community Edition
- Microsoft SQL Server Express
- Microsoft SQL Server Managment Studio
- СУБД MS SQL Server 2012
- TensorFlow
- Theano - библиотека численного вычисления в Python
- Keras - библиотека по глубинному обучению на Python
- Apache Hadoop
- HBase.

Отечественное ПО:

- Яндекс.Браузер
- Архиватор 7z
- Astra Linux

Ресурс информационно-телекоммуникационной сети «Интернет»:

- Образовательный портал ВГТУ
- http://www.edu.ru/
- https://metanit.com/

Информационно-справочные системы:

- http://window.edu.ru
- https://wiki.cchgeu.ru/

Современные профессиональные базы данных:

- https://proglib.io
- https://msdn.microsoft.com/ru-ru/
- https://docs.microsoft.com/

Информационные технологии, используемые при осуществлении образовательного процесса по дисциплине:

- лекции с применением мультимедийных средств;

- обучение прикладным информационным технологиям, ориентированным на специальность, в рамках лабораторных работ с применением лицензионного программного обеспечения.

9 МАТЕРИАЛЬНО-ТЕХНИЧЕСКАЯ БАЗА, НЕОБХОДИМАЯ ДЛЯ ОСУЩЕСТВЛЕНИЯ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА

Для проведения лекционных занятий необходима аудитория, оснащенная оборудованием для лекционных демонстраций и проекционной аппаратурой.

Для проведения лабораторных работ необходима лаборатория с ПК, оснащенными программами для проведения лабораторного практикума и обеспечивающими возможность доступа к локальной сети кафедры и Интернет, из следующего перечня:

- 408 (Лаборатория разработки программных систем)
- 412 (Лаборатория микропроцессорной техники)
- 415 (Лаборатория распределённых вычислений)
- 419 (Лаборатория телекоммуникационных систем)
- 417 (Лаборатория проектирования вычислительных комплексов и сетей)

Лаборатории расположены по адресу: 394018, г. Воронеж, Плехановская, 11 (учебный корпус №2).

10 МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

По дисциплине «Технологии работы с большими данными Big Data» читаются лекции, проводятся лабораторные занятия.

Основой изучения дисциплины являются лекции, на которых излагаются наиболее существенные и трудные вопросы, а также вопросы, не нашедшие отражения в учебной литературе.

Лабораторные работы выполняются на лабораторном оборудовании в соответствии с методиками, приведенными в указаниях к выполнению работ.

Большое значение по закреплению и совершенствованию знаний имеет самостоятельная работа студентов. Информацию обо всех видах самостоятельной работы студенты получают на занятиях.

Контроль усвоения материала дисциплины производится проверкой и защитой лабораторных работ. Освоение дисциплины оценивается на зачете.

Вид учебных занятий	Деятельность студента (*особенности деятельности студента инвалида и лица с ОВЗ, при наличии таких обучающихся*)
Лекция	Написание конспекта лекций: кратко, схематично, последовательно фиксировать основные положения, выводы, формулировки, обобщения; помечать важные мысли, выделять ключевые слова, термины. Проверка терминов, понятий с помощью энциклопедий, словарей, справочников с выписыванием толкований в тетрадь. Обозначение вопросов, терминов, материала, которые вызывают трудности, поиск ответов в рекомен-

	дуемой литературе. Если самостоятельно не удается разобраться в материале, необходимо сформулировать вопрос и задать преподавателю на лекции или на практическом занятии.
Лабораторные занятия	Лабораторные работы позволяют научиться применять теоретические знания, полученные на лекции при решении конкретных задач. Чтобы наиболее рационально и полно использовать все возможности лабораторных занятий для подготовки к ним необходимо: разобрать лекцию по соответствующей теме, ознакомится с соответствующим разделом учебного пособия, проработать дополнительную литературу и источники, изучить методическое обеспечение лабораторной работы.
Самостоятельная работа	Самостоятельная работа студентов способствует глубокому усвоения учебного материала и развитию навыков самообразования. Самостоятельная работа предполагает следующие составляющие: - работа с текстами: учебниками, справочниками, дополнительной литературой, а также проработка конспектов лекций; - работа над темами для самостоятельного изучения; - участие в работе студенческих научных конференций, олимпиад; - подготовка к лабораторным занятиям; - оформление отчетов по лабораторным работам; - подготовка к промежуточной аттестации.
Подготовка к зачету	При подготовке к зачету необходимо ориентироваться на конспекты лекций, рекомендуемую литературу и решение индивидуальных заданий на лабораторных занятиях.

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

№ п/п	Перечень вносимых изменений	Дата внесения изменений	Подпись заведующего кафедрой, ответственной за реализацию ОПОП