

Документ подписан простой электронной подписью
Информация о владельце:
ФИО: Вишневский Дмитрий Александрович
Должность: Ректор
Дата подписания: 30.04.2025 11:55:50
Уникальный программный ключ:
03474917c4d012283e5ad996a48a5e700b1de

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
(МИНОБРНАУКИ РОССИИ)

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ДОНБАССКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «ДонГТУ»)

Факультет информационных технологий и автоматизации
производственных процессов
Кафедра интеллектуальных систем и информационной
безопасности



УТВЕРЖДАЮ
И.о. проректора
по учебной работе

Д.В. Мулов

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Интеллектуальный анализ больших данных
(наименование дисциплины)

10.05.03 Информационная безопасность автоматизированных систем
(код, наименование направления/специальности)

Квалификация специалист по защите информации
(бакалавр/специалист/магистр)

Форма обучения очная
(очная, очно-заочная, заочная)

Алчевск, 2024

1 Цели и задачи изучения дисциплины

Цели дисциплины. Целью изучения дисциплины «Интеллектуальный анализ больших данных» является изучение интеллектуальных методов и моделей для анализа больших данных.

Задачи изучения дисциплины:

- приобретение студентами знаний о технологиях подготовки, хранения, обработки и анализа больших данных;
- приобретение практических навыков работы большими данными.

2 Место дисциплины в структуре образовательной программы

Логико-структурный анализ дисциплины – курс входит часть БЛОКА 1, формируемую участниками образовательных отношений «Дисциплины (модули)» подготовки студентов по специальности 10.05.03 Информационная безопасность автоматизированных систем (Безопасность открытых информационных систем).

Дисциплина реализуется кафедрой интеллектуальных систем и информационной безопасности. Основывается на базе дисциплин: «Высшая математика», «Методы анализа данных».

Является основой для изучения следующих дисциплин: «Выполнение и защита выпускной квалификационной работы».

Для изучения дисциплины необходимы компетенции, сформированные у студента для решения профессиональных задач деятельности, связанных с научно-исследовательской работы.

Курс является фундаментом для ориентации студентов в сфере научных исследований.

Общая трудоемкость освоения дисциплины составляет 5 зачетных единиц, 180 ак.ч. Программой дисциплины предусмотрены лекционные (18 ч.), лабораторные (36 ч.) занятия и самостоятельная работа студента (126 ч.).

Дисциплина изучается на 5 курсе в 9 семестре. Форма промежуточной аттестации – экзамен.

3 Перечень результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения ОПОП ВО

Процесс изучения дисциплины «Базы данных» направлен на формирование компетенций, представленных в таблице 1.

Таблица 1 –Компетенции, обязательные к освоению

Код	Наименование специальности, направления подготовки	Компетенция (код, содержание)	Индикатор (код, наименование)
10.05.03	Информационная безопасность автоматизированных систем	ПК-1 Способен разрабатывать системы защиты информации автоматизированных систем	ПК-1.2 Выполняет проектирование и реализацию системы защиты информации автоматизированных систем

4 Объём и виды занятий по дисциплине

Общая трудоёмкость учебной дисциплины составляет 5 зачётных единиц, 180 ак.ч.

Самостоятельная работа студента (СРС) включает проработку материалов лекций, подготовку к практическим занятиям, текущему контролю, выполнение индивидуального задания, самостоятельное изучение материала и подготовку к экзамену.

При организации внеаудиторной самостоятельной работы по данной дисциплине используются формы и распределение бюджета времени на СРС для очной формы обучения в соответствии с таблицей 2.

Таблица 2 – Распределение бюджета времени на СРС

Вид учебной работы	Всего ак.ч.	Ак.ч. по семестрам
		9
Аудиторная работа, в том числе:	54	54
Лекции (Л)	18	18
Практические занятия (ПЗ)	-	-
Лабораторные работы (ЛР)	36	36
Курсовая работа/курсовой проект	-	-
Самостоятельная работа студентов (СРС), в том числе:	126	126
Подготовка к лекциям	9	9
Подготовка к лабораторным работам	18	18
Подготовка к практическим занятиям / семинарам	-	-
Выполнение курсовой работы / проекта	—	—
Расчетно-графическая работа (РГР)	-	-
Реферат (индивидуальное задание)	12	12
Домашнее задание	15	15
Подготовка к контрольным работам	-	-
Подготовка к коллоквиуму	-	-
Аналитический информационный поиск	18	18
Работа в библиотеке	18	18
Подготовка к экзамену	36	36
Промежуточная аттестация – экзамен (Э)	Э	Э
Общая трудоемкость дисциплины		
ак.ч.	180	180
з.е.	5	5

5 Содержание дисциплины

С целью освоения компетенции, приведенной в п.3 дисциплина разбита на 4 темы:

- тема 1 (Введение в большие данные. Классификация задач анализа данных. Жизненный цикл аналитики данных);
- тема 2 (Высокопроизводительные вычисления);
- тема 3 (Масштабирование и многоуровневое хранение данных);
- тема 4 (Сложные методы аналитики).

Виды занятий по дисциплине и распределение аудиторных часов для очной формы приведены в таблице 3.

Таблица 3 – Виды занятий по дисциплине и распределение аудиторных часов (очная форма обучения)

№ п/п	Наименование темы (раздела) дисциплины	Содержание лекционных занятий	Трудоемкость в ак.ч.	Темы практических занятий	Трудоемкость в ак.ч.	Тема лабораторных занятий	Трудоемкость в ак.ч.
1	2	3	4	5	6	7	8
1	Введение в большие данные. Задачи анализа данных. Жизненный цикл аналитики данных	Перегрузка информацией и Data Mining. Типы закономерностей. Модели вместо законов Системы и модели. Модели информационно-развивающихся систем. Виды знаний и способы их представления. Классы систем Data Mining	2 2	—	—	Подбор наборов больших данных для анализа	4
2	Высокопроизводительные вычисления	Классификация архитектур вычислительных систем (по числу потоков команд и данных). Архитектурные свойства высокопроизводительных ВС. Показатели эффективности структуры ВС. Типовые структуры ВС. Вычислительные кластеры (computer cluster).	4	—	—	Библиотеки Python для анализа больших данных	6
3	Масштабирование и многоуровневое хранение данных	NoSQL. Масштабируемость. Репликация. CAP – теорема. Основы NoSQL. СУБД, поддерживающие NoSQL.	6	—	—	Изучение технологии NoSQL	8

Окончание таблицы 3

1	2	3	4	5	6	7	8
4	Сложные методы аналитики	Методы матричного анализа. Оптимизация. Вероятность. Основные вероятностные формулы. Закон арксинуса. Математическая статистика как некорректная обратная задача теории вероятностей. Многомерный нормальный закон. Генерация случайных чисел. Метод наименьших квадратов в линейной модели измерений. Множественный регрессионный анализ. Главные компоненты и факторный анализ. Дискриминантный анализ. Анализ канонических корреляций. Дискриминантные информанты и классификация. Оценка вероятностей ошибочной классификации. Классификация на основе линейных дискриминантных форм. Кластеризация. Выбор метрики. Метод k средних. Метод опорных векторов.	4	—	-	Глубокий анализ больших данных	6
Всего аудиторных часов			18	-		36	

6 Фонд оценочных средств для проведения текущего контроля успеваемости и промежуточной аттестации по итогам освоения дисциплины

6.1 Критерии оценивания

В соответствии с Положением о кредитно-модульной системе организации образовательного процесса ФГБОУ ВО «ДонГТУ» (https://www.dstu.education/images/structure/license_certificate/polog_kred_modul.pdf) при оценивании сформированности компетенций по дисциплине используется 100-балльная шкала.

Перечень работ по дисциплине и способы оценивания знаний приведены в таблице 4.

Таблица 4 – Перечень работ по дисциплине и способы оценивания знаний

Вид учебной работы	Способ оценивания	Количество баллов
Выполнение лабораторных работ	Предоставление отчетов	60-100
Итого	–	60 - 100

Экзамен проставляется автоматически, если студент набрал в течении семестра не менее 60 баллов и отчитался за каждую контрольную точку. Минимальное количество баллов по каждому из видов текущей работы составляет 60% от максимального.

Экзамен по дисциплине «Интеллектуальный анализ больших данных» проводится по результатам работы в семестре. В случае, если полученная в семестре сумма баллов не устраивает студента, во время зачетной сессии студент имеет право повысить итоговую оценку либо в форме устного собеседования по приведенным ниже вопросам (п.п. 6.5), либо в результате тестирования.

Шкала оценивания знаний при проведении промежуточной аттестации приведена в таблице 5.

Таблица 5 –Шкала оценивания знаний

Сумма баллов за все виды учебной деятельности	Оценка по национальной шкале зачёт/экзамен
0-59	Не зачтено/неудовлетворительно
60-73	Зачтено/удовлетворительно
74-89	Зачтено/хорошо
90-100	Зачтено/отлично

6.2 Домашние задания

Задание 1. Используя MapReduce выполнить подсчет количества разнородных элементов в наборе данных, выданных преподавателем: в следующей последовательности, согласно алгоритма MapReduce:

Шаг 1: Сохраните набор данных в 4 разделах (как в HDFS). Цель научиться балансировать нагрузку на ресурсы.

Шаг 2: Сопоставьте данные. Цель понять суть кластеризации и присвоения ключей.

Шаг 3: Сортировка и перемешивание. Цель научиться балансировать нагрузку на ресурсы.

Шаг 4: Сгенерировать пары "ключ-значение".

Задание 2. Используя Hadoop Distributed File System (HDFS) посчитать, сколько раз встречается слово в тексте.

Шаг 1: Загрузите и установите VirtualBox.

Шаг 2: Загрузите и установите образ виртуальной машины (VM) Cloudera.

Шаг 3: Запустите Cloudera VM.

Шаг 4: Загрузите файл в HDFS с текстом, выданный преподавателем;

Шаг 5: Запустите приложение WordCount.

Шаг 6: Скопируйте результаты WordCount из Hadoop Distributed File System (HDFS).

Задание 3. Выполните визуализацию кластеров, полученных в результате выполнения Задания 1.

6.3 Темы для рефератов (презентаций) – индивидуальное задание

1. Математические модели в анализе данных: применение статистики и вероятности.

2. Машинное обучение и анализ данных: алгоритмы и методы.

3. Большие данные и их анализ: вызовы и перспективы.

4. Визуализация данных: инструменты и техники.

5. Анализ текстовых данных: методы обработки и классификации.

6. Анализ временных рядов: прогнозирование и тренды.
7. Анализ социальных сетей: извлечение информации и моделирование.
8. Анализ географических данных: пространственная статистика и картография.
9. Этические аспекты анализа данных: конфиденциальность и защита личной информации.
10. Примеры использования больших данных в социологии, в сферах коммуникаций и маркетинга
11. Основные источники больших данных по различным отраслям промышленности
12. Принципы и алгоритмы работы с большими данным.
13. Современные технологии обработки больших данных
14. Возможности и ограничения использования различных пакетов и систем при анализе больших данных.
15. Прикладные задачи с применение технологий больших данных.
16. Визуализация больших данных.

6.4 Оценочные средства для самостоятельной работы и текущего контроля успеваемости

Тема 1 Введение в большие данные. Классификация задач анализа данных. Жизненный цикл аналитики данных

- 1) Изложите краткую историю больших данных.
- 2) Опишите подходы в работе с данными.
- 3) Какова этика работы с данными?
- 4) Опишите прогресс нейронных сетей.
- 5) В чем связь анализа данных и искусственного интеллекта?
- 6) Опишите принципы обработки больших данных.
- 7) Хранение и обработка структурированных данных. Базовые принципы.
- 8) Каковы базовые принципы использования инструментов для обработки структурированных данных: Excel и SQL?
- 9) Как осуществляется хранение и обработка неструктурированных данных?
- 10) Какие Вы знаете технологии для обработки неструктурированных данных?
- 11) Как производится кодировка категориальных данных?

Тема 2 Высокопроизводительные вычисления

- 1) Что такое Business Intelligence и Data Science?

2) Каковы базовые принципы, общее, отличия технологий для обработки больших данных: Python, Apache Hadoop, Apache Spark, Apache Storm.

4) Какие Вы знаете нейросетевые архитектуры?

5) Что такое машинное обучение?

Тема 3 Масштабирование и многоуровневое хранение данных

1) Что такое майнинг данных?

2) Что такое Deep Learning?

3) Что такое нейросетевые архитектуры?

4) Что такое обучение с подкреплением?

5) Что такое сверточные нейронные сети?

Тема 4 Сложные методы аналитики

1) Что такое линейная регрессия: определение, формулы?

2) Что такое логистическая регрессия: определение, формулы?

3) Что такое решающие деревья: определение, схема?

4) Как используется глубокое обучение в обработке текстов?

5) Как используется машинное обучение в лингвистике?

6.5 Вопросы для подготовки к экзамену

1) Что означает термин «Big Data» в информационных технологиях?

2) Что является основной целью обработки Big Data?

3) Кто и в каком году впервые ввел термин «Big Data»?

4) Какие главные характеристики Big Data?

5) Какие данные занимают больше мировой памяти относительно остальных?

6) Какие понятия содержит в себе принцип трех "V"?

7) Что является примером квази-структурированных данных?

8) Чем характеризуются "Большие данные"?

9) Что является главным результатом процесса Business Intelligence?

10) Что означает термин «Business Intelligence» в информационных технологиях?

11) Расшифруйте аббревиатуру OLAP?

12) Что относится к средствам предоставления информации в Business Intelligence?

13) Что относится к средствам интеграции в «Business Intelligence»?

14) Какие цели ставит перед собой Data Science?

15) Что такое жизненный цикл аналитики данных?

16) Дайте определение термину «предиктивное моделирование»?

- 17) Что такое ETL?
- 18) Какова роль BI-аналитика в проекте?
- 19) Что такое Apache Hadoop? В чем преимущества решений на базе Hadoop?
- 20) Что такое MapReduce? Какими достоинствами и недостатками обладает MapReduce?
- 21) Какому основному принципу следует HDFS? Какой размер блока по умолчанию в HDFS? Какие функции выполняет NameNode в HDFS?
- 22) Какой узел отвечает за репликацию данных в Hadoop? Какие компоненты содержит Slave узел в Hadoop? Какие компоненты содержит Master узел в Hadoop?
- 23) Какие компоненты являются частями HDFS?
- 24) Для чего используется автономный режим Hadoop?
- 25) Какой режим необходим для того, чтобы на локальной машине использовать Hadoop как кластер, состоящий из одного узла?
- 26) Что является отличительной особенностью NoSQL? В каком случае стоит применять NoSQL хранилища?
- 27) Что, согласно теореме CAP, возможно обеспечить в любой реализации распределённых вычислений?
- 28) Какое свойство означает, что транзакции не нарушают согласованность данных, то есть они переводят базу данных из одного корректного состояния в другое?
- 29) Какой способ хранения данных используется в MongoDB?
- 30) Что относится к плюсам репликации?
- 31) Что относится к преимуществам нереляционных БД?
- 32) Что такое шардинг?
- 33) Какие три свойства фигурируют в определении теоремы CAP?
- 34) Для чего нужна визуализация? В чем состоят основные задачи визуализации?
- 35) Какие традиционные виды визуализации Вы знаете?
- 36) Что такое дедупликация данных?
- 37) Какие требования предъявляются к визуализации? Какие типы визуализации Вы знаете?
- 38) Чем анализ больших данных отличается от традиционного анализа?
- 39) Какие основные типы Data Mining Вы знаете?
- 40) Какие категории Web Mining можно выделить? В чем основная задача Web Content Mining?

- 41) В чем состоят основные задачи интеллектуального анализа текстов?
- 42) К каким алгоритмам классификации относится метод ближайших соседей?
- 43) Что является целью кластеризации?
- 44) С помощью какого алгоритма можно найти ассоциативное правило?
- 45) Что подразумевается под определением "статистический вывод"?
- 46) Чем отличаются ошибки первого и второго рода?
- 47) Что является результатом решения задачи регрессии?
- 48) Что такое α -error?

6.6 Примерные темы курсовых работ

Курсовая работа не предусмотрена.

7. Учебно-методическое и информационное обеспечение дисциплины

7.1 Рекомендуемая литература

Основная литература

1. Келлехер, Д. Наука о данных: базовый курс : учебное пособие / Д. Келлехер. – Москва : Альпина Паблишер, 2020. – 224 с. URL: https://ugolok.vercel.app/books/ai_ds_bd/dzhon_kelleher_brendan_tirni_nauka_o_dannih_bazovii_kurs_alipina.pdf (Дата обращения 26.08.2024).
2. Уэс Маккинни. Python и анализ данных: Первичная обработка данных с применением pandas, NumPy и Jupiter / пер. с англ. А. А. Слинкина. 3-е изд. – М.: МК Пресс, 2023. – 536 с.: ил. URL: gstu.by/sites/default/files/files/resources/2021/09/makkini.pdf?ysclid=m0dp0a8sbk386511381 (Дата обращения 26.08.2024).

Дополнительная литература

1. Белов, В.С. Информационно-аналитические системы: основы проектирования и применения: учебно-практическое пособие / В.С. Белов. – Москва: Евразийский открытый институт, 2010. – 111с. URL: https://shpora1.do.am/_ld/2/255_-_pdf (Дата обращения 26.08.2024).

Учебно-методические материалы и пособия

1. Бизянов, Е.Е. Методы анализа данных: лабораторный практикум / Е.Е. Бизянов, А.С. Закутный ; Каф. Специализированных компьютерных систем. – Алчевск: ГОУ ВО ЛНР ДонГТИ, 2023. – 91 с. URL: <https://library.dstu.education/download.php?rec=132246>.

7.2 Базы данных, электронно-библиотечные системы, информационно-справочные и поисковые системы

1. Научная библиотека ДонГТУ : официальный сайт.— Алчевск. — URL: library.dstu.education.— Текст : электронный.
2. Научно-техническая библиотека БГТУ им. Шухова : официальный сайт. — Белгород. — URL: <http://ntb.bstu.ru/jirbis2/>.— Текст : электронный.
3. Консультант студента : электронно-библиотечная система.— Москва. — URL: <http://www.studentlibrary.ru/cgi-bin/mb4x> .— Текст : электронный.
4. Сайт кафедры ИСИБ <http://scs.dstu.education>.

8 Материально-техническое обеспечение дисциплины

Материально-техническая база обеспечивает проведение всех видов деятельности в процессе обучения, соответствует требованиям ФГОС ВО.

Материально-техническое обеспечение представлено в таблице 9.

Таблица 9 – Материально-техническое обеспечение

Наименование оборудованных учебных кабинетов	Адрес (местоположение) учебных кабинетов
Специальные помещения: <i>Мультимедийная аудитория. (60 посадочных мест), оборудованная специализированной (учебной) мебелью (скамья учебная – 20 шт., стол – 1 шт., доска аудиторная – 1 шт.), учебное ПК (монитор + системный блок), мультимедийная стойка с оборудованием – 1 шт., широкоформатный экран.</i> Аудитории для проведения лекций: <i>Компьютерные классы (22 посадочных места), оборудованный учебной мебелью, компьютерами с неограниченным доступом к сети Интернет, включая доступ к ЭБС:</i> ПК – 12 шт.; Доска – 1 шт.	ауд. <u>207</u> корп. <u>4</u> ауд. <u>211</u> корп. <u>4</u>

Лист согласования рабочей программы дисциплины

Разработал

И.о. заведующего кафедрой
интеллектуальных систем и
информационной безопасности
(должность)


(подпись)

Е.Е. Бизянов
(Ф.И.О.)

И.о. заведующего кафедрой
интеллектуальных систем и
информационной безопасности
(наименование кафедры)


(подпись)

Е.Е. Бизянов
(Ф.И.О.)

Протокол № 1 заседания кафедры
интеллектуальных систем и
информационной безопасности

от 27.08.20 24 г.

И.о. декана факультета
информационных технологий и
автоматизации производственных
процессов


(подпись)

В.В. Дьячкова
Ф.И.О.)

Согласовано

Председатель методической комиссии
По специальности 10.05.03
Информационная безопасность
Автоматизированных систем


(подпись)

Е.Е.Бизянов
Ф.И.О.)

Начальник учебно-методического центра


(подпись)

О.А.Коваленко
Ф.И.О.)

Лист изменений и дополнений

Номер изменения, дата внесения изменения, номер страницы для внесения изменений	
ДО ВНЕСЕНИЯ ИЗМЕНЕНИЙ:	ПОСЛЕ ВНЕСЕНИЯ ИЗМЕНЕНИЙ:
Основание:	
Подпись лица, ответственного за внесение изменений	