

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

Инжиниринг данных
(наименование дисциплины)

09.04.01 Информатика и вычислительная техника
(код, наименование специальности)

Искусственный интеллект и цифровые двойники предприятий
(программа)

Квалификация магистр
(бакалавр/специалист/магистр)

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

Алчевск, 2024

1 Цели и задачи изучения дисциплины

Цели дисциплины. Целью изучения дисциплины «Инжиниринг данных» является предоставить студентам теоретические знания и практические навыки в области представления и обработки знаний в информационных системах, поиска и принятия решений в системах обработки информации.

Задачи изучения дисциплины:

- приобретения навыков формального представления знаний средствами различных моделей и программной реализации элементов систем хранения и обработки знаний и экспертных систем (ЭС);
- изучение основных принципов создания интеллектуальных информационных систем и систем поддержки принятия решений в различных предметных областях.

Дисциплина направлена на формирование общепрофессиональной (ПК - 4) компетенции выпускника.

2 Место дисциплины в структуре образовательной программы

Логико-структурный анализ дисциплины – курс входит обязательную часть БЛОКА 1 «Дисциплины (модули)» подготовки студентов по специальности 09.04.01 Информатика и вычислительная техника (Искусственный интеллект и цифровые двойники предприятий).

Дисциплина реализуется кафедрой интеллектуальных систем и информационной безопасности. Основывается на базе дисциплин: «Анализ больших данных», «Цифровые двойники предприятий» изученных обучающимися при прохождении подготовки по программе бакалавриата (специалитета) и магистратуры.

Является основой для изучения следующих дисциплин: «Интеллектуальные системы поддержки принятия решений», «Экспертные системы», «Выполнение и защита выпускной квалификационной работы».

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

Курс является фундаментом для ориентации студентов в сфере научных исследований.

Общая трудоемкость освоения дисциплины составляет 3 зачетных единицы, 108 ак.ч. Программой дисциплины предусмотрены лекционные (18 ч.), лабораторные (18 ч.) занятия и самостоятельная работа студента (72 ч.), в том числе курсовая работа.

Дисциплина изучается на 1 курсе во 2 семестре. Форма промежуточной аттестации дисциплины – экзамен, курсовая работа – дифференцированный зачет.

3 Перечень результатов обучения по дисциплине, соотнесённых с планируемыми результатами освоения образовательной программы

Процесс изучения дисциплины «Инжиниринг данных» направлен на формирование компетенции, представленной в таблице 1.

Таблица 1 –Компетенции, обязательные к освоению

Содержание компетенции	Код компетенции	Код и наименование индикатора достижения компетенции
Способен руководить проектами по созданию комплексных систем на основе аналитики больших данных в различных отраслях	ПК-4	ПК-4.2. Применяет варианты использования больших данных, определений, словарей и эталонной архитектуры больших данных при руководстве проектами по построению комплексных систем на основе аналитики больших данных в различных отраслях

4 Объём и виды занятий по дисциплине

Общая трудоёмкость учебной дисциплины составляет 3 зачётных единиц, 108 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 дифференцированному зачету.

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
		2
Аудиторная работа, в том числе:	36	36
Лекции (Л)	18	18
Практические занятия (ПЗ)	-	-
Лабораторные работы (ЛР)	18	18
Курсовая работа/курсовой проект	-	-
Самостоятельная работа студентов (СРС), в том числе:	72	72
Подготовка к лекциям	4	4
Подготовка к лабораторным работам	16	16
Подготовка к практическим занятиям / семинарам	-	-
Выполнение курсовой работы / проекта	20	20
Расчетно-графическая работа (РГР)	-	-
Реферат (индивидуальное задание)	-	-
Домашнее задание	-	-
Подготовка к контрольным работам	-	-
Подготовка к коллоквиуму	-	-
Аналитический информационный поиск	-	-
Работа в библиотеке	-	-
Подготовка к экзамену (диф.зачету)	32	32
Промежуточная аттестация – экзамен (Э), диф.зачет (ДЗ)	Э, ДЗ	Э, ДЗ
Общая трудоемкость дисциплины		
ак.ч.	108	108
з.е.	3	3

5 Содержание дисциплины

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

- тема 1 (Работа с данными в Python);
- тема 2 (Подготовка данных для систем машинного обучения);
- тема 3 (Параллельная и распределенная обработка данных).

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	Работа с данными в Python	Введение. Вопросы, решаемые при представлении знаний. Библиотеки для работы с данными в различных форматах в Python: файлы CSV, JSON, HTML. Работа с базами данных в Python. Работа с изображениями, видео и звуковыми файлами. Форматы хранения больших данных и работа с ними: Parquet, Avro. Графы знаний.	6	-	-	Библиотеки для работы с данными в Python Работа с базами данных в Python.	2 2
2	Подготовка данных для систем машинного обучения	Сбор данных и формирование набора данных для систем машинного обучения. Загрузка данных из интернет и социальных сетей. Методы очистки и подготовки данных. Очистка и подготовка данных на Python. Разметка данных. Общедоступные платформы для хранения данных. Подход Data-Centric AI.	6	-	-	Работа с текстовыми файлами разных форматов в Python Работа с изображениями, видео и звуковыми файлами в Python. Работа с файлами для хранения больших данных в Python.	2 2 2

Окончание таблицы 3

1	2	3	4	5	6	7	8
3	Параллельная и распределенная обработка данных	Архитектура центров обработки данных, кластеры для параллельных и распределенных вычислений. Экосистема для распределенного хранения и обработки больших объемов данных: ApacheHadoop, HDFS. Распределенная обработка данных в ApacheSpark. Архитектура ApacheSpark: ResilientDistributedDataset (RDD), действия трансформации. Работа с данными с использованием SparkDataFrame. Источники данных для SparkDataFrame. Обработка данных в SparkDataFrame. Использование SQL в SparkDataFrame.	6	-	-	Создание собственных наборов данных в Python. Очистка и подготовка данных. Работа с данными в ApacheSpark. Использование SQL в ApacheSpark.	4 2 2
Всего аудиторных часов		18		-		18	

6 Фонд оценочных средств для проведения текущего контроля успеваемости и промежуточной аттестации студентов по дисциплине

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
Выполнение курсовой работы	Предоставление материалов согласно календарного плана	36 – 60
Защита курсовой работы	Защита курсовой работы	24 – 40
Итого по курсовой работе	—	60 - 100

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

Дифференцированный зачет по курсовой работе проводится по результатам работы в семестре. В случае если полученная в семестре сумма баллов не устраивает студента, на зачетной неделе 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 Домашнее задание

Домашнее задание не предусмотрено.

6.3 Темы для рефератов (презентаций) – индивидуальное задание

Реферат (индивидуальное задание) не предусмотрен.

6.4 Оценочные средства для самостоятельной работы и текущего контроля успеваемости

Тема 1 (Работа с данными в Python)

- 1) Какие вопросы решаются при представлении знаний?
- 2) Какие библиотеки для работы с данными в различных форматах в Python Вы знаете?
- 3) Какие форматы хранения больших данных Вы знаете?
- 4) Какую работу с файлами для хранения больших данных можно провести в Python.
- 5) Что такое графы знаний.?

Тема 2 (Подготовка данных для систем машинного обучения)

- 1) Как формируется набор данных для систем машинного обучения?
- 2) Как производится загрузка данных из интернет и социальных сетей?
- 3) Какие методы очистки и подготовки данных Вы знаете?
- 4) Как производится разметка данных?
- 5) Какие общедоступные платформы для хранения данных Вы знаете?

Тема 3 (Параллельная и распределенная обработка данных)

- 1) Что из себя представляют центры обработки данных?
- 2) Что из себя представляют кластеры для параллельных вычислений?
- 3) Что из себя представляют кластеры для распределенных вычислений?
- 4) Что такое экосистема для распределенного хранения больших объемов данных?
- 5) Что такое экосистема для распределенной обработки больших объемов данных?

6.5 Вопросы для подготовки к экзамену

- 1) 1. Что такое реляционная база данных?
- 2) Какие примеры реляционных СУБД можете привести?
- 3) Что такое CASE инструмент?
- 4) Какие ключевые понятия проектирования реляционной базы данных Вам известны?
- 5) Что из себя представляет реляционная модель?
- 6) Что такое ключ и значения NULL?
- 7) Что такое отношение?
- 8) Что такое сущность-связь?
- 9) Что такое автономная транзакция?
- 10) Что из себя представляет процедура ведения логгирования
- 11) Что такое Ad-hoc запросы?
- 12) Что такое последовательности?
- 13) Как используются последовательности в коде приложения?
- 14) Что из себя представляет оператор SELECT в реляционной алгебре?
- 15) Как сгенерировать SQL скрипт в CASE средстве на основе придуманной логической модели и создать физическую модель данных?
- 16) Что из себя представляет оператор PROJECT в реляционной алгебре?
- 17) Что такое пользователь и схема в Oracle
- 18) Что из себя представляет операция CartesianProduct в реляционной алгебре?
- 19) Что из себя представляет отношение NaturalJoin в реляционной алгебре?
- 20) Как написать SQL запрос bulk-load загрузки данных?
- 21) Что из себя представляет операция ThetaJoin в реляционной алгебре?
- 22) Как написать анонимный блок bulk-load загрузки данных при помощи цикла?
- 23) Что из себя представляет Unionоператор в реляционной алгебре?
- 24) Как написать все DML выражения на таблицу?
- 25) Что из себя представляет Difference оператор в реляционной алгебре?
- 26) Как написать пакет для выбранной таблицы с процедурой получения данных по идентификатору?
- 27) Что из себя представляет Intersection оператор в реляционной алгебре?
- 28) Как создать B-Tree индекс на таблицу?
- 29) Что из себя представляет Rename оператор в реляционной алгебре?
- 30) Как написать триггер, генерирующий уникальный ключ на поле?
- 31) Как осуществляется представление выражения в реляционной алгебре?

- 32) Как написать триггер, логирующий действия пользователя на данных?
- 33) Как создать Bitmap индекс на таблицу?
- 34) Что такое нормальная форма Бойса Кодда?
- 35) Как написать анонимный блок bulk-load загрузки данных при помощи цикла?
- 36) Что такое функциональные зависимости?
- 37) Как осуществляется генерация идентификатора ключа таблицы при помощи агрегирующей функции?
- 38) Как осуществляется генерация идентификатора ключа таблицы при помощи системного времени?

6.6 Тематика и содержание курсовой работы

Курсовая работа выполняется во втором семестре.

Тематикой курсовой работы является подготовка набора данных для индивидуальной задачи. Индивидуальное задание студент выбирает самостоятельно.

Подготовить собственный набор данных. Выбрать задачу в одном из направлений создания системы искусственного интеллекта (компьютерное зрение, обработка естественного языка) и подготовить для этой задачи набор данных для обучения с учителем. Собрать и очистить данные, провести разметку. Готовый набор данных разместить на одной из общедоступных платформ для хранения данных по своему выбору. Подготовить документацию к созданному набору данных.

Перечень задач, для которых рекомендуется подготовить набор данных:

- классификация объектов на изображениях;
- определение положения объектов на изображениях;
- определение положения объектов в видео;
- классификация текста на русском языке;
- определение эмоциональной окраски текста на русском языке.

Создайте набор данных и проведите его исследование.

7 Учебно-методическое и информационное обеспечение дисциплины

7.1 Рекомендуемая литература

Основная литература

1. Агальцов В. П. Базы данных: в 2 книгах. Книга 2. Распределенные и удаленные базы данных: учебник / В.П. Агальцов. — Москва : ФОРУМ : ИНФРА-М, 2021. — 271 с. — (Высшее образование: Бакалавриат). - ISBN 978-5 8199-0713-9. – Текст: электронный. – [Электронный ресурс] – URL:<https://znanium.com/catalog/product/1514118>. Режим доступа: по подписке.(Дата обращения 26.08.2024).

2. Мартишин С.А. Проектирование и реализация баз данных в СУБД MySQL с использованием MySQLWorkbench: учебное пособие / С.А. Мартишин, В.Л. Симонов, М.В. Храпченко. — Москва : ФОРУМ : ИНФРА-М, 2022. — 160 с. — (Высшее образование). - ISBN 978-5-8199-0517-3. – [Электронный ресурс] – URL: <https://znanium.com/catalog/product/1815962>- Режим доступа: по подписке.(Дата обращения 26.08.2024).

Дополнительная литература

1. Рашка С. Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения / С. Рашка. - Москва : ДМК Пресс, 2017. - 418 с. - ISBN 978-5-97060-409-0. - [Электронный ресурс]. – URL: <https://www.studentlibrary.ru/book/ISBN9785970604090.html> Режим доступа : по подписке.(Дата обращения 26.08.2024).

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

1. Научная библиотека ДонГТУ : официальный сайт.— Алчевск. — URL: <https://library.dontu.ru>.— 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. Университетская библиотека онлайн : электронно-библиотечная система.— URL: http://biblioclub.ru/index.php?page=main_ub_red. — Текст : электронный.

5. Сайт кафедры ИСИБ <https://scs.dontu.ru>

Материально-техническое обеспечение представлено в таблице 6.

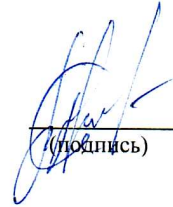
Таблица 6 – Материально-техническое обеспечение

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:	ауд. <u>207</u> корп. <u>4</u>
<i>Компьютерные классы (22 посадочных места), оборудованный учебной мебелью, компьютерами с неограниченным доступом к сети Интернет, включая доступ к ЭБС:</i>	ауд. <u>217</u> корп. <u>3</u> ауд. <u>211</u> корп. <u>4</u>

Лист согласования РПД

Разработал:

ст. преподаватель кафедры
интеллектуальных систем и
информационной безопасности
(должность)


(подпись)

Р.Н. Погорелов
(Ф.И.О.)

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


(подпись)

Е.Е. Бизянов
(Ф.И.О.)

Протокол № 1 заседания кафедрыот 27.08. 2024г.

И.о. декана факультета
информационных технологий
и автоматизации производственных
процессов:
(наименование факультета)


(подпись)

В.В. Дьячкова
(Ф.И.О.)

Согласовано

Председатель методической комиссии по
направлению подготовки 09.04.01
Информатика и вычислительная техника


(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	