МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Астраханский государственный университет имени В. Н. Татищева»
(Астраханский государственный университет им. В. Н. Татищева)

СОГЛАСОВАНО
Руководитель ОПОП

А.Н. Марьенков

«05» мяа 2025 г.

УТВЕРЖДАЮ
И.о. зав. кафедрой информационных технологии

О. Н. Выборнова

«05» мая 2025 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Структурирование, разметка и обогащение данных

Составитель(и)	**Синельщиков А.В., доцент кафедры информационных технологий**
Согласовано с работодателями:	**Механич А.П., Заместитель директора ГБУ АО «Инфраструктурный центр электронного правительства»; Проталинский И.О., Технический директор ООО «Бест плюс»**
Направление подготовки / специальность	**09.04.04 Программная инженерия**
Направленность (профиль) ОПОП	**Проектирование и разработка систем искусственного интеллекта**
Квалификация (степень)	**Магистр**
Форма обучения	**Очная**
Год приёма	**2025**
Курс	**2**
Семестр(ы)	**4**

Астрахань – 2025 г.

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)

1.1. Целями освоения дисциплины (модуля): «Структурирование, разметка и обогащение данных» является формирование у слушателей компетенций в области анализа изображений и видео, а также анализа естественного языка с помощью методов искусственного интеллекта.

1.2. Задачи освоения дисциплины (модуля):

- изучение подходов и приобретение практических навыков в выборе и применении методов структурирования знаний для предметных областей;
- изучение подходов и приобретение практических навыков в выборе и применении методов представления знаний с помощью логических и продукционных методов, семантических сетей и фреймов, объектно-ориентированных методов;
- приобретение практических навыков выбора и применения методов обработки и распространения знаний для разработки программных компонентов систем, основанных на знаниях, и приложений

2. МЕСТО ДИСЦИПЛИНЫ (МОДУЛЯ) В СТРУКТУРЕ ОПОП

2.1. Учебная дисциплина (модуль) «Структурирование, разметка и обогащение данных» относится к элективным дисциплинам и осваивается в 4 семестре для очной формы обучения.

2.2. Для изучения данной учебной дисциплины (модуля) необходимы следующие знания, умения, навыки, формируемые предшествующими учебными дисциплинами (модулями):

- Методы машинного обучения;
- Архитектура систем искусственного интеллекта;
- Методология инженерии программных средств искусственного интеллекта;
- Обработка и анализ данных;

Знания

- Синтаксис Python и принципы объектно-ориентированного программирования.
- Теоретические основы и этапы жизненного цикла обработки данных: сбор, очистка, преобразование, агрегация и хранение.
- Основы работы с ключевыми библиотеками для анализа данных и научных вычислений.
- Понимание роли качественной подготовки данных в процессе построения ML-модели, включая разделение данных, концепции переобучения и недообучения, а также основные метрики качества.
- Знание ключевых структур данных, используемых в анализе данных (массивы, таблицы, словари).

Умения

- Выполнять полный цикл предварительной обработки данных (Data Preprocessing): загрузка, очистка (обработка пропусков, выбросов), преобразование (нормализация, кодирование) и агрегация данных.
- Проводить исследовательский анализ данных (EDA): выявлять закономерности, строить гипотезы и интерпретировать результаты визуализации данных.
- Применять специализированные библиотеки для манипулирования разнородными наборами данных.
- Выполнять векторизованные вычисления для эффективной реализации математических операций и обработки данных.
- Корректно разделять выборку и применять адекватные метрики для оценки качества моделей.

Навыки

- Владение интерактивными средами (по типу "notebook") для быстрого прототипирования, визуализации данных и проведения экспериментов по их обработке.
- Навыки отладки и валидации кода, связанного с алгоритмами обработки данных и

реализацией моделей.
- Владение системами контроля версий на базовом уровне для управления кодом проектов по анализу данных.
- Навыки работы со средами разработки (IDE) и виртуальными окружениями для управления зависимостями проектов.

2.3. Последующие учебные дисциплины (модули) и (или) практики, для которых необходимы знания, умения, навыки, формируемые данной учебной дисциплиной (модулем):
- Выпускная квалификационная работа;

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

Процесс изучения дисциплины (модуля) направлен на формирование элементов следующих компетенций в соответствии с ФГОС ВО и ОПОП ВО по данному направлению подготовки:
а) профессиональных:

ПК-2. Способен разрабатывать алгоритмы и программные средства для решения задач в области создания и применения искусственного интеллекта.

ПК-5. Способен руководить процессами разработки программного обеспечения.

Таблица 1 – Декомпозиция результатов обучения

Код компетенции	Код и наименование индикатора достижения компетенции	Планируемые результаты освоения дисциплины (модуля)		
		Знать (1)	Уметь (2)	Владеть (3)
ПК–2	ПК-2.1 Знать современные тенденции, технологии и версии программного обеспечения (ПО).	современные тенденции, технологии и актуальные версии программного обеспечения в сфере искусственного интеллекта	анализировать, сравнивать и выбирать релевантные технологии и ПО для решения задач в области ИИ	навыками мониторинга и систематизации информации о новых разработках для их практического применения.
	ПК-2.2 Уметь применяет инструментальные среды, программно-технические платформы для решения задач в области создания и применения искусственного интеллекта.	назначение и функциональные возможности современных инструментальных сред и программно-технических платформ для ИИ	обоснованно выбирать и применять эти средства для реализации алгоритмов и решения практических задач	навыками использования выбранных платформ и сред в процессе создания систем искусственного интеллекта
	ПК-2.3 Владеть навыками разработки оригинальных	методологии и жизненный цикл разработки программного	проектировать, кодировать, тестировать и документировать	навыками самостоятельной разработки и внедрения

	программных средств для решения задач в области создания и применения искусственного интеллекта.	обеспечения, а также принципы создания оригинальных алгоритмов в области ИИ	разрабатываемые программные средства	оригинальных программных решений для прикладных задач искусственного интеллекта
ПК–5	ПК-5.1 Знать инструменты и методы управления выпуском и поставкой проектов в области IT.	инструменты и методы управления выпуском и поставкой IT-проектов, применимые к задачам обработки данных	выбирать и использовать эти инструменты для организации процессов структурирования, разметки и обогащения данных	навыками интеграции процессов управления данными в общий цикл разработки и поставки программного обеспечения.
	ПК-5.2 Уметь проводить анализ мнений и замечаний заказчиков по выполнению проекта.	методы сбора и анализа обратной связи от заказчиков, а также стандарты качества, применяемые к проектам по обработке данных	проводить систематический анализ мнений и замечаний, выявляя их влияние на процессы структурирования, разметки и обогащения данных	навыками документирования результатов анализа и разработки корректирующих мер для управления качеством данных в соответствии с требованиями заказчика.
	ПК-5.3 Владеть навыками формулирования решений по внесению изменений в IT-проекты по согласованию с заказчиками.	методы управления изменениями, процедуры согласования с заказчиками и способы оценки влияния изменений на проекты по обработке данных	анализировать запросы заказчика, формулировать технически и экономически обоснованные решения по внесению изменений в процессы структурирования или разметки данных	навыками ведения переговоров, аргументации предложенных решений и их документального оформления для согласования с заказчиком

4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Общая трудоемкость дисциплины в соответствии с учебным планом составляет 3 зачетных единиц (108 часа).

Трудоемкость отдельных видов учебной работы студентов очной формы обучения приведена в таблице 2.1.

Таблица 2.1. Трудоемкость отдельных видов учебной работы по формам обучения

Вид учебной и внеучебной работы	для очной формы обучения
Объем дисциплины в зачетных единицах	3
Объем дисциплины в академических часах	108
Контактная работа обучающихся с преподавателем (всего), в том числе (час.):	25,25
- занятия лекционного типа, в том числе:	12

Вид учебной и внеучебной работы	для очной формы обучения
- практическая подготовка (если предусмотрена)	–
- занятия семинарского типа (семинары, практические, лабораторные), в том числе: - практическая подготовка (если предусмотрена)	12 –
- в ходе подготовки и защиты курсовой работы[1]	–
- консультация (предэкзаменационная)[2]	1
- промежуточная аттестация по дисциплине[3]	0,25
Самостоятельная работа обучающихся (час.)	82,75
Форма промежуточной аттестации обучающегося (зачет/экзамен), семестр (ы)	Экзамен – 4 семестр

Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий и самостоятельной работы, для каждой формы обучения представлено в таблице 2.2.

Таблица 2.2. Структура и содержание дисциплины (модуля)

для очной формы обучения

Раздел, тема дисциплины (модуля)	Контактная работа, час.							СР, час.	Итого часов	Форма текущего контроля успеваемости, форма промежуточной аттестации
	Л		ПЗ		ЛР		КР / КП			
	Л	в т.ч. ПП	ПЗ	в т.ч. ПП	ЛР	в т.ч. ПП				
Семестр 4.										
Тема 1. Основы структурирования данных и форматы представления	***2***				***2***			***16***	***20***	Отчет по ЛПР; вопросы к экзамену.
Тема 2. Методологии и технологии разметки данных (аннотирования)	***2***				***2***			***16***	***20***	Отчет по ЛПР; вопросы к экзамену.
Тема 3. Техники и источники обогащения данных	***2***				***2***			***16***	***20***	Отчет по ЛПР; вопросы к экзамену.
Тема 4. Инструментальные средства и платформы для подготовки данных	***2***				***2***			***16***	***20***	Отчет по ЛПР; вопросы к экзамену.
Тема 5. Управление качеством и жизненным циклом	***4***				***4***			***18.75***	***26.75***	Отчет по ЛПР; вопросы к экзамену.

[1] Числовые данные в данной строке соответствуют трудоемкости, указанной в учебном плане в столбце «КР/КП» Если курсовая работа не предусмотрена – необходимо удалить строку «Контактная работа в ходе подготовки и защиты курсовой работы».

[2] Числовые данные в данной строке соответствуют трудоемкости, указанной в учебном плане в столбце «Конс. (для гр.)»

[3] Числовые данные в данной строке соответствуют трудоемкости, указанной в учебном плане в столбце «КПА»

Раздел, тема дисциплины (модуля)	Контактная работа, час.							СР, час.	Итого часов	Форма текущего контроля успеваемости, форма промежуточной аттестации
	Л		ПЗ		ЛР		КР / КП			
	Л	в т.ч. ПП	ПЗ	в т.ч. ПП	ЛР	в т.ч. ПП				
данных в проектах.										
Консультации									***1***	
Контроль промежуточной аттестации									***0.25***	**Экзамен**
ИТОГО за семестр:	***12***	***0***	***0***	***0***	***12***	***0***	***0***	***82.75***	***108***	
Итого за весь период	***12***	***0***	***0***	***0***	***12***	***0***	***0***	***82.75***	***108***	

Примечание: Л – лекция; ПЗ – практическое занятие, семинар; ЛР – лабораторная работа; ПП – практическая подготовка; КР / КП – курсовая работа / курсовой проект; СР – самостоятельная работа

Таблица 3 – Матрица соотнесения разделов, тем учебной дисциплины (модуля) и формируемых в них компетенций

Разделы, темы дисциплины (модуля)	Кол-во часов	Компетенции		Общее количество компетенций
		ПК-2	ПК-5	
Тема 1. Основы структурирования данных и форматы представления	***20***	+	+	2
Тема 2. Методологии и технологии разметки данных (аннотирования)	***20***	+	+	2
Тема 3. Техники и источники обогащения данных	***20***	+	+	2
Тема 4. Инструментальные средства и платформы для подготовки данных	***20***	+	+	2
Тема 5. Управление качеством и жизненным циклом данных в проектах.	***26.75***	+	+	2
Консультации	***1***	+	+	2
Контроль промежуточной аттестации	***0,25***	+	+	
Итого	**108**			**2**

Краткое содержание каждой темы дисциплины (модуля)

Тема 1. Основы структурирования данных и форматы представления

Типы данных (структурированные, полуструктурированные, неструктурированные), модели данных (реляционные, NoSQL, графовые), форматы представления (JSON, XML, CSV, Parquet, Protobuf), сериализация и десериализация, принципы нормализации и денормализации.

Тема 2. Методологии и технологии разметки данных (аннотирования)

Задачи разметки (классификация, детекция, сегментация, NER), типы разметки (ручная, полуавтоматическая, автоматическая), краудсорсинг, разметка "Human-in-the-loop" (человек-в-цикле), метрики качества разметки (Inter-Annotator Agreement).

Тема 3. Техники и источники обогащения данных

Внутреннее обогащение (генерация признаков, feature engineering), внешнее обогащение

(использование открытых данных, API), парсинг веб-данных (web scraping), интеграция данных из различных источников (data fusion), использование баз знаний и онтологий.

Тема 4. Инструментальные средства и платформы для подготовки данных

Обзор платформ для разметки (Label Studio, CVAT, Labelbox), библиотеки Python для обработки данных (Pandas, NumPy, Dask), инструменты ETL/ELT (Apache Airflow, NiFi), облачные сервисы (AWS Glue, Google Data Prep), системы версионирования данных (DVC)..

Тема 5. Управление качеством и жизненным циклом данных в проектах

Метрики качества данных (полнота, точность, актуальность, непротиворечивость), методы очистки данных (data cleaning), профилирование данных (data profiling), управление данными (Data Governance), построение конвейеров (pipelines) обработки данных, обеспечение безопасности и конфиденциальности.

5. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ПРЕПОДАВАНИЮ И ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

5.1. Указания для преподавателей по организации и проведению учебных занятий по дисциплине (модулю)

Учебная деятельность студента в процессе изучения строится из контактных форм работы с преподавателем (аудиторные занятия, зачет) и самостоятельной работы.

Для успешного освоения дисциплины является обязательным посещение всех занятий, выполнение самостоятельной работы, которая назначается преподавателем.

Методическая поддержка дисциплины обеспечивается использованием дистанционных технологий. Студентам предлагается информационный ресурс, расположенный по адресу: http://moodle.asu-edu.ru, на сервере дистанционного обучения АГУ. На сервере размещен методический материал по данной дисциплине, в содержание которого входит:

– теоретический материал;

– задания и указания по выполнению лабораторных работ;

– вопросы к экзамену.

Аудиторные занятия проводятся на основе теоретического материала, опубликованного на образовательном портале, это позволяет студентам изучить пропущенный материал или самостоятельно разобраться с темой, не освоенной на занятии.

5.2. Указания для обучающихся по освоению дисциплины (модулю)

Самостоятельная работа студентов по дисциплине Структурирование, разметка и обогащение данных предполагает:

– изучение обязательных литературных источников;

– поиск и обзор дополнительных источников литературы и электронных источников информации по программе курса;

– самоконтроль изученного учебного материала в виде тестирования;

– подготовка рефератов и их презентаций;

– выполнение практических и лабораторных работ;

– подготовка к экзамену.

Таблица 4 – Содержание самостоятельной работы обучающихся

для очной формы обучения

Вопросы, выносимые на самостоятельное изучение	Кол-во часов	Форма работы
Тема 1.	***16***	Устный опрос

1. Сравнение производительности бинарных форматов (Parquet, Avro, Protobuf) при работе с большими данными (Big Data). 2. Исследование и применение схем валидации данных (Data Schema Validation) для JSON (JSON Schema) и XML (XSD).		
Тема 2. 1. Изучение методов активного обучения (Active Learning) для снижения затрат на ручную разметку. 2. Практическое применение и расчет метрик согласия разметчиков (Inter-Annotator Agreement), таких как Каппа Коэна и Альфа Криппендорфа.	***16***	Устный опрос
Тема 3. 1. Освоение инструментов веб-скрапинга (например, Scrapy, BeautifulSoup) для сбора данных из открытых источников. 2. Использование API публичных баз знаний (например, Wikidata, DBpedia) для семантического обогащения наборов данных.	***16***	Устный опрос
Тема 4. 1. Развертывание и настройка собственного экземпляра платформы разметки данных (например, Label Studio или CVAT). 2. Практическое использование DVC (Data Version Control) для версионирования наборов данных и моделей машинного обучения.	***16***	Устный опрос
Тема 5. 1. Внедрение автоматизированного профилирования и валидации данных с использованием библиотек (например, Great Expectations или Pandera). 2. Изучение техник анонимизации и псевдонимизации данных для соблюдения требований конфиденциальности (например, GDPR, ФЗ-152)	***18.75***	Устный опрос

5.3. Виды и формы письменных работ, предусмотренных при освоении дисциплины (модуля), выполняемые обучающимися самостоятельно

Отчет по лабораторным работам.

Результатом работы, выполняемой студентами, является электронный отчет по выполнению лабораторно-практической работы, тематика которых представлена в таблице 4,

Электронный отчет представляет собой файл формата Word (PDF, ODT), содержащий диаграммы, схемы и текстовые пояснения. Файл передается на проверку преподавателю путем загрузки на ресурс http://moodle.asu-edu.ru в соответствующий заданию раздел.

Задания к лабораторно-практическим занятиям размещены на образовательном портале http://moodle.asu-edu.ru. Рекомендуется заранее ознакомиться с темой, основными вопросами, рекомендациями, требованиями к представлению отчета и критериями оценивания заданий.

В процессе подготовки к лабораторно-практическим занятиям, необходимо обратить особое внимание на самостоятельное изучение рекомендованной литературы. Самостоятельная работа с учебниками, учебными пособиями, научной, справочной литературой, материалами периодических изданий и Интернета является наиболее эффективным методом получения дополнительных знаний, позволяет значительно активизировать процесс овладения информацией,

способствует более глубокому усвоению изучаемого материала.

6. ОБРАЗОВАТЕЛЬНЫЕ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

6.1. Образовательные технологии

Цели дисциплины достигаются путем сочетания контактной и самостоятельной работы студентов: проведение практических занятий, проведения лабораторных занятий на ПК и организации самостоятельной работы студентов.

Для самостоятельного изучения теоретического материала дисциплины рекомендуется использовать интернет-ресурсы, информационные базы, методические разработки, специальную учебную и научную литературу.

В рамках организации самостоятельной работы студентам рекомендуется:

– дополнительная подготовка к занятиям

– подготовка к текущей и промежуточной аттестации (экзамену).

Таблица 5 – Образовательные технологии, используемые при реализации учебных занятий

Раздел, тема дисциплины (модуля)	Форма учебного занятия		
	Лекция	Практическое занятие, семинар	Лабораторная работа
Тема 1. Основы структурирования данных и форматы представления	Классическая лекция	Не предусмотрено	Лабораторная работа 1
Тема 2. Методологии и технологии разметки данных (аннотирования)	Классическая лекция	Не предусмотрено	Лабораторная работа 2
Тема 3. Техники и источники обогащения данных	Классическая лекция	Не предусмотрено	Лабораторная работа 3
Тема 4. Инструментальные средства и платформы для подготовки данных	Классическая лекция	Не предусмотрено	Лабораторная работа 4
Тема 5. Управление качеством и жизненным циклом данных в проектах.	Классическая лекция	Не предусмотрено	Лабораторная работа 5, 6

6.2. Информационные технологии

При реализации различных видов учебной и внеучебной работы используются следующие информационные технологии:

- использование образовательного сайта http://moodle.asu-edu.ru, как элемента взаимодействия участников образовательного процесса (технологии дистанционного обучения);
- использование электронных учебников электронных библиотечных систем, доступ к которым предоставляется университетом;
- использование как источников информации сайтов, находящихся в Интернете в открытом доступе (электронные библиотеки, журналы, книги, психологические тесты);
- использование возможностей корпоративной электронной почты (рассылка заданий, материалов, ответы на вопросы).

6.3. Программное обеспечение, современные профессиональные базы данных и информационные справочные системы

6.3.1. Программное обеспечение

Наименование программного обеспечения	Назначение
Adobe Reader	Программа для просмотра электронных документов
Платформа дистанционного обучения LMS Moodle	Виртуальная обучающая среда
Mozilla FireFox	Браузер
Google Chrome	Браузер
VirtualBox	Программный продукт виртуализации операционных систем
VMware (Player)	Программный продукт виртуализации операционных систем

6.3.2. Современные профессиональные базы данных и информационные справочные системы

Наименование программного обеспечения	Назначение
Python	Язык программирования

7. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

7.1. Паспорт фонда оценочных средств

При проведении текущего контроля и промежуточной аттестации по дисциплине (модулю) Структурирование, разметка и обогащение данных проверяется сформированность у обучающихся компетенций, указанных в разделе 3 настоящей программы. Этапность формирования данных компетенций в процессе освоения образовательной программы определяется последовательным освоением дисциплин (модулей) и прохождением практик, а в процессе освоения дисциплины (модуля) – последовательным достижением результатов освоения содержательно связанных между собой разделов, тем.

Таблица 6 – Соответствие разделов, тем дисциплины (модуля), результатов обучения по дисциплине (модулю) и оценочных средств

Контролируемые разделы, темы дисциплины (модуля)	Код контролируемой компетенции	Наименование оценочного средства
Тема 1. Основы структурирования данных и форматы представления	ПК-2, ПК-5	Отчет по ЛПР; вопросы к экзамену.
Тема 2. Методологии и технологии разметки данных (аннотирования)	ПК-2, ПК-5	Отчет по ЛПР; вопросы к экзамену.
Тема 3. Техники и источники обогащения данных	ПК-2, ПК-5	Отчет по ЛПР; вопросы к экзамену.
Тема 4. Инструментальные средства и платформы для подготовки данных	ПК-2, ПК-5	Отчет по ЛПР; вопросы к экзамену.
Тема 5. Управление качеством и жизненным циклом данных в проектах.	ПК-2, ПК-5	Отчет по ЛПР; вопросы к экзамену.

Для оценивания результатов обучения в виде **знаний** используются следующие типы контроля:

- индивидуальное собеседование (устный опрос).
- письменные работы (отчеты по ЛПР).

Тестовые задания охватывают содержание всего пройденного материала. Индивидуальное собеседование ё проводится по разработанным вопросам к экзамену.

7.2. Описание показателей и критериев оценивания компетенций, описание шкал оценивания

В системе Moodle балл за выполнение лабораторно-практической работы выставляется в 100-балльной шкале комплексно с учетом степени подготовки студента к выполнению работы, объема выполненной работы на занятии и оформлении отчета в соответствии с перечисленными критериями. В зависимости от выставленного максимального балла (табл. 6) перерасчет за каждый отчет ЛР начисляемых баллов производится автоматически. Итоговый балл за отчеты по лабораторным работам является числом от 0 до 50 баллов.

Таблица 7 – Показатели оценивания результатов обучения в виде знаний

Шкала оценивания	Критерии оценивания
5 «отлично»	демонстрирует глубокое знание теоретического материала, умение обоснованно излагать свои мысли по обсуждаемым вопросам, способность полно, правильно и аргументированно отвечать на вопросы, приводить примеры
4 «хорошо»	демонстрирует знание теоретического материала, его последовательное изложение, способность приводить примеры, допускает единичные ошибки, исправляемые после замечания преподавателя
3 «удовлетворительно»	демонстрирует неполное, фрагментарное знание теоретического материала, требующее наводящих вопросов преподавателя, допускает существенные ошибки в его изложении, затрудняется в приведении примеров и формулировке выводов
2 «неудовлетворительно»	демонстрирует существенные пробелы в знании теоретического материала, не способен его изложить и ответить на наводящие вопросы преподавателя, не может привести примеры

Оценка выставляется по шкале от 0 до 50 баллов. Итоговая оценка по предмету вычисляется как сумма баллов, полученных за ответ на зачете и балл, полученный за отчеты по лабораторным работам. Результат рассчитывается в итоговый балл по шкале от 0 до 100 баллов.

Таблица 8 – Показатели оценивания результатов обучения в виде умений и владений

Шкала оценивания	Критерии оценивания
5 «отлично»	демонстрирует способность применять знание теоретического материала при выполнении заданий, последовательно и правильно выполняет задания, умеет обоснованно излагать свои мысли и делать необходимые выводы
4 «хорошо»	демонстрирует способность применять знание теоретического материала при выполнении заданий, последовательно и правильно выполняет задания, умеет обоснованно излагать свои мысли и делать необходимые выводы, допускает единичные ошибки, исправляемые после замечания преподавателя
3 «удовлетворительно»	демонстрирует отдельные, несистематизированные навыки, испытывает затруднения и допускает ошибки при выполнении заданий, выполняет задание по подсказке преподавателя, затрудняется в формулировке выводов
2 «неудовлетворительно»	не способен правильно выполнить задания

7.3. Контрольные задания и иные материалы, необходимые для оценки результатов обучения по дисциплине (модулю)

Лабораторная работа №1. Анализ и преобразование полуструктурированных данных (JSON, XML)

Цель: Получение практических навыков парсинга, анализа и взаимной конвертации данных,

представленных в форматах JSON и XML, в реляционное (табличное) представление.

Задачи:

1. Изучить библиотеки Python для работы с JSON (json) и XML (ElementTree).
2. Реализовать чтение и парсинг вложенных JSON и XML структур.
3. Разработать скрипт для извлечения заданных полей и их преобразования в плоскую CSV-таблицу.
4. Обработать возможные ошибки и пропуски данных при конвертации.

Краткое содержание: Студенту предоставляется набор JSON-файлов (например, данные о товарах с вложенными характеристиками) и XML-файл (например, заказы с позициями). Необходимо разработать скрипт, который извлекает указанные атрибуты (например, название товара, цена, артикул, имя клиента, дата заказа), объединяет их (если требуется) и сохраняет результат в виде одного или нескольких CSV-файлов, пригодных для загрузки в базу данных.

Лабораторная работа №2. Очистка и профилирование "сырых" данных с использованием Pandas

Цель: Освоение базовых техник очистки (Data Cleaning) и первичного анализа (Data Profiling) "сырых" наборов данных для повышения их качества.

Задачи:

1. Загрузить "грязный" набор данных (CSV или Excel) в Pandas DataFrame.
2. Провести профилирование: определить типы данных, найти и оценить количество пропусков (NaN), выявить дубликаты.
3. Обнаружить аномалии (выбросы) в числовых данных и некорректные значения в категориальных.
4. Выполнить очистку: применить стратегии заполнения пропусков (среднее, медиана, мода, константа), удалить дубликаты, скорректировать типы данных, обработать выбросы (удаление или замена).

Краткое содержание: Используя библиотеку Pandas и Jupyter Notebook, загрузить предоставленный "грязный" набор данных (например, логи посещений или данные о клиентах). Провести детальный анализ качества данных. Применить не менее 5 различных техник очистки к разным столбцам. Сохранить очищенный датасет и предоставить Jupyter Notebook как отчет с описанием найденных проблем и шагов по их устранению (с обоснованием выбора методов).

Лабораторная работа №3. Разметка данных для задач машинного обучения в Label Studio (или CVAT)

Цель: Получение навыков работы с современными инструментами аннотирования данных и подготовки датасетов для задач Computer Vision (CV) или Natural Language Processing (NLP).

Задачи:

1. Развернуть локально (или использовать облачную версию) платформу Label Studio/CVAT.
2. Создать новый проект и настроить конфигурацию разметки (Labeling Config) в соответствии с задачей.
3. Загрузить набор данных (изображения или тексты).
4. Выполнить ручную разметку части данных.
5. Экспортировать аннотации в одном из стандартных форматов (COCO, YOLO, JSON).

Краткое содержание: Студент выбирает одну из двух задач:

- **CV (Детекция объектов):** Загрузить 100 изображений (например, уличные сцены). Разметить все объекты класса "автомобиль" и "пешеход" с помощью Bounding Boxes.
- **NLP (NER):** Загрузить 50 текстовых фрагментов (новости). Разметить именованные сущности: "Организация", "Персона", "Локация".

Лабораторная работа №4. Обогащение набора данных с использованием внешних API

Цель: Изучение методов обогащения (Data Enrichment) путем интеграции данных из внешних веб-сервисов (API) для повышения прогностической ценности модели.

Задачи:

1. Выбрать и изучить документацию публичного API (например, API геокодирования, API погоды, API курсов валют, API баз знаний).
2. Получить ключ доступа (если требуется).
3. Написать скрипт на Python (используя requests), который по списку ключей (ID, адресов, дат) получает данные из API.
4. Объединить (join) полученные данные с исходным набором данных (в Pandas DataFrame).

Краткое содержание: Имеется CSV-файл со списком адресов (например, ресторанов). Необходимо написать программу, которая для каждой строки файла обращается к API геокодера (например, DaData или OpenStreetMap Nominatim), получает географические координаты (широту и долготу) и записывает их в новые столбцы исходного файла.

Лабораторная работа №5. Сбор и структурирование данных из веб-источников (Web Scraping)

Цель: Освоение техник парсинга статических и динамических веб-страниц для сбора и последующего структурирования данных.

Задачи:

1. Изучить библиотеки requests и BeautifulSoup (для статических сайтов) или Selenium (для динамических).
2. Проанализировать HTML/CSS структуру целевого веб-сайта (используя DevTools браузера).
3. Реализовать парсер для извлечения конкретных данных (например, название, цена, характеристики товара, текст статьи).
4. Обеспечить обработку пагинации (переход по страницам) и базовую обработку ошибок.
5. Сохранить собранные данные в структурированном виде (JSON или CSV).

Краткое содержание: Выбрать веб-сайт (например, интернет-магазин или сайт с объявлениями). Разработать скрейпер, который собирает информацию (например, название, цена, рейтинг) не менее чем с 5 страниц каталога (суммарно не менее 50 элементов). *Важно: соблюдать правила robots.txt и этические нормы скрапинга (устанавливать задержки, User-Agent).*

Лабораторная работа №6. Построение конвейера валидации качества данных (Data Validation Pipeline)

Цель: Изучение подходов к автоматизированной проверке качества данных (Data Validation) с использованием специализированных фреймворков.

Задачи:

1. Изучить фреймворк Great Expectations (или Pandera).
2. Инициализировать проект (Data Context) для существующего набора данных (можно взять из Лаб. 2).
3. Определить набор "ожиданий" (Expectations) к данным: проверка на null, уникальность ключей, попадание в диапазон, соответствие формату (e.g., email, дата) и т.д.
4. Создать "контрольную точку" (Checkpoint) и запустить валидацию.
5. Сгенерировать и проанализировать отчет о качестве данных (Data Docs).

Краткое содержание: Взять очищенный набор данных из Лабораторной работы №2. С помощью Great Expectations определить не менее 10 "ожиданий" (правил) для ключевых столбцов. Затем взять "грязный" набор данных (исходный) и прогнать его через этот же Checkpoint. Проанализировать сгенерированный HTML-отчет, чтобы увидеть, какие именно проверки не прошли "грязные" данные.

Перечень вопросов и заданий, выносимых на экзамен

1. Сравнение типов данных: В чем фундаментальное различие между структурированными, полуструктурированными и неструктурированными данными? Приведите по два примера для каждого типа.
2. Формат JSON: Опишите синтаксис, основные типы данных (объект, массив, примитивы) и преимущества использования формата JSON.
3. Формат XML: Сравните JSON и XML. В каких сценариях XML может быть предпочтительнее? (Назовите не менее 2-х).
4. Бинарные форматы: Почему для аналитики больших данных (Big Data) часто предпочитают форматы Parquet или Avro вместо CSV?
5. Схемы данных: Что такое схема данных (Data Schema) и зачем нужна ее валидация (например, с помощью JSON Schema)?
6. Нормализация и Денормализация: Дайте определение нормализации и денормализации. Когда денормализация является оправданной стратегией при структурировании данных?
7. Задачи разметки: Перечислите и кратко опишите 4 основные задачи, которые решаются с помощью разметки данных (например, для CV и NLP).
8. Качество разметки (IAA): Что такое "согласие между разметчиками" (Inter-Annotator Agreement, IAA)? Назовите и опишите одну из метрик (например, Каппа Коэна).
9. Типы разметки: Сравните подходы: ручная разметка, полуавтоматическая разметка и краудсорсинг.
10. "Человек-в-цикле" (HITL): Объясните концепцию "Human-in-the-Loop". Какую роль человек играет в этом цикле при разметке?
11. Активное обучение (Active Learning): Что такое "активное обучение" и как этот метод помогает снизить объем данных, требующих ручной разметки?

12. Определение обогащения: Что такое обогащение данных? Приведите пример внутреннего (на основе существующих данных) и внешнего обогащения.
13. Feature Engineering: Как "генерация признаков" (Feature Engineering) связана с обогащением данных?
14. Обогащение через API: Опишите общий алгоритм обогащения набора данных (например, списка адресов) с использованием внешнего API (например, API геокодирования).
15. Веб-скрапинг: Что такое веб-скрапинг (web scraping)? Назовите 2-3 технические и 1-2 этические проблемы, связанные с этим методом сбора данных.
16. Слияние данных (Data Fusion): Какие сложности возникают при интеграции (слиянии) данных из нескольких разнородных источников?
17. Очистка данных (Pandas): Назовите 5 наиболее частых операций по очистке данных, которые выполняются с помощью библиотеки Pandas (например, обработка пропусков, удаление дубликатов...).
18. Платформы разметки: Сравните назначение и основные возможности платформ Label Studio и CVAT.
19. ETL и ELT: Дайте определение ETL (Extract, Transform, Load). В чем ключевое отличие ETL от подхода ELT?
20. Версионирование данных (DVC): Зачем необходимо версионирование данных (Data Version Control, DVC) в проектах машинного обучения? Чем оно отличается от Git?
21. Конвейеры (Pipelines): Что такое конвейер обработки данных (data pipeline)? Назовите его основные этапы.
22. Метрики качества: Назовите и дайте краткое определение 5 основным метрикам качества данных (например, полнота, точность, актуальность...).
23. Профилирование данных: Что такое профилирование данных (Data Profiling) и какие задачи оно решает?
24. Валидация данных: Каково назначение инструментов автоматической валидации данных, таких как "Great Expectations"?
25. Управление данными (Data Governance): Дайте определение понятию "Управление данными" (Data Governance).
26. Анонимизация: В чем разница между анонимизацией и псевдонимизацией данных?
27. Управление выпуском: Как версионирование данных (DVC) и конвейеры (pipelines) помогают в управлении выпуском (релизом) моделей в IT-проекте?
28. Анализ обратной связи: Приведите пример, как замечание заказчика (например, "ваша система плохо распознает объекты ночью") транслируется в задачи по разметке и обогащению данных.
29. Управление изменениями: Опишите процесс внесения изменений в проект (например, добавление нового класса для разметки), требующий согласования с заказчиком.
30. Жизненный цикл данных: Опишите полный жизненный цикл данных в проекте ИИ, начиная от сбора "сырых" данных и заканчивая их архивацией.

Таблица 9 – Примеры оценочных средств с ключами правильных ответов

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
ПК-2. Способен разрабатывать алгоритмы и программные средства для решения задач в области создания и применения искусственного интеллекта.				
1.	Задание закрытого типа	Какой формат данных является полуструктурированным, человекочитаемым и в основном использует пары "ключ-значение" и массивы? A) CSV B) XML C) JSON D) Parquet	C) JSON	1

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
2.		Как называется метрика качества данных, означающая отсутствие противоречий между различными источниками данных или в рамках одного источника? A) Полнота B) Актуальность C) Непротиворечивость D) Точность	C) Непротиворечивость	1
3.		Что такое "согласие между разметчиками" (Inter-Annotator Agreement, IAA)? A) Метод автоматической разметки данных с помощью ИИ. B) Степень совпадения результатов разметки у нескольких специалистов (аннотаторов) при работе с одними и теми же данными. C) Программный интерфейс (API) для платформы разметки. D) Процесс обучения и аттестации нового разметчика.	B) Степень совпадения результатов разметки у нескольких специалистов (аннотаторов) при работе с одними и теми же данными.	1
4.	Задание открытого типа	Объясните, почему для аналитики больших данных (Big Data) бинарные форматы, такие как Parquet или Avro, часто предпочтительнее текстовых форматов, таких как CSV или JSON?	Предпочтение отдается бинарным форматам по нескольким ключевым причинам: 1. Компактность (Сжатие): Бинарные форматы хранят данные гораздо эффективнее, чем текстовые, и часто имеют встроенные алгоритмы сжатия (например, Snappy, Gzip), что значительно уменьшает объем хранимых данных и экономит место. 2. Скорость чтения/записи: Поскольку данные хранятся в оптимизированном бинарном виде, их сериализация и десериализация (чтение/запись) происходят намного быстрее, что критично для высокопроизводительных вычислений. 3. Колоночное хранение (для Parquet): Parquet хранит данные по колонкам, а не по строкам (как CSV). Это дает огромное преимущество в аналитических запросах (OLAP), когда нужно агрегировать данные только по нескольким столбцам из сотен (например, `SELECT SUM(price) ...`). Системе не нужно считывать весь файл, а только нужные колонки. 4. Хранение схемы: Бинарные форматы (особенно Avro) хранят схему данных (типы данных, названия полей) вместе с самими данными, что предотвращает ошибки при чтении и обеспечивает целостность данных.	5
5.		Вы внедряете процесс разметки изображений для детекции объектов (Bounding Boxes). Какие три основные проблемы или источника ошибок вы ожидаете от команды разметчиков и как вы будете их контролировать?	Основные проблемы и методы контроля: 1. Проблема (Субъективность/Неполнота): Разметчики могут пропускать объекты (особенно мелкие или частично перекрытые) или, наоборот, размечать "мусор". - Контроль: Четкая и детальная инструкция по разметке с примерами всех пограничных случаев (что считать объектом, что нет, как размечать перекрытия). 2. Проблема (Низкая точность границ): Bounding Box (рамка) может быть слишком "свободной"	5

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
			(захватывает много фона) или слишком "тесной" (обрезает часть объекта). - Контроль: Внедрение "золотого набора" (Gold Standard) — эталонно размеченных данных, на которых разметчики тренируются. Также используется аудит (выборочная проверка) работы старшим разметчиком/экспертом. 3. Проблема (Несогласованность): Два разных разметчика один и тот же случай размечают по-разному (например, один размечает "человека", а другой "человека" и "рюкзак" на нем отдельно). - Контроль: Использование перекрытия (Overlap) — когда одна и та же задача дается 2-3 разным разметчикам. Затем рассчитываются метрики согласия (IAA, Inter-Annotator Agreement), например, "Каппа Коэна". Если согласие низкое, инструкция дорабатывается, а разметчики проходят повторный инструктаж.	
6.		Сравните концепции "Активного обучения" (Active Learning) и "Человек-в-цикле" (Human-in-the-Loop, HITL) в контексте разметки данных. В чем их сходство и в чем ключевое различие?	- Сходство: Оба подхода являются стратегиями полуавтоматической разметки, которые используют комбинацию машинного обучения и ручного труда человека (эксперта) для создания набора данных. - Human-in-the-Loop (HITL): Это общая концепция, при которой модель и человек работают совместно. Модель делает предсказание (например, предлагает рамку), а человек (разметчик) либо подтверждает его, либо исправляет. Это ускоряет разметку по сравнению с полностью ручной "с нуля". - Active Learning (Активное обучение): Это конкретная стратегия внутри HITL. Ее цель — не ускорить разметку, а снизить ее объем. Модель, обученная на малом количестве данных, сама выбирает из всего неразмеченного пула те данные, на которых она больше всего "не уверена" (например, предсказания с низкой вероятностью). Она отдает на ручную разметку человеку только эти, самые сложные и информативные примеры, что позволяет достичь высокой точности модели при меньших затратах на разметку.	5
ПК-5. Способен руководить процессами разработки программного обеспечения.				
7.	Задание закрытого типа	Процесс добавления в набор данных географических координат (широты и долготы) на основе почтовых адресов является примером: A) Профилирования данных B) Очистки данных C) Обогащения данных D) Анонимизации данных	C) Обогащения данных	1
8.		Какой инструмент в первую очередь предназначен для версионирования наборов данных и моделей (Data Version Control) в проектах машинного обучения? A) Pandas B) Apache Airflow C) DVC D) Label Studio	C) DVC	1

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
9.		К какому типу данных относятся необработанные текстовые документы, фотографии и аудиофайлы? A) Структурированные B) Полуструктурированные C) Неструктурированные D) Реляционные	C) Неструктурированные	1
10.	Задание открытого типа	Опишите, как бы вы действовали для обогащения набора данных о клиентах (CSV-файл), содержащего ID клиента и его почтовый адрес. Назовите не менее двух типов данных, которыми можно обогатить этот набор, и укажите источники их получения.	Действия по обогащению: 1. Выбор API и источников: - Источник 1 (Геокодирование): Использовать API геокодера (например, DaData, OpenStreetMap Nominatim, Яндекс.Карты). - Источник 2 (Демография/Инфраструктура): Использовать публичные данные Росстата, данные 2ГИС или открытые наборы данных (например, OpenData). 2. Процесс (для Геокодирования): - Написать скрипт (например, на Python с библиотекой `requests`), который будет итерироваться по каждой строке CSV-файла. - Внутри цикла он будет брать значение из столбца "почтовый адрес" и отправлять его в виде HTTP-запроса к API геокодера. - Из ответа API (обычно в формате JSON) извлекать нужные поля. - Тип данных 1: Добавить геокоординаты (широта и долгота) в новые столбцы CSV-файла. 3. Процесс (для Демографии/Инфраструктуры): - Используя полученные на шаге 2 координаты или адрес (район, округ), обратиться к другому API или локальной базе данных. - Тип данных 2: Добавить социально-демографические или инфраструктурные признаки (например, "средний доход по району", "плотность населения", "количество конкурирующих магазинов в радиусе 1 км"). 4. Контроль: Обработать ошибки API (лимиты запросов, ненайденные адреса) и обеспечить сохранение результата.	5
11.		Что такое профилирование данных (Data Profiling) и какие три ключевые характеристики данных оно помогает выявить?	Профилирование данных — это процесс первичного анализа "сырого" набора данных для получения общего представления (статистического "портрета") о его содержимом и качестве. Это первый шаг перед очисткой данных. Ключевые характеристики, которые оно выявляет: 1. Структура и типы: Определяет фактические типы данных в столбцах (число, строка, дата) и сравнивает их с ожидаемыми. Помогает найти столбцы, где числа хранятся как строки ("1 000 руб") или даты в разных форматах. 2. Полнота (Пропуски): Рассчитывает количество и процент пропущенных значений (NaN, NULL) в каждом столбце, чтобы оценить объем "грязных" данных. 3. Распределение и Аномалии: - Для числовых данных: выявляет основные статистики (среднее, медиана, min, max) и	5

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
			помогает обнаружить выбросы (аномалии), например, возраст 150 лет. - Для категориальных данных: показывает распределение уникальных значений (кардинальность), выявляет наиболее и наименее частые значения, а также ошибки (например, "Москва", "Мск", "г. Москва" как три разных значения).	
12.		Объясните назначение и принцип работы инструментов автоматической валидации данных, таких как Great Expectations или Pandera.	- Назначение: Эти инструменты предназначены для автоматического тестирования и документирования качества данных. Их главная задача — предотвратить попадание "плохих" (некачественных, некорректных) данных в следующие этапы конвейера обработки (pipeline), например, в обучение модели или в аналитическое хранилище. - Принцип работы: 1. Определение "Ожиданий" (Expectations): Разработчик (аналитик) один раз описывает набор правил (утверждений), которым _должны_ соответствовать данные. Это и есть "ожидания". Например: - `expect_column_values_to_not_be_null('user_id')` (столбец ID не должен иметь пропусков). - `expect_column_values_to_be_unique('user_id')` (ID должны быть уникальны). - `expect_column_values_to_be_between('age', 18, 99)` (возраст в допустимом диапазоне). - `expect_column_values_to_be_in_set('status', ['new', 'active', 'closed'])` (статус только из списка). 2. Валидация (Validation): Инструмент запускает этот набор "ожиданий" на новом батче (порции) данных. 3. Отчет (Data Docs): По результатам проверки генерируется детальный отчет (часто в виде HTML-страницы), который показывает, какие именно проверки (ожидания) прошли, а какие провалились, и на каких данных. 4. Действие: Если валидация не пройдена, конвейер (pipeline) может быть автоматически остановлен, а ответственным отправляется уведомление.	5

Полный комплект оценочных материалов по дисциплине (модулю) (фонд оценочных средств) хранится в электронном виде на кафедре, утверждающей рабочую программу дисциплины (модуля), и в Центре мониторинга и аудита качества обучения.

7.4. Методические материалы, определяющие процедуры оценивания результатов обучения по дисциплине (модулю)

Таблица 10 – Технологическая карта рейтинговых баллов по дисциплине (модулю)

№ п/п	Контролируемые мероприятия	Количество мероприятий / баллы	Максимальное количество баллов	Срок представления
Основной блок				
1.	Ответ на занятии	6	6	По расписанию
2.	Выполнение лабораторной работы	7	84	По расписанию
Всего			**90**	-
Блок бонусов				
1.	Посещение занятий	1	1	По расписанию
2.	Своевременное выполнение всех заданий	9	9	По расписанию
Всего			**10**	-
ИТОГО			**100**	-

Таблица 11 – Система штрафов (для одного занятия)

Показатель	Балл
Опоздание (два и более)	-2
Не готов к практической части занятия	-3
Нарушение учебной дисциплины	-2
Пропуски лекций без уважительных причин (за одну лекцию)	-2
Пропуск занятий без уважительной причины (за одно занятие)	-2
Нарушение правил техники безопасности	-1
Отсутствие конспектов лекций, семинарских занятий, первоисточников при начислении баллов не учитываются	0

Таблица 12 – Шкала перевода рейтинговых баллов в итоговую оценку за семестр по дисциплине (модулю)

Сумма баллов	Оценка по 4-балльной шкале
90–100	5 (отлично)
85–89	4 (хорошо)
75–84	
70–74	
65–69	3 (удовлетворительно)
60–64	
Ниже 60	2 (неудовлетворительно)

При реализации дисциплины (модуля) в зависимости от уровня подготовленности обучающихся могут быть использованы иные формы, методы контроля и оценочные средства, исходя из конкретной ситуации.

8. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

8.1. Основная литература:

1. DAMA International. DAMA-DMBOK: Свод знаний по управлению данными. Второе издание / Пер. с англ. — Москва : Олимп-Бизнес, 2020. — 806 с. — ISBN 978-5-9693-0402-6.
2. Макгрегор, С. Обработка данных на Python. Data Wrangling и Data Quality / С. Макгрегор ; пер. с англ. А. В. Снатенкова. — Санкт-Петербург : Питер, 2023. — 256 с. — (Серия «Библиотека программиста»). — ISBN 978-5-4461-1934-0.
3. Чжен, Э. Машинное обучение. Конструирование признаков. Принципы и техники для аналитиков / Э. Чжен, А. Казари ; пер. с англ. М. А. Райтмана. — Москва : Эксмо, 2022. — 352 с. — (Top Engineering). — ISBN 978-5-04-121430-8.
4. Хейдт, М. Изучаем Pandas / М. Хейдт ; пер. с англ. — 2-е изд. — Москва : ДМК Пресс, 2022. — 510 с. — ISBN 978-5-97060-943-7.
5. Рейс, Д. Основы инженерии данных / Д. Рейс, М. Хаусли ; пер. с англ. — Санкт-Петербург : Питер, 2024. — 368 с. — (Серия «Бестселлеры O'Reilly»). — ISBN 978-5-4461-2016-2.
6. Большакова, Е. И. Автоматическая обработка текстов на естественном языке и анализ данных : учебное пособие / Е. И. Большакова, К. В. Воронцов, Н. Э. Ефремова [и др.]. — Москва : Изд. дом Высшей школы экономики, 2017. — 269 с. — ISBN 978-5-7598-1580-0.
7. Садаладж, П. NoSQL. Методология разработки нереляционных баз данных / П. Садаладж, М. Фаулер ; пер. с англ. — 2-е изд. — Санкт-Петербург : Питер, 2020. — 192 с. — (Серия «Бестселлеры O'Reilly»). — ISBN 978-5-4461-1406-2.

8.2. Дополнительная литература:

8. Берсон, А. Управление мастер-данными. Лучшие практики / А. Берсон, Л. Дубов ; пер. с англ. — Москва : ДМК Пресс, 2019. — 560 с. — ISBN 978-5-97060-681-8.
9. Бурков, А. Машинное обучение: проверенный подход / А. Бурков ; пер. с англ. — Санкт-Петербург : Питер, 2021. — 288 с. — (Серия «Бестселлеры O'Reilly»). — ISBN 978-5-4461-1634-9.
10. Грофф, Д. Р. SQL: полное руководство / Д. Р. Грофф, П. Н. Вайнберг, Э. Д. Оппель ; пер. с англ. — 3-е изд. — Москва : Вильямс, 2014. — 960 с. — ISBN 978-5-8459-1654-9.
11. Николенко, С. Глубокое обучение / С. Николенко, А. Кадурин, Е. Архангельская. — Санкт-Петербург : Питер, 2020. — 480 с. — (Серия «Библиотека программиста»). — ISBN 978-5-496-02534-7.
12. Хюйен, Ч. Проектирование систем машинного обучения. Паттерны и лучшие практики / Ч. Хюйен ; пер. с англ. — Санкт-Петербург : Питер, 2023. — 608 с. — (Серия «Бестселлеры O'Reilly»). — ISBN 978-5-4461-1913-5.
13. Редмонд, Э. Семь баз данных за семь недель. Введение в современные базы данных и идеологию NoSQL / Э. Редмонд, Д. Р. Уилсон ; пер. с англ. — Москва : ДМК Пресс, 2018. — 384 с. — ISBN 978-5-97060-641-2.
14. Солем, Я. Э. Программирование компьютерного зрения на Python / Я. Э. Солем ; пер. с англ. — Москва : ДМК Пресс, 2016. — 312 с. — ISBN 978-5-97060-312-1.
15. Джуба, С. Изучаем PostgreSQL 10 / С. Джуба ; пер. с англ. — Москва : ДМК Пресс, 2018. — 364 с. — ISBN 978-5-97060-621-4.
16. Болье, А. Изучаем SQL / А. Болье ; пер. с англ. — 3-е изд. — Санкт-Петербург : Питер, 2024. — 368 с. — (Серия «Бестселлеры O'Reilly»). — ISBN 978-5-4461-2015-5.
17. Брюс, П. Практическая статистика для специалистов Data Science / П. Брюс, Э. Брюс, П. Гедек ; пер. с англ. — 2-е изд. — Санкт-Петербург : БХВ-Петербург, 2021. — 352 с. — ISBN 978-5-9775-6675-7.
18. Лотман, Ю. М. Структура художественного текста / Ю. М. Лотман. — Москва : Азбука, 2019. — 416 с. — ISBN 978-5-389-15822-2.
19. Клеппман, М. Высоконагруженные приложения. Проектирование масштабируемых и отказоустойчивых систем / М. Клеппман ; пер. с англ. — Санкт-Петербург : Питер, 2022. — 720 с. — (Серия «Бестселлеры O'Reilly»). — ISBN 978-5-4461-1711-7.

20. Аггарвал, Ч. Нейронные сети и глубокое обучение. Учебный курс / Ч. Аггарвал ; пер. с англ. — Санкт-Петербург : Диалектика, 2021. — 880 с. — ISBN 978-5-907365-24-3.

8.3. Интернет-ресурсы, необходимые для освоения дисциплины (модуля)

Электронно-библиотечная система (ЭБС) ООО «Политехресурс» «Консультант студента», http://www.studentlibrary.ru.

9. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Учебные аудитории, библиотеки АГУ, центр мониторинга и аудита качества образования, компьютерные классы, мультимедийные аудитории.

10. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Рабочая программа дисциплины (модуля) при необходимости может быть адаптирована для обучения (в том числе с применением дистанционных образовательных технологий) лиц с ограниченными возможностями здоровья, инвалидов. Для этого требуется заявление обучающихся, являющихся лицами с ограниченными возможностями здоровья, инвалидами, или их законных представителей и рекомендации психолого-медико-педагогической комиссии. При обучении лиц с ограниченными возможностями здоровья учитываются их индивидуальные психофизические особенности. Обучение инвалидов осуществляется также в соответствии с индивидуальной программой реабилитации инвалида (при наличии).

Для лиц с нарушением слуха возможно предоставление учебной информации в визуальной форме (краткий конспект лекций; тексты заданий, напечатанные увеличенным шрифтом), на аудиторных занятиях допускается присутствие ассистента, а также сурдопереводчиков и тифлосурдопереводчиков. Текущий контроль успеваемости осуществляется в письменной форме: обучающийся письменно отвечает на вопросы, письменно выполняет практические задания. Доклад (реферат) также может быть представлен в письменной форме, при этом требования к содержанию остаются теми же, а требования к качеству изложения материала (понятность, качество речи, взаимодействие с аудиторией и т. д.) заменяются на соответствующие требования, предъявляемые к письменным работам (качество оформления текста и списка литературы, грамотность, наличие иллюстрационных материалов и т. д.). Промежуточная аттестация для лиц с нарушениями слуха проводится в письменной форме, при этом используются общие критерии оценивания. При необходимости время подготовки к ответу может быть увеличено.

Для лиц с нарушением зрения допускается аудиальное предоставление информации, а также использование на аудиторных занятиях звукозаписывающих устройств (диктофонов и т. д.). Допускается присутствие на занятиях ассистента (помощника), оказывающего обучающимся необходимую техническую помощь. Текущий контроль успеваемости осуществляется в устной форме. При проведении промежуточной аттестации для лиц с нарушением зрения тестирование может быть заменено на устное собеседование по вопросам.

Для лиц с ограниченными возможностями здоровья, имеющих нарушения опорно-двигательного аппарата, на аудиторных занятиях, а также при проведении процедур текущего контроля успеваемости и промежуточной аттестации могут быть предоставлены необходимые технические средства (персональный компьютер, ноутбук или другой гаджет); допускается присутствие ассистента (ассистентов), оказывающего обучающимся необходимую техническую помощь (занять рабочее место, передвигаться по аудитории, прочитать задание, оформить ответ, общаться с преподавателем).