Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»

УТВЕРЖДАЮ

Директор Института магистратуры

_____Иванова Е.А.

«22» февраля 2022 г.

Рабочая программа дисциплины

Технологии анализа больших данных

Направление 09.04.03 Прикладная информатика

магистерская программа

09.04.03.03 Машинное обучение и технологии больших данных

Для набора 2022 года

Квалификация

магистр

Кафедра Информационных систем и прикладной информатики

Составители рабочей программы:

к.э.н., доцент Аручиди Наталья Александровна

СОДЕРЖАНИЕ

I. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Цели освоения дисциплины:

– формирование у обучающихся способности осуществлять обработку больших объемов данных для решения профессиональных задач, эффективно применять методы, технологии и инструментальные средства анализа больших данных в профессиональной деятельности.

Задачи освоения дисциплины:

– развитие у обучающихся умения выбирать и применять методы, технологии и инструментальные средства для решения задач анализа больших данных;
– формирование знаний, умений и навыков владения технологиями хранения, обработки и анализа больших данных, методами построения информационных систем на основе нереляционных баз данных и распределенных систем хранения

II. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Дисциплина относится к модулю обязательных профессиональных дисциплин обязательной части образовательной программы.

Для изучения данной дисциплины необходимы знания, умения и навыки, формируемые предшествующими элементами образовательной программы:

Наименование дисциплины (модуля), практики	Требуемые знания, умения, навыки
Математические методы анализа больших данных	*Знания:* – Знает классы методов и алгоритмов машинного обучения. – Знает математические модели, методы и алгоритмы для обработки и анализа больших данных. – Знает принципы построения моделей глубоких нейронных сетей и глубокого машинного обучения. – Знает подходы к применению моделей на основе нечеткой логики в системах искусственного интеллекта. *Умения:* – Умеет ставить задачи и адаптировать методы и алгоритмы машинного обучения. – Умеет выбирать и применять математические модели, методы и алгоритмы для решения прикладных задач анализа больших данных. – Умеет руководить выполнением коллективной проектной деятельности для создания, поддержки и использования систем искусственного интеллекта на основе моделей глубоких нейронных сетей и нечетких моделей и методов.
Экспертные системы и базы знаний	*Знания:* – Знает принципы построения моделей глубоких нейронных сетей и глубокого машинного обучения. – Знает основные постулаты искусственного интеллекта, модели представления данных и знаний. – Знает подходы к применению моделей на основе нечеткой логики в системах искусственного интеллекта. – Знает фундаментальные правила построения рекомендательных систем и систем поддержки принятия решений, основанных на интеллектуальных принципах, методы и подходы к планированию и реализации проектов по созданию систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Рекомендательные системы и системы

Наименование дисциплины (модуля), практики	Требуемые знания, умения, навыки
	поддержки принятия решений». – Знает методологические подходы к выбору и применению методов обработки и распространения знаний с помощью с помощью дедукции, индукции и абдукции, согласования экспертных оценок и нечеткого вывода. *Умения:* – Умеет строить «мягкие» модели, используя методы правдоподобного вывода. – Умеет представлять знания в виде продукционных систем, семантических сетей и фреймов. – Умеет руководить выполнением коллективной проектной деятельности для создания, поддержки и использования систем искусственного интеллекта на основе моделей глубоких нейронных сетей и нечетких моделей и методов. – Умеет решать задачи по выполнению коллективной проектной деятельности для создания, поддержки и использования систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Рекомендательные системы и системы поддержки принятия решений» со стороны заказчика.
Методы машинного обучения	*Знания:* – Знает классы методов и алгоритмов машинного обучения. – Знает возможности современных инструментальных средств и систем программирования для решения задач машинного обучения. – Знает функциональность современных инструментальных средств и систем программирования в области создания моделей и методов машинного обучения. *Умения:* – Умеет ставить задачи и адаптировать методы и алгоритмы машинного обучения. – Умеет проводить сравнительный анализ и осуществлять выбор инструментальных средств для решения задач машинного обучения. – Умеет применять современные инструментальные средства и системы программирования для разработки новых методов и моделей машинного обучения.

Знания, умения и навыки, формируемые данной дисциплиной, потребуются при освоении следующих элементов образовательной программы:

– производственная практика, проектно-технологическая практика;
– производственная практика, преддипломная практика.

III. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Освоение дисциплины направлено на формирование следующих компетенций в соответствии с образовательной программой:

Перечень планируемых результатов обучения по дисциплине, соотнесённых с индикаторами достижения компетенций

Компетенция	Индикаторы достижения компетенции	Результаты обучения
ПК-4. Способен руководить проектами по созданию комплексных систем на основе аналитики больших данных в различных отраслях со стороны заказчика	ПК-4.1. Руководит проектами по построению комплексных систем на основе аналитики больших данных в различных отраслях со стороны заказчика	*Знания:* – Знает методологию и принципы руководства проектами по созданию, поддержке и использованию комплексных систем на основе аналитики больших данных со стороны заказчика. – Знает специфику сфер и отраслей, для которых реализуется проект по аналитике больших данных *Умения:* – Умеет решать задачи по руководству коллективной проектной деятельностью для создания, поддержки и использования комплексных систем на основе аналитики больших данных со стороны заказчика. – Умеет выявлять небольшие по масштабу проекты аналитики, которые потенциально могут представлять интерес для ряда подразделений / служб или для организации в целом. – Умеет выявлять области деловой деятельности, которые потенциально могут получить отдачу от аналитики.
ПК-6. Способен управлять этапами жизненного цикла методологической и технологической инфраструктуры анализа больших данных в организации	ПК-6.1. Управляет получением, хранением, передачей, обработкой больших данных	*Знания:* – Знает архитектуры и модели баз и хранилищ данных, адаптированные к технологиям больших данных. – Знает технологии, методы и инструментальные средства обработки больших данных. – Знает рекомендации по использованию, опыт использования и интеграции современных инструментальных средств сбора, хранения, обработки и анализа больших данных. – Знает рекомендации по использованию и опыт использования разнородных источников данных и информации в задачах анализа больших данных. – Знает производителей программного обеспечения и инфраструктуры технологий больших данных. *Умения*:

		– Умеет проводить интеграцию систем хранения и обработки данных. – Умеет пользоваться методами и инструментами получения, хранения, передачи, обработки больших данных. – Умеет выбирать NoSQL СУБД для решения прикладных задач. – Умеет проектировать архитектуры информационных систем на основе нереляционных баз данных и распределенных систем хранения.
	ПК-6.2. Управляет качеством больших данных	*Знания:* – Знает источники больших данных. – Знает базовые характеристики и метрики качества больших данных. – Знает методы и технологии управления качеством больших данных. – Знает методы оценки рисков использования больших данных. *Умения*: – Умеет определять происхождение данных и оценивать источники больших данных. – Умеет измерять и оценивать качество больших данных. – Умеет проводить оценку и устранение рисков использования больших данных.

IV. СОДЕРЖАНИЕ И СТРУКТУРА ДИСЦИПЛИНЫ

Трудоёмкость дисциплины составляет 5 зачётных единиц, 180 часов.

Форма промежуточной аттестации: дифференцированный зачёт

4.1. Содержание дисциплины, структурированное по темам

№ п/п	Темы дисциплины	Семестр	Виды учебной работы и их трудоёмкость, часы (в том числе с использованием онлайн-курсов)				Наименования оценочных средств
			Контактная работа			Самостоя-тельная работа	
			Лекции	Практические занятия	Лабораторные занятия		
Модуль 1. Основы построения и использования систем больших данных							
1	Раздел 1. «Основы систем больших данных». Понятие больших данных. Особенности сбора, хранения, обработки и анализа больших массивов данных. Источники больших данных. Базовые характеристики и метрики качества больших данных. Измерение и оценка качества больших данных. Методы и технологии управления качеством больших данных. Методы оценки рисков использования больших данных. Инженер по качеству данных, дата стюард. Data Steward. Использование больших данных в науке, бизнесе, государственном управлении.	3	4	-	-	22	Тест № 1
2	Раздел 2. «Методы работы с распределенными информационными системами. Управление качеством данных». Использование фреймворка MapReduce в распределенной среде. Реализации MapReduce. Состав и возможности программного комплекса Apache Hadoop. Языки поисковых запросов для Hadoop. Управление качеством данных с использованием ETL-конвейеров Apache Airflow.	3	4	8	-	42	Практическая работа №1 Тест № 1
Модуль 2. Разработка и использование приложений на основе распределенных баз данных							
4	Раздел 4. «Базы данных NoSQL». Варианты построения распределенных баз данных, репликация, фрагментация. Согласованность. CAP-теорема. Классы NoSQL баз данных. Примеры СУБД NoSQL. Семейства столбцов. Графовые СУБД. Neo4j	3	4	4	-	42	Практическая работа №2 Тест № 2

№ п/п	Темы дисциплины	Семестр	Виды учебной работы и их трудоёмкость, часы (в том числе с использованием онлайн-курсов)				Наименования оценочных средств
			Контактная работа			Самостоя-тельная работа	
			Лекции	Практические занятия	Лабораторные занятия		
5	Раздел 5. «Документно-ориентированные распределенные СУБД». Понятие агрегата. Современные документо-ориентированные СУБД. Запросы к СУБД на языке JSON. MongoDB. Использование фреймворка MapReduce в документо-ориентированных СУБД.	3	4	4	-	42	Практическая работа №3 Тест № 2
Итого часов			**16**	**16**	**-**	**148**	–

4.2. План внеаудиторной самостоятельной работы

№ п/п	Темы дисциплины	Семестр	Вид самостоятельной работы	Сроки выполнения (нед.)	Затраты времени (часы)	Учебно-методическое обеспечение
Модуль 1. Основы построения и использования систем больших данных						
1	Раздел 1. «Основы систем больших данных». Понятие больших данных. Особенности сбора, хранения, обработки и анализа больших массивов данных. Источники больших данных. Использование больших данных в науке, бизнесе, государственном управлении.	3	– проработка и повторение материала лекционных занятий;	1–2	22	[1], [3], [4]
2	Раздел 2. «Методы работы с распределенными информационными системами. Управление качеством данных». Использование фреймворка MapReduce в распределенной среде. Реализации MapReduce. Состав и возможности программного комплекса Apache Hadoop. Языки поисковых запросов для Hadoop. Управление качеством данных с использованием ETL-конвейеров Apache Airflow.	3	– проработка и повторение материала лекционных занятий; – подготовка к практическим занятиям	3-10	42	[1], [4]
Модуль 2. Разработка и использование приложений на основе распределенных баз данных						

№ п/п	Темы дисциплины	Семестр	Вид самостоятельной работы	Сроки выполнения (нед.)	Затраты времени (часы)	Учебно-методическое обеспечение
4	Раздел 4. «СУБД NoSQL». Варианты построения распределенных баз данных, репликация, фрагментация. Согласованность. CAP-теорема. Классы NoSQL баз данных. Примеры СУБД NoSQL. Семейства столбцов. Графовые СУБД. Neo4j	3	– проработка и повторение материала лекционных занятий; – подготовка к практическим занятиям	11-14	42	[2]
5	Раздел 5. «Документно-ориентированные распределенные СУБД». Понятие агрегата. Современные документо-ориентированные СУБД. Запросы к СУБД на языке JSON. MongoDB. Использование фреймворка MapReduce в документо-ориентированных СУБД	3	– проработка и повторение материала лекционных занятий; – подготовка к практическим занятиям	15-18	42	[2]
Общая трудоёмкость самостоятельной работы по дисциплине					**148**	**–**

4.3. Содержание учебного материала

Модуль 1 «Основы построения и использования систем больших данных»
Тема 1.1 «Основы систем больших данных»
Понятие Больших данных. Особенности сбора, хранения, обработки и анализа больших массивов данных. Источники больших данных. Использование больших данных в науке, бизнесе, государственном управлении.
Тема 1.2 «Методы работы с распределенными информационными системами. Управление качеством данных»
Использование фреймворка MapReduce в распределенной среде. Реализации MapReduce. Состав и возможности программного комплекса Apache Hadoop. Языки поисковых запросов для Hadoop. Управление качеством данных с использованием ETL-конвейеров Apache Airflow.

Модуль 2 «Разработка и использование приложений на основе распределенных баз данных»
Тема 2.1 «СУБД NoSQL»
Варианты построения распределенных баз данных, репликация, фрагментация. Согласованность. CAP-теорема. Классы NoSQL баз данных. Примеры СУБД NoSQL. Семейства столбцов. Графовые СУБД. Neo4j
Тема 2.2 «Документно-ориентированные распределенные СУБД»
Понятие агрегата. Современные документо-ориентированные СУБД. Запросы к СУБД на языке JSON. MongoDB. Использование фреймворка Map-Reduce в документо-ориентированных СУБД.

Перечень тем практических занятий

№ п/п	Тема практического занятия	Количество часов
Модуль 1. Основы построения и использования систем больших данных		
1	Методы работы с распределенными информационными системами. Apache Spark, Hadoop, Mahout, MapReduce. Управление качеством данных. Apache Airflow.	8
Модуль 2. Разработка и использование приложений на основе распределенных баз данных		
2	СУБД NoSQL. Графовые СУБД. Neo4j.	4
3	Документно-ориентированные распределенные СУБД. MongoDB.	4
Всего часов		**16**

V. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

Активные формы обучения, применяемые на практических занятиях, способствуют разнообразному (индивидуальному, групповому, коллективному) изучению (усвоению) учебных вопросов (проблем), активному взаимодействию обучающихся и преподавателя, живому обмену мнениями между ними, нацеленному на выработку правильного понимания содержания изучаемой темы и способов ее практического использования.

Наряду с традиционными образовательными технологиями, для реализации дисциплины будут использоваться технологии электронного обучения и дистанционные образовательные технологий в электронной информационно-образовательной среде университета.

Аудиторные занятия и другие формы контактной работы обучающихся с преподавателем могут проводиться с использованием платформ Microsoft Teams, Moodle (BigBlueButton) и др., что позволяет обеспечить онлайн и офлайн взаимодействие преподавателя с обучающимися в рамках дисциплины

Основными методами текущего контроля являются электронный учёт и контроль учебных достижений студентов (использование средств сервиса балльно-рейтинговой системы; ведение

электронного журнала успеваемости, проведение электронного тестирования и применение других средств контроля с использованием системы электронного обучения).

VI. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Основная литература

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

2. Григорьев, Ю. А. Реляционные базы данных и системы NoSQL : учебное пособие / Ю. А. Григорьев, А. Д. Плутенко, О. Ю. Плужникова. – Благовещенск : Амурский государственный университет, 2018. – 425 с. – ISBN 978-5-93493-308-2. – Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. – URL: https://www.iprbookshop.ru/103912.html

3. Алексеев Д. С. Технологии интеллектуального анализа данных [Электронный ресурс]: учебное пособие / Алексеев Д. С. – Кострома: КГУ им. Н.А. Некрасова, 2020. - 141 с. – URL: https://e.lanbook.com/book/160082.

4. Воронов, В. И. Data Mining - технологии обработки больших данных : учебное пособие / В. И. Воронов, Л. И. Воронова, В. А. Усачев. – Москва : Московский технический университет связи и информатики, 2018. – 47 с. – ISBN 2227-8397. – Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. – URL: https://www.iprbookshop.ru/81324.html

6.2. Дополнительная литература

5. Билл, Фрэнкс Укрощение больших данных : как извлекать знания из массивов информации с помощью глубокой аналитики / Фрэнкс Билл ; перевод А. Баранов. – Москва : Манн, Иванов и Фербер, 2014. – 340 с. – ISBN 978-5-00057-146-0. – Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. – URL: https://www.iprbookshop.ru/39433.html

6. Е.И. Николаев. Базы данных в высокопроизводительных информационных системах : учебное пособие / авт.-сост. Е. И. Николаев ; Северо-Кавказский федеральный университет. – Ставрополь : Северо-Кавказский Федеральный университет (СКФУ), 2016. – 163 с. – URL: https://biblioclub.ru/index.php?page=book&id=466799

6.3. Периодические издания

- IEEE Spectrum https://spectrum.ieee.org/
- Научный журнал «Машинное обучение и анализ данных» http://jmlda.org/ru/journal
- Journal of Big Data https://journalofbigdata.springeropen.com/

6.4. Перечень ресурсов сети Интернет

- ЭБС IPR Books http://www.iprbookshop.ru/
- ЭБС «Университетская библиотека онлайн» http://biblioclub.ru.
- Образовательная платформа Юрайт https://urait.ru/
- IBM Academic Initiative http://ictis.sfedu.ru/ibm_academic_initiative/ (учебные материалы)
- http://github.com/
- http://habr.com/
- http://www.kdnuggets.com/
- Python, Свободное ПО, https://www.python.org/
- https://www.jetbrains.com/pycharm/

VII. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

При реализации дисциплины используются следующие помещения, оборудование и

программное обеспечение:

Лаборатория машинного обучения и технологий больших данных

Персональные компьютеры (8 шт.), проектор, экран. Windows 10, Microsoft Office 365, Adobe Acrobat Reader (Бесплатное проприетарное ПО, https://acrobat.adobe.com/ru/ru/acrobat/pdf-reader/volume-distribution.html), Google Chrome (Свободное ПО, https://google.com/chrome/browser/), Mozilla Firefox, Бесплатное ПО (GNU GPL), https://firefox.com/, Foxit (Бесплатное проприетарное ПО, https://www.foxitsoftware.com/ru/), i2 Analyst's Notebook (Бесплатная лицензия для образовательных целей), https://developer.ibm.com/academic/), Notepad++, Бесплатное ПО (GNU GPL 2), https://notepad-plus-plus.org/, Total Commander 7.x , WinRAR, XAMPP, Бесплатное ПО (GNU GPL), http://www.apachefriends.org/en/xampp.html, Team Foundation Server 2015, Visual Studio 2015, Android Studio, Операционная система на базе Linux; Офисный пакет Open Office, актуальные версии браузеров Google Chrome (Свободное ПО, https://google.com/chrome/browser/), Mozilla Firefox, Бесплатное ПО (GNU GPL), https://firefox.com/, Edge, Safari с поддержкой протокола WebRTC, PyCharm 2017.1.2 https://www.jetbrains.com/pycharm/ Свободное ПО, https://www.python.org/, Evolus Pencil, Свободное ПО (GNU GPL 2), https://pencil.evolus.vn/

VIII. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Дисциплина включает в себя лекционные и практические занятия, а также самостоятельную работу обучающихся.

Организация образовательного процесса по дисциплине осуществляется с использованием системы электронного обучения.

Все лекционные занятия проводятся с визуализацией учебного материала в форме презентаций лекционного материала, которые доступны в системе электронного обучения.

Лекционная часть курса включает следующие компоненты системы знаний учебной дисциплины: понятийный аппарат (тезаурус курса), теоретические утверждения, разъяснения и комментарии; междисциплинарные точки зрения; описание рассматриваемых разделов; ретроспективный и перспективный взгляды на изучаемую проблематику.

Практические занятия по всем модулям дисциплины требуют предварительной теоретической подготовки по соответствующим темам: проработка лекционного материала, ознакомление и изучение отдельных источников основной и дополнительной литературы.

Лекционные и практические занятия могут проводиться с применением дистанционных образовательных технологий с использованием платформ Microsoft Teams, Cisco, Moodle (BigBlueButton) и др.

Проведение лекционных и практических занятий осуществляется с постановкой проблемных вопросов, допускающих возникновение дискуссий, что предполагает активное включение студентов в образовательный процесс.

В организации процесса обучения используются как традиционные, характерные лекционно-семинарской форме обучения, так и инновационные (интерактивные, имитационные, проектные) технологии.

Используемые технологии обеспечивают:

– формирование компетенций, осознанное усвоение знаний, качественное освоение умений их применять и формирование заинтересованного отношения к изучаемым объектам в единстве;

– продуктивность познавательной деятельности, научный поиск, создание субъективно и объективно новых знаний или других продуктов;

– ориентацию на студентов, стимулирование их активности, самостоятельности, инициативы и ответственности;

– контекстный характер обучения, то есть привязку к реальным профессиональным задачам;

– вовлеченность студентов в выполняемую деятельность, возможность проявить и развить свой интеллектуальный, творческий, личностный, деловой потенциал.

Самостоятельная работа направлена на повышение качества обучения, углубление и закрепление знаний студента, развитие аналитических навыков по проблематике учебной дисциплины, активизацию учебно-познавательной деятельности студентов и снижение аудиторной нагрузки.

Максимальное количество баллов по каждому виду контрольных мероприятий указано в учебной карте дисциплины.

IX. УЧЕБНАЯ КАРТА ДИСЦИПЛИНЫ

Курс 2, семестр 3, очная форма обучения

№ п/п	Виды контрольных мероприятий (наименования оценочных средств)	Количество баллов	
		Текущий контроль	Рубежный контроль
Модуль 1. Основы построения и использования систем больших данных			
1	Практическая работа № 1	20	–
2	Тест № 1		20
Модуль 2. Разработка и использование приложений на основе распределенных баз данных			
1	Практическая работа № 2	20	–
2	Практическая работа № 3	20	–
3	Тест № 2		20
Всего		**60**	**40**
Бонусные баллы		**Не предусмотрены**	
Промежуточная аттестация в форме дифференцированного зачёта		Оценка по дисциплине выставляется по сумме баллов за текущий контроль и рубежный контроль: – 85–100 баллов – оценка «отлично»; – 71–84 балла – оценка «хорошо»; – 60–70 баллов – оценка «удовлетворительно»; – менее 60 баллов – оценка «неудовлетворительно»	

X. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

10.1. Паспорт фонда оценочных средств

№ п/п	Индикатор достижения компетенции	Наименование оценочного средства
1	ПК-4.1. Руководит проектами по построению комплексных систем на основе аналитики больших данных в различных отраслях со стороны заказчика	– практическая работа №1; – тест №1, тест №2
2	ПК-6.1. Управляет получением, хранением, передачей, обработкой больших данных	– практическая работа №1; – практическая работа №2; – практическая работа №3; – тест №1, тест №2
3	ПК-6.2. Управляет качеством больших данных	– практическая работа №1; – тест №1, тест №2

10.2. Практическая работа №1

Задание 1.1.: Настройка облачной среды Microsoft Azure и рабочей области машинного обучения

1. Зарегистрироваться в программе Microsoft Dreamspark. Получить код подтверждения Microsoft Azure for Students (Добавить в корзину, Оформить заказ). Следовать инструкции по регистрации в Microsoft Azure. Пройти верификацию по номеру телефона.

2. Авторизоваться в Студии машинного обучения https://studio.azureml.net/

3. Выполнить обучающие тьюториалы:

3.1. Создание первого эксперимента (на примере набора Данные о ценах на автомобили) https://docs.microsoft.com/ru-ru/azure/machine-learning/studio/create-experiment

3.2. Пошаговое руководство по разработке решения для прогнозной аналитики в службе машинного обучения Azure для оценки кредитных рисков.

https://docs.microsoft.com/ru-ru/azure/machine-learning/studio/walkthrough-develop-predictive-solution

Действия с 1 по 4.

1. Создание рабочей области машинного обучения
2. Отправка существующих данных
3. Создание нового эксперимента
4. Обучение и анализ моделей

Включить в отчет схемы экспериментов, подтверждающие выполнение тьюториалов.

4. Создать эксперимент и обучить модель на примере набора данных (по вариантам). Необходимо сравнить между собой (Evaluate Model) не менее трех алгоритмов.

Содержание отчета:

- Описание набора данных.
- Схема эксперимента.
- Оценка результатов эксперимента.

Задание 1.2. Разворачивание кластера в HDInsight

Создание кластеров под управлением Linux в HDInsight с помощью портала Azure
https://docs.microsoft.com/ru-ru/azure/hdinsight/hdinsight-hadoop-create-linux-clusters-portal
Выполнить задания:
1) Создание кластера Apache Spark в Azure HDInsight
https://docs.microsoft.com/ru-ru/azure/hdinsight/spark/apache-spark-jupyter-spark-sql

2) Создание приложений машинного обучения Apache Spark в Azure HDInsight

https://docs.microsoft.com/ru-ru/azure/hdinsight/spark/apache-spark-ipython-notebook-machine-learning

3) Использование Spark MLlib для создания приложения машинного обучения и анализа набора данных

https://docs.microsoft.com/ru-ru/azure/hdinsight/spark/apache-spark-machine-learning-mllib-ipython

Задание 1.3. Библиотека машинного обучения Mahout для Apache Hadoop

Mahout http://mahout.apache.org/ – это библиотека машинного обучения для Apache Hadoop. Mahout содержит алгоритмы для обработки данных, такие как фильтрация, классификация и кластеризация.

Задание: Подключиться к Hadoop через SSH. Реализовать систему рекомендаций с помощью библиотеки Mahout.

Подключение к HDInsight (Hadoop) с помощью SSH

https://docs.microsoft.com/ru-ru/azure/hdinsight/hdinsight-hadoop-linux-use-ssh-unix

OpenSSH (бета-версия). В Fall Creators Update выберите Параметры > Приложения и возможности > Управление дополнительными компонентами > Добавить функцию и Клиент OpenSSH.

https://blogs.msdn.microsoft.com/powershell/2017/12/15/using-the-openssh-beta-in-windows-10-fall-creators-update-and-windows-server-1709/

либо PuTTY

https://www.chiark.greenend.org.uk/~sgtatham/putty/

Создание списка рекомендуемых фильмов с помощью Apache Mahout и Hadoop в HDInsight (SSH) на платформе Linux

https://docs.microsoft.com/ru-ru/azure/hdinsight/hadoop/apache-hadoop-mahout-linux-mac

Задание 1.4. Платформа Hadoop MapRedice

Hadoop MapReduce – это программная платформа для написания заданий, обрабатывающих большие объемы данных. Входные данные разбиваются на независимые блоки, которые затем обрабатываются параллельно на узлах кластера.

Использование MapReduce в Hadoop в HDInsight

https://docs.microsoft.com/ru-ru/azure/hdinsight/hadoop/hdinsight-use-mapreduce

Задание: реализовать приложение для подсчета слов в записных книжках Леонардо да Винчи на Java _либо_ на Python, используя команды Hadoop через SSH.

Разработка программ MapReduce на Java для Hadoop в HDInsight

https://docs.microsoft.com/ru-ru/azure/hdinsight/hadoop/apache-hadoop-develop-deploy-java-mapreduce-linux

Разработка программ MapReduce с потоковой передачей Python для HDInsight

https://docs.microsoft.com/ru-ru/azure/hdinsight/hadoop/apache-hadoop-streaming-python

Критерии оценки:

17-20 баллов выставляется студенту, если он своевременно выполнил все задачи, предусмотренные в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на дополнительные вопросы, связанные не только с процессом выполнения практической работы, но и с пониманием совершенных действий и решенных задач.

14-16 баллов выставляется студенту, если он выполнил от 70% задач, предусмотренных в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания

учебной дисциплины, относящейся к практической работе. Сумел ответить на вопросы, связанные с процессом выполнения практической работы.

12-13 баллов выставляется студенту, если он более чем на 50% выполнил поставленные в практической работе задачи, способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

0 баллов выставляется студенту, если он более чем на 50% не выполнил поставленные в практической работе задачи, не способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

10.3. Практическая работа №2

Графовая СУБД Neo4j

В качестве основного источника справочной информации по графовой СУБД Neo4j и языку Cypher используйте ресурс http://docs.neo4j.org/.

Задание 1.1. Разработка базы данных социальной сети

Используя графовую СУБД Neo4j, разработайте базу данных модельной социальной сети. Узел социальной сети имеет следующие атрибуты: ФИО, пол, возраст, город, список групп по интересам (например: спорт, игры, компьютеры). Семантика связи между узлами – дружеские отношения соответствующих персон («А является другом Б»).

Для этого:

1. Разработайте скрипт с запросами на языке Cypher, которые загружают базу данных социальной сети в СУБД Neo4j. База должна содержать не менее 10 узлов и не менее 15 связей между ними.
2. Загрузите базу данных в СУБД Neo4j при помощи консольного клиента (Neo4jShell.bat).

Задание 1.2. Разработка запросов к графам (2 часа)

Разработайте и протестируйте запросы на выборку данных из созданной графовой базы данных.

Для этого:

1. Изучите справочную информацию о языке Cypher по теме Reading Clauses и разработайте следующие запросы:
   1) Выдать упорядоченный список ФИО персон.
   2) Выдать список ФИО мужчин с указанием возраста, упорядоченный по убыванию возраста.
   3) Выдать упорядоченный список ФИО друзей персоны заданными ФИО.
   4) Выдать упорядоченный список ФИО друзей друзей персоны заданными ФИО.
   5) Выдать упорядоченный по алфавиту список ФИО персон, в котором для каждой персоны указано количество друзей.
2. Изучите справочную информацию о языке Cypher по теме Functions и разработайте следующие запросы:
   1) Выдать упорядоченный список групп социальной сети.
   2) Выдать упорядоченный список групп персоны с заданными ФИО.
   3) Выдать список групп социальной сети с указанием количества членов каждой группы, упорядоченный по убыванию количества членов группы.

4) Выдать список ФИО персон, в котором для каждой персоны указано количество групп, в которые она входит, упорядоченный по убыванию количества групп.
5) Выдать общее количество групп, в которых состоят друзья друзей персоны с заданными ФИО.

Задание 1.3. Визуализация графа

Разработайте и протестируйте запросы на обновление свойств узлов графа и выполните с их помощью визуализацию графа.

Для этого:

1. Изучите справочную информацию по средствам визуализации Neo4j. Отоб- разите граф.
2. Добавьте свойство *friend* всем друзьям одной персоны и настройте профиль визуализации таким образом, чтобы узлы со свойством *friend* отображались красным цветом.
3. Добавьте свойство *twoHandFriend* всем друзьям друзей одной персоны и настройте профиль визуализации таким образом, чтобы узлы со свойством *twoHandFriend* отображались желтым цветом.
4. Добавьте свойство *manyFriends* персонам, имеющим больше 5 друзей, свойство *fewFriends* – персонам, у которых меньше 3 друзей. Настройте профиль визуализации таким образом, чтобы узлы со свойством *manyFriends* отображались зеленым цветом, узлы со свойством *fewFriends* отображались красным цветом, а все остальные узлы – желтым цветом.
5. Добавьте свойство *group1* персонам, состоящим в некоторой группе, *group2* – персонам, состоящим в другой группе, а свойство *bothGroups* -- персонам, состоящим в обеих группах. Настройте профиль визуализации таким обра- зом, чтобы узлы со свойством *group1* отображались синим цветом, узлы со свойством *group2* отображались красным цветом, а узлы со свойством *bothGroups* – фиолетовым цветом.

Критерии оценки:

17-20 баллов выставляется студенту, если он своевременно выполнил все задачи, предусмотренные в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на дополнительные вопросы, связанные не только с процессом выполнения практической работы, но и с пониманием совершенных действий и решенных задач.

14-16 баллов выставляется студенту, если он выполнил от 70% задач, предусмотренных в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на вопросы, связанные с процессом выполнения практической работы.

12-13 баллов выставляется студенту, если он более чем на 50% выполнил поставленные в практической работе задачи, способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

0 баллов выставляется студенту, если он более чем на 50% не выполнил поставленные в практической работе задачи, не способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

10.4. Практическая работа №3

Документ-ориентированная СУБД MongoDB

В качестве основного источника справочной информации по документ- ориентированной СУБД MongoDB используйте ресурсы http://docs.mongodb.org/manual/ и http://jsman.ru/mongo-book/.

Задание 1.1. Разработка базы данных с использованием СУБД MongoDB
Используя СУБД MongoDB, разработайте базу данных, предназначенную для хранения логов веб-сервера. Лог включает в себя следующие поля: адрес ресурса (URL), IP-адрес пользовательского компьютера, отметка времени начала просмотра ресурса, длительность просмотра ресурса.

Для этого:

1. Разработайте консольную утилиту для преобразования лога веб-сервера в формате CSV (Comma Separated Values), в формат JSON. Лог должен со- держать поля со следующими названиями: URL, IP, timeStamp, timeSpent.
2. Разработайте запросы для загрузки полученных данных в формате JSON в СУБД MongoDB.

Задание 1.2. Разработка запросов в СУБД MongoDB

1. Разработайте и протестируйте запросы на выборку данных из созданных коллекций.
2. Разработайте и протестируйте функции MapReduce для анализа посещаемости ресурсов web-сервера.

Для этого:

1. Разработайтеследующие запросы, используя встроенные в СУБД MongoDB средствами выборки:
   1) Выдать упорядоченный список URL ресурсов.
   2) Выдать упорядоченный список IP-адресов пользователей, посетивших ресурс с заданным URL.
   3) Выдать упорядоченный список URL ресурсов, посещенных в заданный временной период.
   4) Выдать упорядоченный список URL ресурсов, посещенных пользователем с заданным IP-адресом.
2. Разработайтеследующие запросы, используя встроенные в СУБД MongoDB средства программирования на основе парадигмы MapReduce:
   1) Выдать список URL ресурсов с указанием суммарной длительности посещения каждого ресурса, упорядоченный по убыванию.
   2) Выдать список URL ресурсов с указанием суммарного количества посещений каждого ресурса, упорядоченный по убыванию.
   3) Выдать список URL ресурсов с указанием количества посещений каждого ресурса в день за заданный период, упорядоченный URL ресурса и убыванию количества посещений.
   4) Выдать список IP-адресов с указанием суммарного количества и суммарной длительности посещений ресурсов, упорядоченный по адресу, убыванию количества и убыванию длительности.

Критерии оценки:

17-20 баллов выставляется студенту, если он своевременно выполнил все задачи, предусмотренные в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на дополнительные вопросы, связанные не только с процессом выполнения практической работы, но и с пониманием совершенных действий и решенных задач.

14-16 баллов выставляется студенту, если он выполнил от 70% задач, предусмотренных в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на вопросы, связанные с процессом выполнения практической работы.

12-13 баллов выставляется студенту, если он более чем на 50% выполнил поставленные в практической работе задачи, способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

0 баллов выставляется студенту, если он более чем на 50% не выполнил поставленные в практической работе задачи, не способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

Методические рекомендации по выполнению практических работ

Целью практических работ является приобретение практических навыков использования математических моделей, методов и алгоритмов в области технологий анализа больших данных; усвоение полученных знаний студентами, а также формирование у них мотивации к самообразованию за счет активизации самостоятельной познавательной деятельности.

Все работы выполняются студентами в рамках 4-х академических часов, которые отведены учебным планом.

Итогом работы является защита полученных результатов. Защита проводится индивидуально в форме собеседования и проверке полученных навыков работы с системой на компьютере.

10.5. Тест № 1

1. **Текст задания:**

Вы собираетесь использовать библиотеку машинного обучения RevoScaleR для распределённой обработки данных в среде Apache Hadoop. Какие источники данных вы будете использовать?

A. База данных Microsoft SQL Server.
B. Файлы данных XDF.
C. Источники данных ODBC.
D. База данных Teradata.

Ответ (ключ):

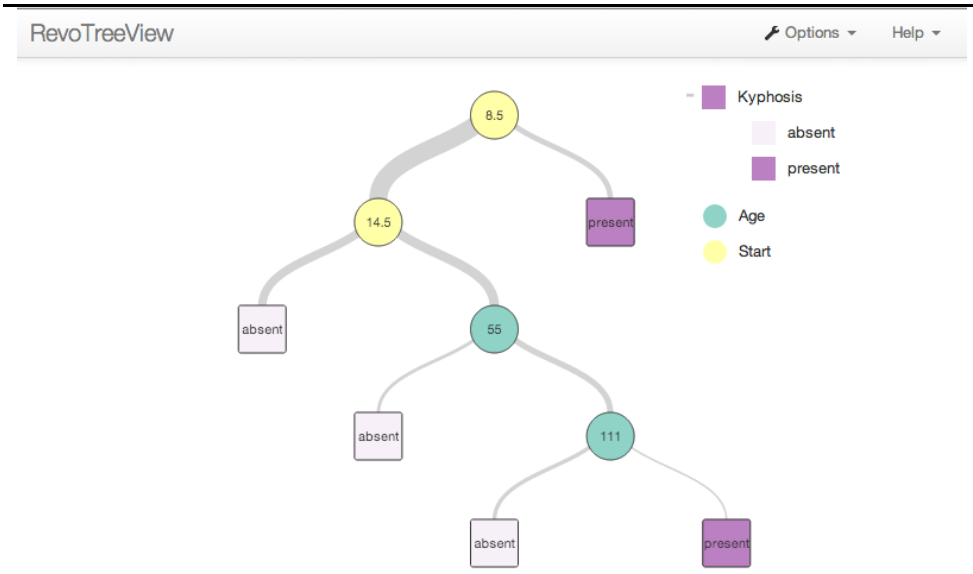
B

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

2. **Текст задания:**

Классификационная модель в пакете rpart в среде R построила дерево решений, приведенное на рисунке ниже:



Сколько листовых вершин в этом дереве?

Ответ (ключ):

5

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

3. **Текст задания:**

В ходе эксперимента в Azure ML модель демонстрирует много ошибок на подтверждающем наборе данных и всего несколько ошибок на обучающих данных. В чем наиболее вероятная причина этих ошибок?

A. Переобучение.
B. Обобщение.
C. Недообучение.
D. Слишком простой предиктор.
E. «Проклятие большой размерности»

Ответ (ключ):

A

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

4. **Текст задания:**

Как ключи и значения представляются и передаются процессу свертки (Reducer) во время стандартной фазы сортировки и перемешивания в MapReduce?

A. Ключи передаются процессу свертки в отсортированном порядке, а значения для ключа сортируются в порядке возрастания.
B. Ключи передаются процессу свертки в отсортированном порядке, а значения для ключа не сортируются.

C. Ключи передаются процессу свертки в случайном порядке, а значения для ключа не сортируются.
D. Ключи передаются процессу свертки в случайном порядке, а значения для ключа сортируются в порядке возрастания.

Ответ (ключ):

B

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

5. **Текст задания:**

В какой момент можно вызвать метод свертки в MapReduce?

A. Как только хотя бы один процесс отображения закончит обработку записей.
B. Зависит от InputFormat, используемого для задания.
C. Не раньше, чем все процессы отображения закончат обработку записей.
D. Как только процесс отображения закончит обработку первой записи.

Ответ (ключ):

C

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

6. **Текст задания:**

Укажите правильную последовательность для записи в HDFS.

A. Клиент HDFS кэширует пакеты данных в памяти.
B. NameNode предоставляет DataNode информацию о расположении реплик блоков.
C. Клиент передает пакет данных первой целевой DataNode.

Ответ (ключ):

A, B, C

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

7. **Текст задания:**

На изображении граница предмета выделена областью красного цвета. Каким типом прерывности (Discontinuity) характеризуется эта область?

A. Прерывность по глубине.
B. Прерывность по цвету поверхности.
C. Прерывность по освещенности.
D. Ни один из перечисленных типов.

Ответ (ключ):

A

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

8. **Текст задания:**

Вы собрали данные о 10 тысячах сообщений в Twitter, и больше никакой информации. Вы хотите создать модель классификации твитов, которая классифицирует каждый из твитов в один из трех классов – положительный, отрицательный или нейтральный.

Какая из моделей может выполнить классификацию твитов для анализа тональности текста?

A. Naïve Bayes.
B. SVM.
C. Naïve Bayes и SVM.
D. Ни одна из перечисленных

Ответ (ключ):

D

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

9. **Текст задания:**

Рассмотрим три таблицы Tab1, Tab2 и Tab3, которые содержат 10, 15 и 20 записей, соответственно. 5 записей являются общими во всех трех таблицах.

Вы решаете объединить три таблицы при помощи перекрестного соединения. Сколько строк будет в декартовом произведении этих таблиц?

Ответ (ключ):

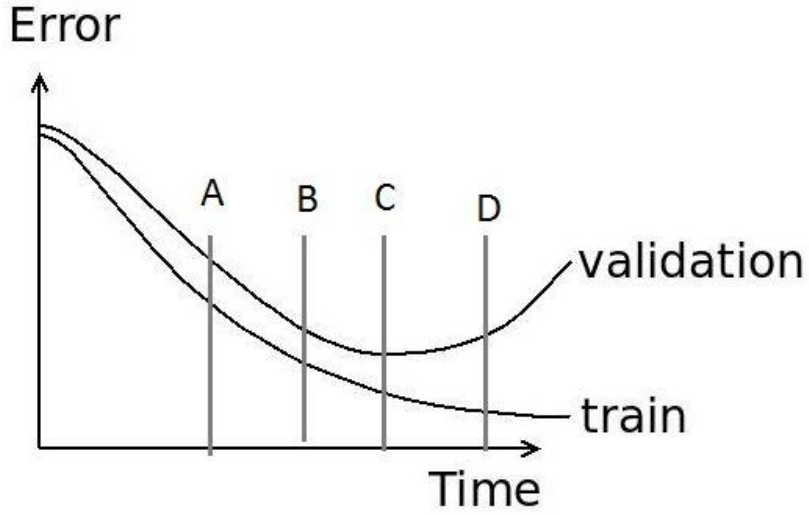
3000

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

10. **Текст задания:**

При обучении нейронной сети в задаче распознавания образов был построен следующий график ошибки обучения и ошибки валидации модели.



Какой момент времени лучше всего подходит для раннего останова?

A. A
B. B
C. C
D. D

Ответ (ключ):

C

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

11. **Текст задания:**

Туристическое агентство продает авиабилеты клиентам в Европейском Союзе. Агентство хочет, чтобы вы выявили закономерности и сделали прогноз о задержках

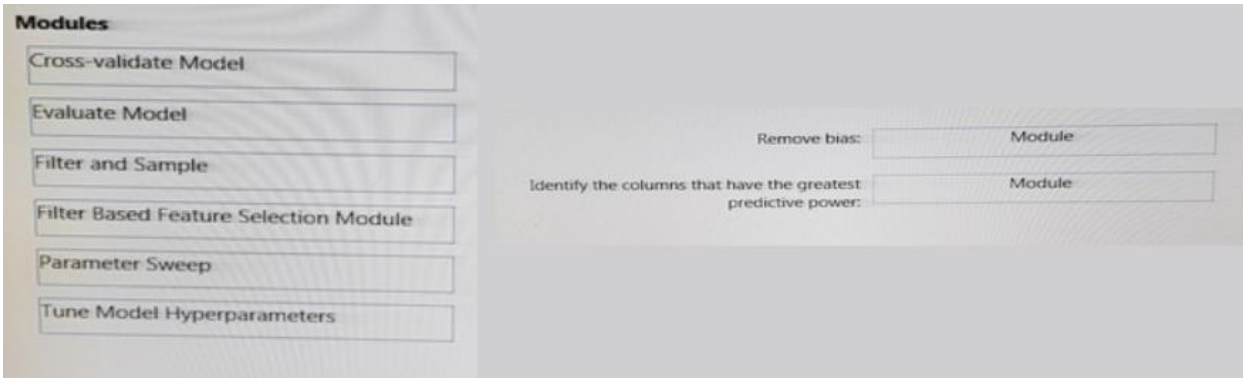
рейсов. Агентство рассматривает возможность внедрения системы, которая будет информировать своих клиентов о возможных задержках из-за погодных условий.

Данные о полетах содержат следующие атрибуты:
- DepartureDate: Время отправления с детализацией по часам.
- Carrier: Код, присвоенный IATA и обычно используемый для идентификации перевозчика.
- OriginAirportID: Идентификационный номер, присвоенный USDOT для идентификации уникального аэропорта (происхождение рейса).
- DestAirportID: Время задержки в минутах.
- DepDet30: Булево значение, указывающее, было ли отправление задержано более, чем на 30 минут (значение 1 указывает, что отправление было задержано на 30 минут или более).
Данные о погоде содержат следующие атрибуты: AirportID, ReadingDate (YYYY/MM/DD
HH), SkyConditionVisibility, WeatherType, Windspeed, StationPressure, PressureChange and HourlyPrecip.

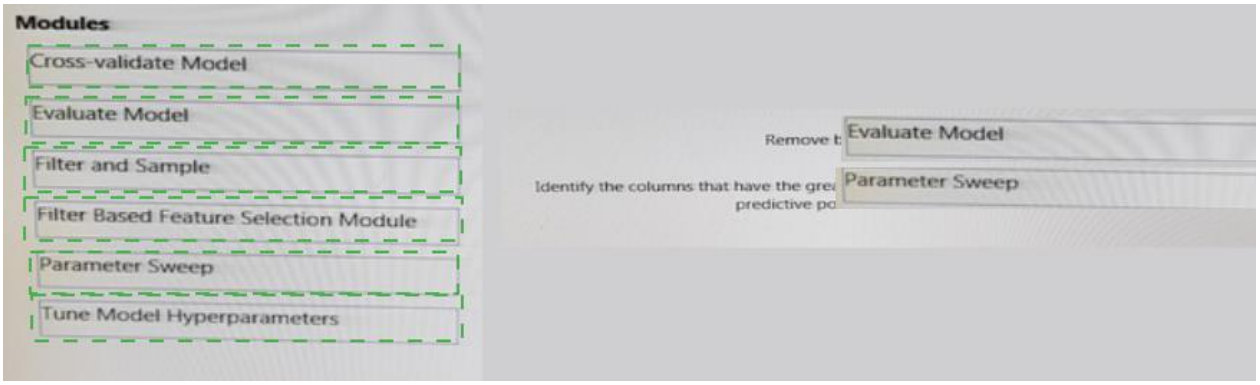
Необходимо устранить смещение (bias) и определить столбцы входного набора данных, которые обладают наибольшей предсказательной силой. Какой модуль в AzureML следует использовать на каждом из этапов?

Устранить смещение: _____

Определить столбцы с наибольшей предсказательной силой: _____



Ответ (ключ):



Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

12. **Текст задания:**

Необходимо создать задание, которое выполняет частотный анализ входных данных. Вы решаете сделать это, написав процесс отображения (Mapper), который использует

TextInputFormat и разбивает каждое значение (строку текста из входного файла) на отдельные символы. Для каждого из этих символов вы будете выдавать символ в качестве ключа и InputWritable в качестве значения. Поскольку это приведет к пропорционально большему объему промежуточных данных, чем входных данных, какие два ресурса могут стать узкими местами?

A. Процессор и сетевой ввод-вывод.
B. Процессор и дисковый ввод-вывод.
C. Процессор и оперативная память.
D. Дисковый ввод-вывод и сетевой ввод-вывод.

Ответ (ключ):

D

Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

13. **Текст задания:**

Многозначность определяется как сосуществование нескольких значений для слова или фразы в текстовом объекте. Какая из следующих моделей, вероятно, является лучшим выбором для устранения этой проблемы?

A. Классификатор «случайный лес».
B. Сверточные нейронные сети.
C. Градиентный бустинг.
D. Ни одна из перечисленных.

Ответ (ключ):

B

Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

14. **Текст задания:**

Ниже приведена входная матрица размерности 7 X 7. Какой будет выходная матрица, если применить сверточный слой 3 X 3 со страйдом 2?

1	**2**	**4**	**1**	**4**	**0**	**1**
0	**0**	**1**	**6**	**1**	**5**	**5**
1	**4**	**4**	**5**	**1**	**4**	**1**
4	**1**	**5**	**1**	**6**	**5**	**0**
1	**0**	**6**	**5**	**1**	**1**	**8**
2	**3**	**1**	**8**	**5**	**8**	**1**
0	**9**	**1**	**2**	**3**	**1**	**4**

A.

4	6	5
6	6	8
9	8	8

B.

4	5	5
6	6	8

9	8	6

C.

4	5	6
3	6	8
9	9	6

D.

4	3	3
3	3	3
4	3	4

Ответ (ключ):

A

Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

15. **Текст задания:**

Предположим, у нас есть 5-слойная нейронная сеть, обучение которой занимает 3 часа на GPU с 4 Гб VRAM. Во время теста на каждую точку данных требуется 2 секунды.

Теперь мы изменяем архитектуру, добавляя исключения после 2-го и 4-го слоев с коэффициентом дропаута 0,2 и 0,3 соответственно.

Каково будет время тестирования этой новой архитектуры?

A. Менее 2 секунд.
B. 2 секунды.
C. Больше 2 секунд.
D. Сложно сказать.

Ответ (ключ):

B

Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

Критерии оценки:

Тестирование оценивается дифференцированно по балльной шкале:

- выполнено без ошибок и недочетов 85-100% от общего объема заданий – выставляется от 18 до 20 баллов;
- выполнено без ошибок и недочетов 71-84% от общего объема заданий – выставляется от 15 до 17 баллов;
- выполнено без ошибок и недочетов 60-70% от общего объема заданий – выставляется от 13 до 14 баллов;
- выполнено без ошибок и недочетов 50-59% от общего объема заданий – выставляется от 10 до 12 баллов.

Если выполнено менее 50% от общего объема заданий – тестирование считается не пройденным, студент должен пройти его повторно.

10.4. Тест № 2

1. **Текст задания:**

У вас есть облачные и локальные ресурсы, которые включают Microsoft SQL Server и среду больших данных в Apache Hadoop. У вас есть 10 миллиардов фактических записей. Для выполнения прогнозных отчетов необходимо построить модели временных рядов. Что вы будете использовать?

A. RxSpark в кластере Hadoop.
B. RxHadoopMR в кластере Hadoop.
C. RxLocalseq в базе данных SQL Server.
D. RxLocalParallel в базе данных SQL Server.

Ответ (ключ):

A

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

2. **Текст задания:**

Вы анализируете поездки на такси в мегаполисе. Azure Data Factory используется для создания конвейеров данных и организации движения данных.

Вы планируете разработать прогнозную модель для 170 миллионов строк (37 Гб) необработанных данных в Apache Hive с помощью Microsoft R Server, чтобы определить, какие факторы влияют на размер чаевых.

Все платформы, которые используются для анализа, одинаковые. Каждый рабочий узел имеет 8-ядерный процессор и 32 Гб оперативной памяти.

Какой тип кластера Azure HDInsight следует использовать для получения результатов как можно быстрее?

A. Hadoop
B. HBase
C. HDInsight Interactive Query
D. Spark

Ответ (ключ):

C

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

3. **Текст задания:**

В таблице приведено описание клинического набора данных по выживаемости при раке молочной железы. Какой фрагмент кода в Spark поможет вам определить средний возраст, при котором был поставлен первоначальный патологический диагноз?

Номер столбца	Название столбца	Тип данных	Описание

Column 0	Complete_TCGA_ID	string	Уникальный идентификатор для пациента
Column 1	Gender	string	Женский
Column 2	Age_Initial_diag	int	Возраст пациента на момент поступления
Column 3	ER_Status	string	Эстроген-рецепторы ER-положительного (ER+) рака молочной железы.
Column 4	PR_Status	string	Клетки рака молочной железы имеют рецепторы прогестерона, рак называется PR-положительным раком молочной железы.
Column 5	HER2_Final_Status	string	HER2-положительный рак молочной железы – это рак молочной железы определяемый при наличии белка, называемого человеческим эпидермальным фактором роста 2-го типа (HER2), который способствует росту раковых клеток. Примерно в 1 из каждых 5 случаев рака молочной железы раковые клетки имеют мутацию гена, который производит избыток белка HER2.
Column 6	Tumor	string	Числа после T (T1, T2, T3 и T4) характеризуют размер опухоли и / или степень распространения в близлежащие структуры. Чем выше число T, тем больше опухоль и / или тем больше она выросла в близлежащие ткани.
Column 7	Node	string	Числа после N (N1, N2 и N3) характеризуют размер, расположение и / или количество близлежащих лимфатических узлов, пораженных раком. Чем выше число N, тем сильнее рак распространяется на близлежащие лимфатические узлы. N0 означает, что близлежащие лимфатические узлы не содержат рака.
Column 8	Node_Coded	string	Положительное значение означает, что рак присутствует. Отрицательное значение означает, что рака нет.
Column 9	Metastasis	string	M0 означает, что рака нет. M1 означает, что рак присутствует.
Column 10	Metastasis_Coded	string	Положительное значение означает, что рак присутствует. Отрицательное значение означает, что рака нет.
Column 11	AJCC_Stage	string	Система стадирования AJCC – это система классификации для описания степени прогрессирования заболевания у онкологических больных.

Column 12	Converted_Stage	string	После лечения любые изменения в стадии AJCC.
Column 13	Survival_Data_Form	string	Н/д
Column 14	Vital_Status	string	Человек жив или мертв / болен.
Column 15	Days_to_Date_ of_Last_Contact	int	Количество дней, прошедших с последнего обращения пациента.

A. val data = sc.textFile("clinical_data_breast_cancer.csv ")
val header = data.first()
val remove_header = data.filter(x => x!=header)
val avg_age = remove_header.map(x => x.split(",")).map(x =>
x(2).toInt).reduce(_+_)/remove_header.count

B. val data = sc.textFile("clinical_data_breast_cancer.csv")
val header = data.first()
val remove_header = data.filter(x => x!=header)
val stages = remove_header.map(x => x.split(",")).map(x =>
(x(11),x(15).toInt)).mapValues((_, 1)).reduceByKey((x, y) => (x._1 + y._1, x._2 +
y._2)).mapValues{ case (sum, count) => (1.0 * sum)/count}.foreach(println)

C. val data = sc.textFile("clinical_data_breast_cancer.csv ")
val header = data.first()
val remove_header = data.filter(x => x!=header)
val status = remove_header.map(x => x.split(",")).map(x =>
(x(14),1)).reduceByKey(_+_).foreach(println)

Ответ (ключ):

A

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

4. **Текст задания:**

Вы разрабатываете Комбайнер, который принимает текстовые ключи и значения IntWritable в качестве входных данных и выдает текстовые ключи и значения IntWritable . Какой интерфейс должен быть реализован вашим классом?

A. Combiner <Text, IntWritable, Text, IntWritable>
B. Reducer <Text, Text, IntWritable, IntWritable>
C. Combiner <Text, Text, IntWritable, IntWritable>
D. Reducer <Text, IntWritable, Text, IntWritable>
F. Mapper <Text, IntWritable, Text, IntWritable>

Ответ (ключ):

A

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

5. **Текст задания:**

Есть каталог файлов со следующей структурой: номер строки, символ табуляции, строка.

Например:

1 fdprdvffrbroxlvzmtbd
2 qwpwdwtdqlkbzwgwmhma
3 vyatquegqbnbtsojinrm

Если вы хотите ввести каждую строку в качестве одной записи в процесс отображения в MapReduce, то какой InputFormat следует использовать для завершения строки: conf.setInputFormat (_____.class) ; ?

A. BDBInputFormat.
B. SequenceFileInputFormat.
C. KeyValueFileInputFormat.
D. SequenceFileAsTextInputFormat.

Ответ (ключ):

B

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

6. **Текст задания:**

В задании MapReduce вы хотите, чтобы каждый из ваших входных файлов был обработан отдельным процессом отображения. Как настроить задание MapReduce таким образом, чтобы отдельный процесс отображения обрабатывал каждый входной файл независимо от того, сколько блоков занимает входной файл?

A. Увеличить значение параметра, который управляет минимальным размером разделения в конфигурации задания.
B. Написать свой MapRunner, который повторяет все пары ключ-значение во всем файле.
C. Написать свой FileInputFormat и переопределить метод isSplitable, чтобы он всегда возвращал false.
D. Установить число процессов отображения равным числу входных файлов, которые следует обработать.

Ответ (ключ):

C

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

7. **Текст задания:**

Вы используете команду Hadoop *fs -put* чтобы сохранить файл размером 300 Мб в ранее пустой каталог, используя блоки HDFS размером 64 Мб. Команда только что закончила запись 200 Мб этого файла. Что другие пользователи увидят, когда они попытаются получить доступ к этому файлу?

A. Они не увидят никакого содержимого, пока весь файл не будет записан и закрыт.

B. Hadoop вернет ConcurrentFileAccessException, когда они пытаются получить доступ к файлу.
C. Они будут видеть текущее состояние файла до последнего завершенного блока.
D. Они будут видеть текущее состояние файла до последнего бита, записанного командой.

Ответ (ключ):

A

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

8. **Текст задания:**

Какой из следующих методов аугментации данных вы бы использовали для решения задачи распознавания объектов?

A. Отражение по горизонтали.
B. Изменение масштаба.
C. Увеличение изображения.
D. Все перечисленные методы.

Ответ (ключ):

D

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

9. **Текст задания:**

Поисковые и генеративные модели – это два популярных метода, используемых для создания чат-ботов. Что будет примером модели поиска и генеративной модели соответственно?

A. Обучение со словарем и модель Word2vec.
B. Обучение на основе правил и модель Sequence-to-Sequence.
C. Модель Word2vec и модель Sentence-to-Vector.
D. Рекуррентная нейронная сеть и сверточная нейронная сеть.

Ответ (ключ):

B

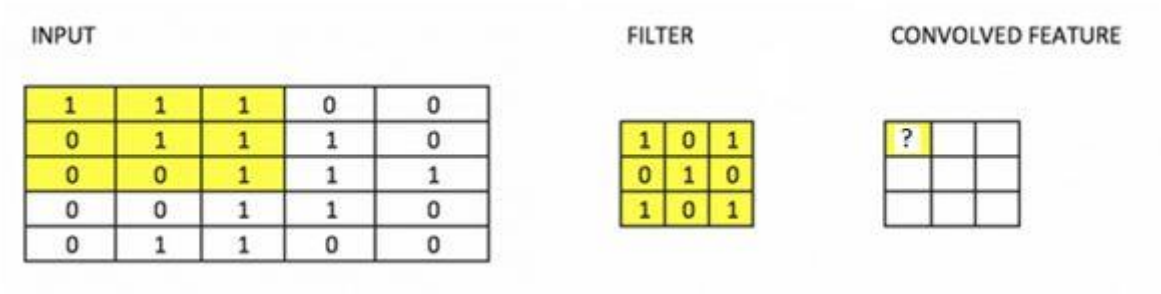
Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

Текст задания:

10
.

Ко входу нейронной сети применяется сверточная функция. Какое значение будет на месте вопросительного знака?



Ответ (ключ):

4

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

11
.

Текст задания:

Выберите правильный порядок для процесса чтения файла в HDFS.

A. Клиент открывает файл, вызывая метод open () объекта FileSystem, соответствующий файловой системе HDFS и вызывающий объект DistributedFileSystem.
B. DistributedFileSystem вызывает namenode через RPC, чтобы определить расположение первых нескольких блоков файла.
C. namenode для каждого блока возвращает адреса всех namenode, у которых есть резервная копия блока, а datanode сортируются по близости к клиенту в сети топологии кластера.
D. DistributedFileSystem возвращает FSDataInputStream, клиент может считывать данные из FSDataInputStream. FSDataInputStream обертывает класс DFSInputStream, который обрабатывает операции ввода-вывода для namenode и datanode.
E. Клиент выполняет метод read (), а DFSInputStream, который уже сохранил информацию о местоположении первых нескольких блоков считываемого файла, подключается к ближайшей datanode, чтобы получить данные.
F. Повторяя вызов метода read (), данные в файле передаются клиенту. Когда считан конец блока, DFSInputStream закрывает поток, указывающий на блок, и вместо этого находит информацию о местоположении следующего блока, а затем повторно вызывает метод read () для продолжения потоковой передачи блока.

Ответ (ключ):

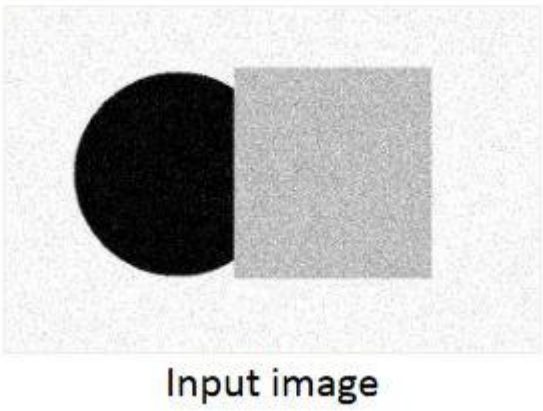
A, B, C, D, E, F

Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

Текст задания:

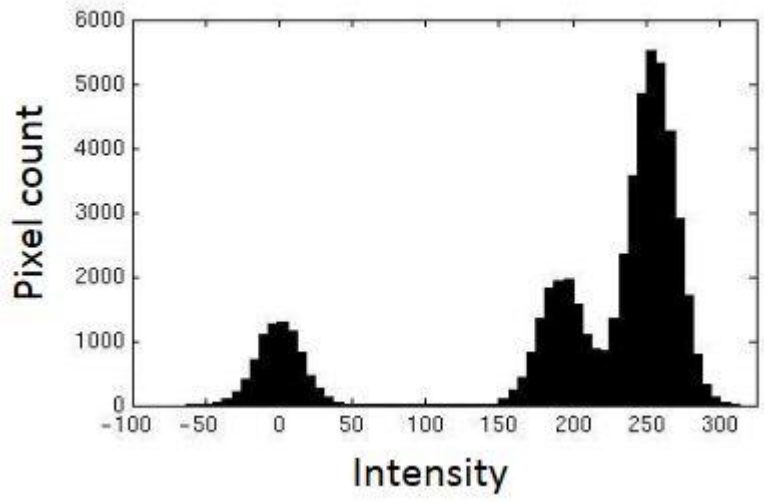
Предположим, у нас есть изображение, приведенное ниже.



Наша задача – сегментировать объекты на изображении. Простой способ сделать это – представить изображение в терминах интенсивности пикселей, а затем кластеризовать их в соответствии со значениями.

Сделав это, мы получили такую структуру:

12
.



Предположим, что мы выбираем кластеризацию методом k-средних для решения проблемы. Изучив график интенсивности, подберите оптимальное значение *k*.

Ответ (ключ):

3

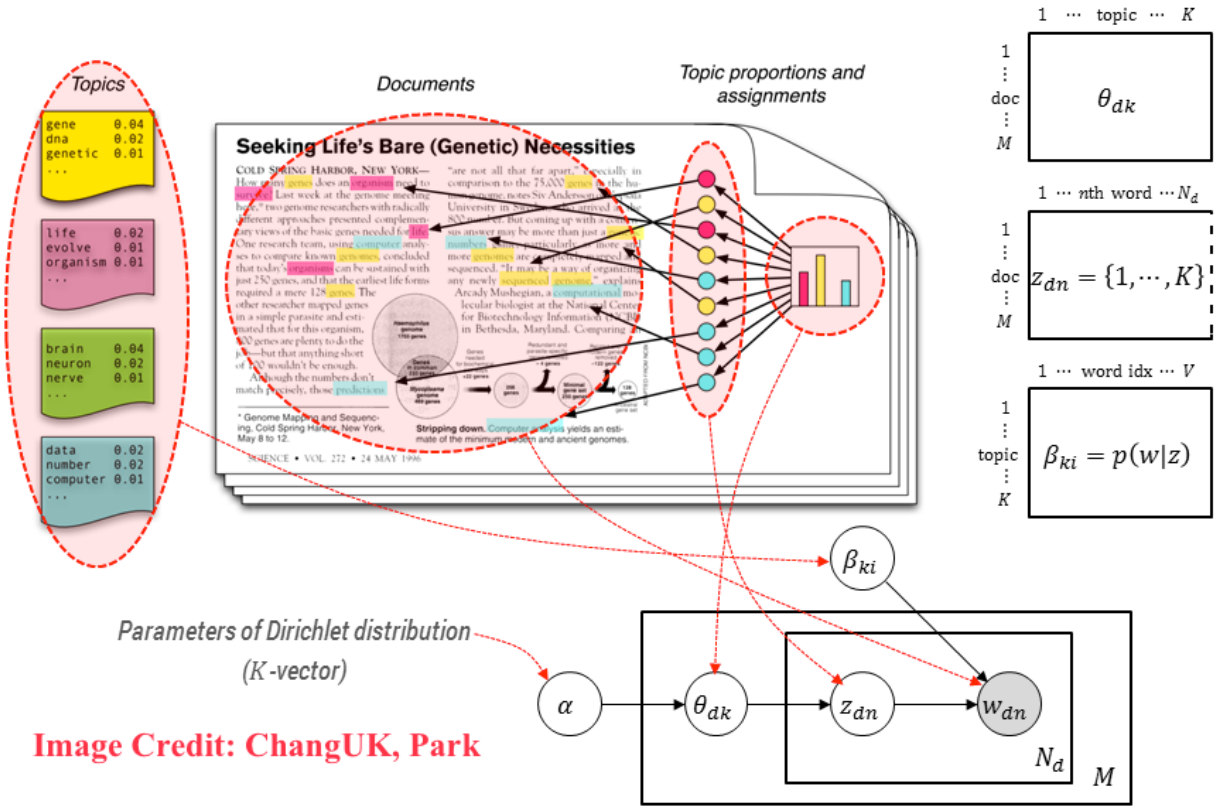
Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

13
.

Текст задания:

Что представляют собой гиперпараметры альфа и бета в модели латентного размещения Дирихле для целей классификации текста?



A. Альфа – количество тем в документах; бета – количество терминов в темах.
B. Альфа – плотность терминов, генерируемых в рамках тем; бета – плотность тем, генерируемых в рамках терминов.
C. Альфа – количество терминов в темах; бета – количество тем в документах.
D. Альфа – плотность тем, генерируемых в документах; бета – плотность терминов, генерируемых в рамках тем.

Ответ (ключ):

D

Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

14. **Текст задания:**

Предположим, вам дана следующая таблица A_V-1.

id	name	salary	dept
1	Oliver	100	DS
2	Harry	200	DS
3	George	800	ALL
4	Noah		INTERN
5	Jacob	1000	ALL

Каким будет результат следующего запроса?

Запрос: SELECT COALESCE(salary,2)+100 AS salary FROM A_V-1;

Ответ (ключ):

Salary
200
300
900
102
1100

Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

15. **Текст задания:**

Инициализация Ксавье чаще всего используется для инициализации весов нейронной сети. Ниже приведена формула инициализации.

$$Var(W) = \frac{2}{n_{in} + n_{out}}$$

Какие из следующих утверждений верные?

1. Если веса в начале невелики, то сигналы на выходе будут слишком малы.
2. Если веса в начале слишком велики, то сигналы на выходе будут слишком большими.
3. Веса берутся из гауссовского распределения.
4. Инициализация Ксавье помогает с проблемой исчезающего градиента.
5. Инициализация Ксавье используется для того, чтобы помочь входным сигналам проникнуть глубоко в сеть.

A. 1, 2, 4
B. 2, 3, 4
C. 1, 3, 4
D. 1, 2, 3
E. 1, 2, 3, 4

Ответ (ключ):

D

Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

Критерии оценки:

Тестирование оценивается дифференцированно по балльной шкале:

- выполнено без ошибок и недочетов 85-100% от общего объема заданий – выставляется от 18 до 20 баллов;
- выполнено без ошибок и недочетов 71-84% от общего объема заданий – выставляется от 15 до 17 баллов;
- выполнено без ошибок и недочетов 60-70% от общего объема заданий – выставляется от 13 до 14 баллов;

- выполнено без ошибок и недочетов 50-59% от общего объема заданий – выставляется от 10 до 12 баллов.

Если выполнено менее 50% от общего объема заданий – тестирование считается не пройденным, студент должен пройти его повторно.