АВТОНОМНАЯ НЕКОММЕРЧЕСКАЯ ОРГАНИЗАЦИЯ ВЫСШЕГО ОБРАЗОВАНИЯ
«ИНСТИТУТ МЕЖДУНАРОДНЫХ ЭКОНОМИЧЕСКИХ СВЯЗЕЙ»
INSTITUTE OF INTERNATIONAL ECONOMIC RELATIONS

Принята на заседании
Учёного совета ИМЭС
(протокол от 26 января 2022 г. № 6)

УТВЕРЖДАЮ
Ректор ИМЭС Ю.И. Богомолова
26 января 2022 г.

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 0162B6870090AF03B544352CE01F9DE4C7
Владелец: БОГОМОЛОВА ЮЛИЯ ИГОРЕВНА
Действителен: с 20.01.2023 до 20.04.2024

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

ПРОЕКТИРОВАНИЕ СИСТЕМ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ (BIG DATA)

по направлению подготовки
09.03.02 Информационные системы и технологии

Направленность (профиль)
«Информационные системы и сетевые технологии»

Москва – 2022

1. АННОТАЦИЯ К ДИСЦИПЛИНЕ

Рабочая программа дисциплины «Проектирование систем обработки больших данных (Big Data)» составлена в соответствии с требованиями ФГОС ВО по направлению подготовки 09.03.02 Информационные системы и технологии, утвержденным приказом Министерства образования и науки Российской Федерации от 19.09.2017 № 926.

Дисциплина «Проектирование систем обработки больших данных (Big Data)» предназначена для практического освоения обучающимися работы с технологиями информационного поиска и обработки больших данных, работы с инструментами анализа данных, основ математической статистики и теории вероятностей, основ математического моделирования.

Место дисциплины в структуре образовательной программы

Настоящая дисциплина включена в учебные планы по программам подготовки бакалавров по направлению 09.03.02 Информационные системы и технологии и входит в обязательную часть Блока 1.

Дисциплина изучается на 3 курсе в 6 семестре.

Цели и задачи дисциплины

Цель изучения дисциплины – формирование у обучающихся необходимых компетенций для успешного освоения образовательной программы, в частности, изучение математических методов и подходов, используемых в программных системах обработки и анализа больших данных, получение практического опыта работы с IT решениями в части обработки и анализа больших данных.

***Задачи изучения дисциплины*:**

- сформировать знания о сущности, структуре и видах математических моделей принятия решений;
- формирование знаний, умений и практического опыта создания и решения моделей, необходимых в сфере управления.

2. ПЕРЕЧЕНЬ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ, СООТНЕСЕННЫХ С ПЛАНИРУЕМЫМИ РЕЗУЛЬТАТАМИ ОСВОЕНИЯ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Процесс изучения дисциплины направлен на формирование следующих компетенций, предусмотренных образовательной программой.

Результаты освоения ООП (содержание компетенций)	**Код компе-тенции**	**Код и наименование индикатора достижения компетенций**	**Перечень планируемых результатов обучения по дисциплине**			**Формы образовательной деятельности**
			выпускник должен знать	**выпускник должен уметь**	**выпускник должен иметь практический опыт**	
Способен решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности	**ОПК -3**	**ОПК-3.1. -** знает: принципы, методы и средства решения стандартных задач профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности	• принципы и методы работы с Большими Данными на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности	• использовать параллельные и масштабируемые алгоритмы обработки информации различных видов при решении практических задач в области информационных систем и технологий на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности	• поиска и обработки информации различных видов с использованием современных компьютерных технологий в соответствии с принятыми идеями и подходами к решению задачи	Контактная работа: Лекции Лабораторные практикумы Самостоятельная работа
		ОПК-3.2. - умеет: решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности	• принцип построения современных компьютерных кластеров для обработки Больших Данных	• использовать современные средства для обработки больших объемов информации	• навыков подготовки отчетов для руководителей	
		ОПК-3.3. - имеет	требования к	готовить обзор	применения	

Результаты освоения ООП (содержание компетенций)	Код компе-тенции	Код и наименование индикатора достижения компетенций	Перечень планируемых результатов обучения по дисциплине			Формы образовательной деятельности
			выпускник должен знать	выпускник должен уметь	выпускник должен иметь практический опыт	
		навыки: подготовки обзоров, аннотаций, составления рефератов, научных докладов, публикаций и библиографии по научно-исследовательской работе с учетом требований информационной безопасности	документации, описывающей работу системы обработки больших данных	системы обработки больших данных	требования информационной безопасности при проектировании систем обработки больших данных	

3. ТЕМАТИЧЕСКИЙ ПЛАН

Наименование тем	Контактная работа обучающихся с преподавателем (по видам учебных занятий)									Самостоятельная работа обучающихся	ТКУ / балл Форма ПА
	Лекции	Семинары	Практикум по решению задач	Ситуационный практикум	Мастер-класс	Лабораторный практикум	Тренинг	Дидактическая	Из них в форме практической подготовки		
Очная форма											
Тема 1. Большие данные: термины, проблемы	8					12				15	Отчет по лабораторному практикуму/20
Тема 2. Обработка структурированной информации	10					12				15	Отчет по лабораторному практикуму/20
Тема 3. Обработка слабоструктурированной информации	10					12				15	Отчет по лабораторному практикуму/20
Тема 4. Обработка визуальной информации	10					14				16	Отчет по лабораторному практикуму/20
Тема 5. Неструктурированная информация на примере корпуса текстов	10					14				16	Отчет по лабораторному практикуму/20
Всего:	**48**					**64**				**77**	**100**
Контроль, час	**27**										**Экзамен**
Объем дисциплины (в академических часах)	**216**										
Объем дисциплины (в зачетных единицах)	**6**										

4. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Тема 1. Большие данные: термины, проблемы.

Основные термины Больших данных. Области применения Больших данных. Рентабельность Больших Данных. Примеры использования: CRM, построение маршрутов Яндекс-навигатором и т.д.

Тема 2. Обработка структурированной информации

Визуальные образы данных: временной ряд, диаграмма разброса, тепловая карта. Задача выявления взаимосвязей, понятие корреляции и детерминации. Выявление многофакторных связей и проклятие размерности.

Тема 3. Обработка слабоструктурированной информации

Звук как временной ряд и изображение как двумерное поле. Пространство времени и пространство частот для звука. Пространство пиксельных координат и текстурных масштабов для изображения. Взаимно однозначное соответствие пространств. Временные и пространственные фильтры, убирание артефактов. Введение в нелинейные фильтры для изображения и сверточные нейросети.

Тема 4. Обработка визуальной информации

Понятие цвета, цветовой треугольник. Представление координат RGB-пикселя в различных цветовых пространствах. Восприятие цвета глазом человека, проблема переносимости и адекватного воспроизведения цвета ПЗС-матрицами, дисплеями и принтерами. Методы улучшения видимости изображения и его участков: изменение контраста, эквализация гистограммы яркости. Комбинирование изображений: технология HDR. Введение в вычислительную фотографию.

Тема 5. Неструктурированная информация на примере корпуса текстов

Задачи обработки текстовой информации: анализ настроений текста, идентификация авторства (деанонимизация). Разделение текста на слова, фильтрация стоп-слов и стоп-символов. Кластеризация слов и предложений, расстояние Левенштейна между словами и расстояние Хэмминга между последовательностями слов. Лемматизация слов. Кластеризация текстов.

5. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

В процессе изучения данной дисциплины используются такие виды учебной работы, как лекция, лабораторный практикум, а также различные виды самостоятельной работы обучающихся по заданию преподавателя, направленные на развитие навыков использования профессиональной лексики, закрепление практических профессиональных компетенций, поощрение интеллектуальных инициатив.

Методические указания для обучающихся при работе над конспектом лекций во время проведения лекции

Лекция – систематическое, последовательное, монологическое изложение преподавателем учебного материала, как правило, теоретического характера.

В процессе лекций рекомендуется вести конспект, что позволит впоследствии вспомнить изученный учебный материал, дополнить содержание при самостоятельной работе с литературой, подготовиться к экзамену.

Следует также обращать внимание на категории, формулировки, раскрывающие содержание тех или иных явлений и процессов, научные выводы и практические рекомендации, положительный опыт в ораторском искусстве. Желательно оставить в рабочих конспектах поля, на которых делать пометки из рекомендованной литературы, дополняющие материал прослушанной лекции, а также подчеркивающие особую важность тех или иных теоретических положений.

Любая лекция должна иметь логическое завершение, роль которого выполняет заключение. Выводы по лекции подытоживают размышления преподавателя по учебным вопросам. Формулируются они кратко и лаконично, их целесообразно записывать. В конце лекции, обучающиеся имеют возможность задать вопросы преподавателю по теме лекции.

Методические рекомендации для обучающихся по выполнению лабораторных практикумов

Лабораторные практикумы выполняются в соответствии с рабочим учебным планом при последовательном изучении тем дисциплины.

Порядок проведения практикума.

1. Получение задания и рекомендаций к выполнению практикума.
2. Настройка инструментальных средств, необходимых для выполнения практикума.
3. Выполнение заданий практикума.
4. Подготовка отчета в соответствии с требованиями.
5. Сдача отчета преподавателю или демонстрация работоспособности результата.

В ходе выполнения практикума необходимо следовать технологическим инструкциям, использовать материал лекций, рекомендованных учебников, источников интернета, активно использовать помощь преподавателя на занятии.

Лабораторные практикумы выполняются в соответствии с рабочим учебным планом при последовательном изучении тем дисциплины.

Прежде чем приступать к выполнению лабораторного практикума, обучающемуся необходимо:

- ознакомиться с соответствующими разделами программы дисциплины по учебной литературе, рекомендованной программой курса;
- получить от преподавателя рекомендации о порядке выполнения заданий;
- настроить под руководством преподавателя инструментальные средства,

необходимые для проведения лабораторного практикума

- получить от преподавателя конкретное задание и информацию о сроках выполнения, требованиях к оформлению, форме представления и критериях оценки результатов работы;

Требования к оформлению результатов практикумов.

При подготовке отчета: изложение материала должно идти в логической последовательности, отсутствие грамматических и синтаксических ошибок, шрифт Times New Roman, размер – 14, выравнивание по ширине, отступ первой строки – 1,25, междустрочный интервал – 1,5, правильное оформление рисунков (подпись, ссылка на рисунок в тексте).

При подготовке презентации: строгий дизайн, минимум текстовых элементов, четкость формулировок, отсутствие грамматических и синтаксических ошибок, воспринимаемая графика, умеренная анимация.

Методические указания для обучающихся по организации самостоятельной работы

Самостоятельная работа обучающихся направлена на самостоятельное изучение отдельных тем/вопросов учебной дисциплины.

Самостоятельная работа является обязательной для каждого обучающегося, ее объем по дисциплине определяется учебным планом.

При самостоятельной работе обучающиеся взаимодействуют с рекомендованными материалами при минимальном участии преподавателя.

Работа с литературой (конспектирование)

Самостоятельная работа с учебниками, учебными пособиями, научной, справочной и популярной литературой, материалами периодических изданий и Интернета, статистическими данными является наиболее эффективным методом получения знаний, позволяет значительно активизировать процесс овладения информацией, способствует более глубокому усвоению изучаемого материала, формирует у обучающихся свое отношение к конкретной проблеме.

Изучая материал по учебной книге (учебнику, учебному пособию, монографии, и др.), следует переходить к следующему вопросу только после полного уяснения предыдущего, фиксируя выводы и вычисления (конспектируя), в том числе те, которые в учебнике опущены или на лекции даны для самостоятельного вывода.

Особое внимание обучающийся должен обратить на определение основных понятий курса. Надо подробно разбирать примеры, которые поясняют определения. Полезно составлять опорные конспекты.

Выводы, полученные в результате изучения учебной литературы, рекомендуется в конспекте выделять, чтобы при перечитывании материала они лучше запоминались.

Вопросы, которые вызывают у обучающегося затруднение при подготовке, должны быть заранее сформулированы и озвучены во время занятий в аудитории для дополнительного разъяснения преподавателем.

Навигация для обучающихся по самостоятельной работе в рамках изучения дисциплины

Наименование темы	Вопросы, вынесенные на самостоятельное изучение	Формы самост. работы	Форма текущего контроля
Тема 1. Большие данные: термины, проблемы	Рентабельность Больших Данных.	Работа с литературой, включая ЭБС, источниками в сети Internet Подготовка к лабораторному практикуму, подготовка отчета по практикуму	Отчет по лабораторному практикуму
Тема 2. Обработка структурированной информации	Выявление многофакторных связей и проклятие размерности.	Работа с литературой, включая ЭБС, источниками в сети Internet Подготовка к лабораторному практикуму, подготовка отчета по практикуму	Отчет по лабораторному практикуму
Тема 3. Обработка слабоструктурированной информации	Временные и пространственные фильтры, убирание артефактов. Введение в нелинейные фильтры для изображения и сверточные нейросети.	Работа с литературой, включая ЭБС, источниками в сети Internet Подготовка к лабораторному практикуму, подготовка отчета по практикуму	Отчет по лабораторному практикуму
Тема 4. Обработка визуальной информации	Методы улучшения видимости изображения и его участков: изменение контраста, эквализация гистограммы яркости. Комбинирование изображений: технология HDR. Введение в вычислительную фотографию.	Работа с литературой, включая ЭБС, источниками в сети Internet Подготовка к лабораторному практикуму, подготовка отчета по практикуму	Отчет по лабораторному практикуму
Тема 5. Неструктурированная информация на примере корпуса текстов	Лемматизация слов. Кластеризация текстов.	Работа с литературой, включая ЭБС, источниками в сети Internet Подготовка к лабораторному практикуму, подготовка отчета по практикуму	Отчет по лабораторному практикуму

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ И УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Перечень основной и дополнительной литературы

Основная литература:

1. Душин, В.К. Теоретические основы информационных процессов и систем : учебник : [16+] / В.К. Душин. – 5-е изд. – Москва : Дашков и К°, 2018. – 348 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/

Дополнительная литература:

1. Информационные системы и технологии управления : учебник / ред. Г.А. Титоренко. – 3-е изд., перераб. и доп. – Москва : Юнити, 2015. – 591 с. : ил., табл., схемы – (Золотой фонд российских учебников). – Режим доступа: по подписке. –

URL: https://biblioclub.ru/.

2. Балдин, К. В. Информационные системы в экономике : учебник / К. В. Балдин, В. Б. Уткин. – 9-е изд., стер. – Москва : Дашков и К°, 2021. – 395 с. : ил., табл. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=684194 – Библиогр. в кн. – ISBN 978-5-394-04038-2. – Текст : электронный.

3. Матяш, С.А. Корпоративные информационные системы : учебное пособие : [16+] / С.А. Матяш. – Москва ; Берлин : Директ-Медиа, 2015. – 471 с. : ил., схем., табл. – Режим доступа: по подписке. – URL: https://biblioclub.ru/

6.2. Перечень ресурсов информационно-коммуникационной сети «Интернет»

№	Наименование ресурса	Ссылка
1.	Сайт вопросов и ответов для программистов	https://ru.stackoverflow.com/
2.	Анализ данных	https://academy.yandex.ru/posts/chem-zanimayutsya-analitiki-dannykh-i-kak-nachat-rabotat-v-etoy-oblasti
3.	Что нужно знать начинающему дата-аналитику	https://vc.ru/books/71290-mozhno-bez-opyta-chto-nuzhno-znat-nachinayushchemu-data-analitiku

6.3. Описание материально-технической базы

Материально-техническое обеспечение дисциплины включает в себя:

Учебная аудитория (Лаборатория информационно-коммуникационных технологий), оборудованная:

комплекты специализированной учебной мебели, мультимедийный проектор, экран, доска классная, принтер, компьютер преподавателя и компьютеры обучающихся с выходом в сеть «Интернет», доступом в электронную информационно-образовательную среду.

Помещение для самостоятельной работы обучающихся – аудитория, оборудованная:

комплекты специализированной учебной мебели, мультимедийный проектор, экран, доска классная, компьютеры с выходом в сеть «Интернет» и доступом в электронную информационно-образовательную среду.

6.4. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, в том числе комплект лицензионного программного обеспечения, электронно-библиотечные системы, современные профессиональные базы данных и информационные справочные системы

Обучающиеся обеспечены доступом к электронной информационно-

образовательной среде из любой точки, в которой имеется доступ к сети «Интернет», как на территории организации, так и вне ее.

лицензионное программное обеспечение:

- Windows (зарубежное, возмездное);
- MS Office (зарубежное, возмездное);
- Adobe Acrobat Reader (зарубежное, свободно распространяемое);
- КонсультантПлюс: «КонсультантПлюс: Студент» (российское, свободно распространяемое);
- 7-zip – архиватор (зарубежное, свободно распространяемое);
- Comodo Internet Security (зарубежное, свободно распространяемое);
- MySQL for Windows – реляционная система управления базами данных (зарубежное, свободно распространяемое);
- Apache NetBeans – свободная интегрированная среда разработки приложений (IDE) на языках программирования Java, Python, PHP, JavaScript, C, C++, Ада и ряда других (зарубежное, свободно распространяемое);
- Android Studio – разработка мобильных приложений (зарубежное, свободно распространяемое)

электронно-библиотечная система:

- Электронная библиотечная система (ЭБС) «Университетская библиотека ONLINE» http://biblioclub.ru/.
- Образовательная платформа «Юрайт». Для вузов и ссузов. Электронная библиотечная система (ЭБС) https://urait.ru/

современные профессиональные баз данных:

- Официальный интернет-портал базы данных правовой информации http://pravo.gov.ru.
- Портал Единое окно доступа к образовательным ресурсам http://window.edu.ru/

информационные справочные системы:

- Портал Федеральных государственных образовательных стандартов высшего образования http://fgosvo.ru.
- Компьютерная справочная правовая система «КонсультантПлюс» (http://www.consultant.ru/).

7. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ

7.1. Описание оценочных средств для проведения текущего контроля успеваемости в процессе освоения дисциплины

№ п/п	Форма учебного занятия, по которому проводится ТКУ	Шкала и критерии оценки, балл
1.	Лабораторный практикум	20-18 – задание практикума и отчет выполнены полностью, корректно оформлен отчет, даны исчерпывающие ответы на дополнительные вопросы; 17-12 – задание практикума выполнено практически полностью, отчет оформлен с небольшими ошибками, даны ответы на дополнительные вопросы; 11-5– задание практикума выполнено частично, отчет оформлен с ошибками, даны ответы на некоторые дополнительные вопросы; 0 – задание практикума не выполнено, отчет не оформлен, не даны ответы на дополнительные вопросы.

Типовые контрольные задания или иные материалы в рамках текущего контроля успеваемости

<u>Типовые задания к лабораторным практикумам</u>

Лабораторный практикум 1.

Способы подготовки и отображения данных в R. Возможности ввода/вывода.

Лабораторный практикум 2.

Решение задач на больших графах.

Лабораторный практикум 3.

Способы анализа данных в R. Получение первичных элементарных характеристик о наборах данных (элементарные статистики).Способы импорта/экспорта данных.

Лабораторный практикум 4.

Решение задач Data Mining. Задачи классификации, кластеризации: деревья решений, RandomForest, k-means.

Лабораторный практикум 5.

Развертывание локального кластера Hadoop. Подсчѐт слов в тексте, с помощью MapReduce.

Лабораторный практикум 6.

Составление спецификации ПО (Software Requirement Specification – SRS). Информация о рамках и стандартах, которые ограничивают опции разработчика при создании системы.

Деанонимизация пользователя путем сравнения постов на 2ch.ru и комментариев под роликами в сети YouTube.

7.2. Описание оценочных средств для проведения промежуточной аттестации

Промежуточная аттестация по дисциплине проводится в форме экзамена.

Процедура оценивания	Шкала и критерии оценки, балл
Экзамен представляет собой выполнение обучающимся заданий билета, включающего в себя: Задание №1 – теоретический вопрос на знание базовых понятий предметной области дисциплины, а также позволяющий оценить степень владения обучающимся принципами предметной области дисциплины, понимание их особенностей и взаимосвязи между ними; Задание №2 – задание на анализ ситуации из предметной области дисциплины и выявление способности обучающегося выбирать и применять соответствующие принципы и методы решения практических проблем, близких к профессиональной деятельности; Задание №3 – задание на проверку умений и навыков, полученных в результате освоения дисциплины	Выполнение обучающимся заданий оценивается по следующей балльной шкале: Задание 1: 0-30 баллов Задание 2: 0-30 баллов Задание 3: 0-40 баллов **-90 и более (отлично)** – ответ правильный, логически выстроен, приведены необходимые формулы, использована профессиональная лексика. Задача решена правильно. Обучающийся правильно интерпретирует полученный результат. **-70 и более (хорошо)–** ответ в целом правильный, логически выстроен, приведены необходимые формулы, использована профессиональная лексика. Ход решения задачи правильный, ответ неверный. Обучающийся в целом правильно интерпретирует полученный результат. **-50 и более (удовлетворительно)–** ответ в основном правильный, логически выстроен, приведены не все необходимые формулы, использована профессиональная лексика. Задача решена частично. **-Менее 50 (неудовлетворительно)**– ответы на теоретическую часть неправильные или неполные. Задача не решена

Типовые задания для проведения промежуточной аттестации обучающихся

Задания на знания

1. Факторный анализ.
2. Дискриминантный анализ.
3. Кластерный анализ.
4. Многомерное шкалирование.
5. Методы контроля качества.
6. Основные направления развития методов обработки и хранения данных.
7. Volume.
8. Закон Мура.
9. Velocity. Variety.
10. Фреймворк Hadoop.

11. Проблема хранения неструктурированных данных.
12. Проблема преобразования данных.
13. Семантические анализаторы.
14. Самообучающиеся автоматы.
15. Языки для Big Data: R.
16. Языки для Big Data: Python.
17. Языки для Big Data: Julia.
18. Аналитика Big Data — реалии и перспективы в России и мире.
19. Data Mining.
20. Краудсорсинг.
21. Смешение и интеграция данных.
22. Статистический анализ.
23. Визуализация аналитических данных.
24. Big data: применение и возможности.
25. Решения на основе Big data.
26. Рынок Big data в России.
27. Big data в банках.
28. Big data в бизнесе.
29. Big data в маркетинге.
30. Имитационное моделирование.
31. Пространственный анализ.
32. Статистический анализ.

Задания на умения

1. Определение больших данных, ключевые характеристики. Примеры задач больших данных. Основные виды данных.

2. Дать краткую сравнительную характеристику инструментария ПО для анализа данных.

3. Охарактеризовать конструкции языка R

4. Перечислить типы языка R, привести примеры.

5. Роль аналитика по данным (Data Scientist).

6. Ключевые компетенции аналитика.

7. Отличия BI от Data Science.

8. «Жизненный цикл» проекта по аналитике больших данных.

9. Типовая архитектура проекта в области больших данных. Перечислить используемые технологии, указать степень вовлеченности каждой из технологий на каждом этапе работы над проектом.

10. Перечислить основные роли исполнителей проекта в области больших данных.

11. Что такое Data Mining? Основные задачи и методы Data Mining. Этапы интеллектуального анализа данных.

12. Методы интеллектуального анализа данных.

13. Что такое ИИ? Декатлон?

14. Роль гипотез в процессе познания. Какие факторы используются для

уточнения гипотез?

15. Основные понятия статистики и дескриптивный анализ:

16. Шкалы измерений. Генеральная совокупность и выборка.

17. Нормальное распределение. Уровень статистической достоверности.

18. Корреляция и регрессионный анализ. Коэффициент корреляции. Графическое представление.

19. Постановка задачи регрессионного анализа.

20. Пояснить термин "Линейная регрессия".

21. Привести примеры использования регрессионного анализа.

22. Классификация и кластеризация – суть и назначение. Метрики.

23. Постановка задачи кластеризации.

24. Методы кластеризации на графах. Отличие от задачи классификации. Привести примеры использования алгоритмов кластеризации.

25. Парадигма Map Reduce. Описать принцип работы. Нарисовать схему.

26. Перечислить слабые и сильные стороны. Обозначить области применимости. Привести примеры использования.

27. Визуализация. Дать определение визуализации. Показать важность визуализации в аналитике больших данных. Привести примеры и инструменты для визуализации.

28. Научные проблемы больших данных. Показать значимость проблем, актуальность, связь с областями математики и инженерии.

29. OLAP и OLTP системы. Разница.

30. Репликация и шардинг.

31. Требования ACID. CAP-теорема, BASE архитектура

32. NoSql. Классификация NoSql хранилищ. Их особенности. Примеры распределенных хранилищ.

Задания на навыки

Задание 1.

Дан набор данных заданной структуры и программа SAS Data step, производящая определенную обработку и вычисления с использованием данного набора. Перепишите эту программу на SAS DS2 с использованием параллельных нитей и созданием пользовательского пакета, чтобы результат обработки сохранился тем же, но код мог выполняться в параллельной среде.

Задание 2

Дан набор заданной структуры, постройте модель прогнозирования отклика с использованием процедуры impstat с алгоритмом random forest с заданным числом деревьев. Примените полученную модель к тестовому набору данных той же структуры, визуализируйте полученный график Lift. Постройте на том же наборе модель с использованием высокопроизводительной версии метода GLM. Примените к тестовому набору. Сравните результаты GLM и Random Forest по AUC.

Задание 3

Дан текстовый корпус документов, лежащих в указанной директории.Создайте в SAS Tex Miner проект, который: выберет файлы с расширением pdf; осуществит парсинг набора с определением частей речи и сохранением в признаковом пространстве только существительных и глаголов; осуществит фильтрацию документов и признаков с использованием заданной схемой определения весов лексем (например, на основе tf-idf); выделит заданное количество ключевых тематик по методу SVD. В ответе укажите топ 5 ключевых слов во второй выявленной тематике. Какой документ имеет наибольший вес в это тематике?

Задание 4.

Дана таблица, табулирующая кривую спроса на сотовый телефон: в первом столбце – цена телефона от 1000р до 140 000р с шагом в 1000р, во втором количество человек в месяц, готовых покупать этот телефон по данной цене. Кривая спроса имеет единственный максимум, который мы назовем «комфортной ценой покупателя». Докажите, что при условии гладкости кривой спроса «комфортная цена продавца», т.е. та, по которой телефон продавать наиболее выгодно, всегда больше «комфортной цены покупателя».

Задание 5.

Имеется беспилотник массой 1500г, на которой можно навесить любое число однотипных двигателей из некоторого прайс-листа (прайс-лист в формате *.xlsx прилагается). В прайс-листе указаны цена каждого двигателя и его тяга. Найдите наиболее дешевую конфигурацию беспилотника.

Задание 6.

В меню пиццерии имеется 12-дюймовая пицца ценой $10 и 24-дюймовая пицца по $30. Верно ли, что можно сэкономить средства, купив две 12-дюймовых пиццы и заплатив $20?