МИНОБРНАУКИ РОССИИ

Бугульминский филиал

Федерального государственного бюджетного образовательного учреждения высшего профессионального образования

«Казанский национальный исследовательский технологический университет»

Хакимова А.А.

**БОЛЬШИЕ ДАННЫЕ**

**Контрольная работа**

**6 семестр**

*Для бакалавриатов направления 09.03.02 «Информационные системы и технологии» заочной формы обучения*

Бугульма, 2023

# 1 ОБЩИЕ ТРЕБОВАНИЯ К СОДЕРЖАНИЮ И ОФОРМЛЕНИЮ КОНТРОЛЬНОЙ РАБОТЫ

Контрольная работа состоит из двух теоретических вопросов и тестовых заданий, выполняется по одному из предложенных вариантов. Номера вопросов определяются по двум последним цифрам зачетной книжки. Контрольная работа, выполненная не по своему варианту, не рецензируется и не допускается к защите.

Контрольная работа включает следующие структурные элементы:

1. Титульный лист.
2. Ответы на два теоретических вопроса. Ответ на каждый вопрос должен составлять по объему 3 – 4 машинописные страницы. Контрольная работа оформляется на листах формата А4 шрифтом Times New Roman, 14, межстрочный интервал – одинарный, красная строка – 1,25см. Поля на странице: верхнее – 2,5см, нижнее – 2см, левое – 3см, правое – 1,5см.
3. Выполнение тестовых заданий

# 2 ВАРИАНТЫ КОНТРОЛЬНОЙ РАБОТЫ

Варианты контрольной работы приведены в таблице 1. Вариант определяется на основании последних двух цифр номера зачетной книжки студента. Предпоследняя цифра показывает номер строки таблицы, последняя – номер столбца таблицы.

Таблица 1 – Варианты контрольной работы

|  |  |
| --- | --- |
|  | Последняя цифра зачетной книжки |
| **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Предпоследняя цифра зачетной книжки | **0** | В.1, 21 | В. 2, 22 | В. 3,23 | В. 4,24 | В.5,25 | В.6,26 | В.7,27 | В.8,28 | В.9,29 | В.10,30 |
| **1** | В. 11,31 | В.12,32 | В.13,33 | В.14,34 | В.15,35 | В.16,36 | В.17,37 | В.18,38 | В.19,39 | В.20,40 |
| **2** | В.1, 21 | В. 2, 22 | В. 3,23 | В. 4,24 | В.5,25 | В.6,26 | В.7,27 | В.8,28 | В.9,29 | В.10,30 |
| **3** | В. 11,31 | В.12,32 | В.13,33 | В.14,34 | В.15,35 | В.16,36 | В.17,37 | В.18,38 | В.19,39 | В.20,40 |
| **4** | В.1, 21 | В. 2, 22 | В. 3,23 | В. 4,24 | В.5,25 | В.6,26 | В.7,27 | В.8,28 | В.9,29 | В.10,30 |
| **5** | В. 11,31 | В.12,32 | В.13,33 | В.14,34 | В.15,35 | В.16,36 | В.17,37 | В.18,38 | В.19,39 | В.20,40 |
| **6** | В.1, 21 | В. 2, 22 | В. 3,23 | В. 4,24 | В.5,25 | В.6,26 | В.7,27 | В.8,28 | В.9,29 | В.10,30 |
| **7** | В. 11,31 | В.12,32 | В.13,33 | В.14,34 | В.15,35 | В.16,36 | В.17,37 | В.18,38 | В.19,39 | В.20,40 |
| **8** | В.1, 21 | В. 2, 22 | В. 3,23 | В. 4,24 | В.5,25 | В.6,26 | В.7,27 | В.8,28 | В.9,29 | В.10,30 |
| **9** | В. 11,31 | В.12,32 | В.13,33 | В.14,34 | В.15,35 | В.16,36 | В.17,37 | В.18,38 | В.19,39 | В.20,40 |

# 3 ТЕОРЕТИЧЕСКИЕ ВОПРОСЫ

1. Что такое большие данные и как они отличаются от обычных данных?

2. Каковы основные шаги в жизненном цикле анализа больших данных?

3. Какие инструменты используются для выполнения каждого этапа жизненного цикла анализа больших данных?

4. Что такое корреляция и как она используется в анализе больших данных?

5. Какие методы регрессионного анализа могут быть применены к большим данным?

6. Каковы основные шаги в задачах классификации данных?

7. Какие методы используются для выполнения задачи классификации больших данных?

8. Что такое кластеризация данных и как она применяется к большим данным?

9. Какие алгоритмы кластеризации можно использовать для анализа больших данных?

10. Какие методы и инструменты можно использовать для подготовки данных перед анализом?

11. Что такое парадигма Map Reduce и как она применяется в анализе больших данных?

12. Какие преимущества предоставляет парадигма Map Reduce при работе с большими данными?

13. Какой процесс выполняется при выполнении Map функции?

14. Какой процесс выполняется при выполнении Reduce функции?

15. Какие типы данных можно обрабатывать с использованием парадигмы Map Reduce?

16. Каким образом масштабируется выполнение задач с использованием парадигмы Map Reduce?

17. Какие инструменты и языки программирования поддерживают парадигму Map Reduce?

18. Как организовать параллельную обработку данных с использованием парадигмы Map Reduce?

19. Какие проблемы могут возникнуть при работе с большими данными, и как их можно решить?

20. Какие методы используются для управления и хранения больших объемов данных?

21. Как оценить качество и достоверность результатов в анализе больших данных?

2. Какие методы используются для визуализации и интерпретации результатов анализа больших данных?

23. Каким образом можно оптимизировать производительность обработки больших данных?

4. Какие метрики могут быть использованы для оценки эффективности алгоритмов анализа больших данных?

25. Какие методы можно использовать для обнаружения выбросов и аномалий в больших данных?

26. Какие техники используются для поиска и извлечения информации из больших наборов данных?

27. Какие методы машинного обучения можно применить к анализу больших данных?

28. Какие алгоритмы машинного обучения наиболее популярны для задач анализа больших данных?

29. Какие методы машинного обучения подходят для задач классификации больших данных?

30. Какие методы машинного обучения подходят для задач кластеризации больших данных?

31. Как оценить качество классификации или кластеризации в больших данных?

32. Какие методы можно использовать для обработки текстовых данных в больших объемах?

33. Какие алгоритмы нейронных сетей могут быть применены для анализа больших данных?

34. Какие методы можно использовать для обнаружения ассоциативных правил в больших данных?

35. Как выполняется препроцессинг данных перед применением алгоритмов анализа в больших данных?

36. Как обработать пропущенные значения в больших наборах данных?

37. Какие методы можно использовать для снижения размерности данных перед анализом?

38. Каким образом можно обеспечить безопасность и конфиденциальность данных при работе с большими объемами?

39. Какие методы можно использовать для ранжирования и рекомендаций в больших данных?

40. Какие исследования и тенденции есть в области анализа больших данных?

# 4 ТЕСТОВЫЕ ЗАДАНИЯ

Тема: Жизненный цикл анализа больших данных

1. Что такое жизненный цикл анализа больших данных?

а) Процесс сбора, обработки, анализа и интерпретации больших объемов данных

б) Периодическое обновление и модификация алгоритмов анализа больших данных

в) Процесс разработки инструментов для анализа больших данных

2. Какой этап жизненного цикла анализа данных отвечает за подготовку данных для анализа?

а) Сбор данных

б) Чистка данных

 в) Анализ данных

3. Какой этап жизненного цикла анализа данных отвечает за разработку моделей и алгоритмов анализа?

а) Сбор данных

б) Чистка данных

в) Анализ данных

4. Какой этап жизненного цикла анализа данных отвечает за интерпретацию и использование полученных результатов?

а) Сбор данных

б) Чистка данных

в) Анализ данных

5. Какой из перечисленных методов может использоваться для обработки и анализа больших данных?

а) Алгоритмы машинного обучения

 б) Методы статистического анализа

в) Оба варианта

6. Какой из перечисленных инструментов является примером для обработки и анализа больших данных?

а) Hadoop

б) Microsoft Excel

в) Photoshop

7. Что такое "большие данные"?

а) Данные, объем которых превышает возможности обычных баз данных

б) Данные, которые сложно обработать и анализировать

в) Данные, которые содержат личную информацию

8. Какие основные задачи решает анализ больших данных?

а) Поиск закономерностей и трендов в данных

б) Оптимизация производственных процессов

в) Разработка новых продуктов и услуг

9. Какие проблемы могут возникнуть при анализе больших данных?

а) Недостаток вычислительных ресурсов

б) Неправильная интерпретация результатов

в) Оба варианта

10. Какие типы данных могут быть использованы при анализе больших данных?

а) Числовые данные

б) Текстовые данные

в) Графические данные

 г) Все варианты

11. Какие этапы включает жизненный цикл анализа больших данных?

а) Сбор данных, чистка данных, анализ данных, интерпретация результатов

б) Сбор данных, обработка данных, визуализация данных

в) Разработка моделей, алгоритмов, интерпретация результатов

12. Что такое "глубокое обучение" (deep learning)?

а) Метод машинного обучения, использующий искусственные нейронные сети

б) Метод анализа данных, основанный на статистических методах

в) Метод обработки текстов и изображений

3. Какие источники данных могут быть включены в процесс сбора данных?

а) Сенсоры IoT (интернет вещей)

б) Социальные сети и онлайн-платформы

в) Базы данных предприятия

14. Какая роль имеет визуализация данных в анализе больших данных?

а) Позволяет лучше понять и интерпретировать результаты анализа

б) Отсутствует роль визуализации данных в анализе больших данных

 в) Упрощает процесс сбора данных

15. Какие методы может использовать аналитик данных для обнаружения аномалий?

а) Статистические методы

б) Методы машинного обучения

в) Оба варианта

16. Какие ограничения могут быть связаны с анализом больших данных?

а) Нехватка квалифицированных специалистов

б) Сложность обработки и хранения больших объемов данных

в) Оба варианта

17. Какие методы используются для хранения и управления большими данными?

а) Хранение в распределенных системах

б) Использование облачных сервисов

в) Базы данных

18. Какие преимущества может иметь использование анализа больших данных?

а) Получение новых знаний и закономерностей

б) Улучшение работоспособности и производительности предприятия

в) Принятие более обоснованных и эффективных решений

19. Какой роль может играть машинное обучение в анализе больших данных?

а) Автоматизация процесса анализа и выявления закономерностей

б) Улучшение процедуры сбора и хранения больших объемов данных

в) Разработка новых моделей и алгоритмов анализа

20. Какие вызовы могут возникнуть при внедрении анализа больших данных в организацию?

а) Необходимость обучения сотрудников новым навыкам и знаниям

б) Сложность интеграции существующей инфраструктуры и систем

в) Оба варианта

Тема: Корреляция и регрессия

1. Что такое корреляция?

а) Статистическая взаимосвязь между двумя переменными

б) Изменение зависимой переменной при изменении независимой

в) Отрицательное влияние одной переменной на другую

г) Мощность статистического теста

2. Регрессионная модель предназначена для:

а) Прогнозирования будущих значений зависимой переменной

б) Описания статистической связи между двумя переменными

в) Оценки силы связи между переменными

г) Проверки гипотезы о различии средних значений

3. Коэффициент детерминации обозначает:

а) Процент изменчивости зависимой переменной, объясненный моделью регрессии

б) Процент изменчивости зависимой переменной, обусловленный другими факторами

в) Степень линейной зависимости между двумя переменными

г) Различие средних значений двух групп

4. Что такое множественная регрессия?

а) Регрессия с большим количеством независимых переменных

б) Регрессия, основанная на больших данных

в) Регрессия, проводимая с помощью компьютерных программ

г) Регрессия, в которой зависимая переменная принимает несколько значений

5. Как можно оценить значимость коэффициента корреляции?

а) Сравнить его со среднеквадратической ошибкой

б) Провести t-тест на значимость

в) Посчитать процент изменчивости зависимой переменной

г) Сравнить его с коэффициентом детерминации

6. В каком диапазоне может находиться коэффициент корреляции Пирсона?

а) От -1 до 0

б) От 0 до 1

в) От -∞ до +∞

г) От 0 до 100

7. Что означает коэффициент корреляции 0.8?

а) Очень высокая связь между переменными

б) Очень слабая связь между переменными

в) Отсутствие связи между переменными

г) Невозможно сказать без дополнительной информации

8. Что такое автокорреляция?

а) Связь между переменными в разных периодах времени

б) Ситуация, когда одна переменная является причиной другой

в) Случайное изменение зависимой переменной

г) Связь между независимыми переменными

9. В чем отличие между простой и множественной линейной регрессией?

а) Простая регрессия не учитывает другие факторы, множественная учитывает их

б) Простая регрессия не может иметь отрицательных коэффициентов, множественная может

в) Простая регрессия предсказывает только одну зависимую переменную, множественная неограничена в количестве

г) Нет отличий, это синонимы

10. Что такое модель линейной регрессии?

а) Математическое выражение, описывающее линейную зависимость между переменными

б) Графическое представление данных

в) Методология проведения статистического анализа

 г) Конкретная выборка данных

11. Как можно провести проверку на гетероскедастичность в регрессионной модели?

 а) Построить диаграмму рассеяния остатков

б) Проверить значимость коэффициентов регрессионной модели

в) Сравнить модель регрессии с альтернативными моделями

г) Проверить нормальность распределения остатков

12. Какова основная цель регрессии?

а) Описать влияние одной переменной на другую

б) Прогнозировать значения зависимой переменной

в) Определить причинно-следственную связь между переменными

г) Вывести статистически значимые результаты

13. Что такое многомерная регрессия?

а) Регрессия, в которой зависимая переменная принимает много разных значений

б) Регрессия с большим количеством независимых переменных

в) Регрессия, основанная на больших объемах данных

г) Возможны все вышеперечисленные варианты

14. Что означает нулевая корреляция между переменными?

а) Все значения переменной равны нулю

б) Переменные полностью независимы друг от друга

в) Одна переменная полностью определяется другой

г) Корреляция не может быть нулевой

15. Что такое коэффициент корреляции Спирмена?

а) Коэффициент, показывающий линейную связь между переменными

б) Коэффициент, показывающий силу монотонной связи между переменными

в) Коэффициент, показывающий силу связи между номинативными переменными

г) Коэффициент, показывающий зависимость переменной от времени

16. Какова основная цель корреляционного анализа?

а) Определить статистическую значимость взаимосвязи между переменными

б) Построить математическую модель, описывающую взаимосвязь

в) Выбрать оптимальные независимые переменные для регрессии

г) Проверить гипотезу о равенстве средних значений двух групп

17. Что такое регрессионные коэффициенты?

а) Статистические параметры, оценивающие влияние независимых переменных на зависимую

б) Стандартные ошибки оценок регрессионных коэффициентов

в) Диапазон значений, которые может принимать зависимая переменная

г) Количество независимых переменных в регрессионной модели

18. Как найти линейную регрессию в программе Excel?

а) Использовать функцию "Линейная регрессия"

б) Использовать функцию "SUM"

в) Использовать функцию "Совершенное сравнение"

г) Нет такой функции в программе Excel

19. Каково значение коэффициента детерминации, если степень связи между переменными равна 0?

а) 0

б) 1

в) -1

г) 0.5

20. Что такое мультиколлинеарность в множественной регрессии?

а) Ситуация, когда две независимые переменные сильно коррелируют между собой

б) Ситуация, когда зависимая переменная имеет множество значений

в) Ситуация, когда нет статистически значимых результатов регрессии

г) Ситуация, когда значения переменных изменяются во времени

Тема: Классификация и кластеризация данных

1. Что такое классификация данных?

а) Процесс разделения данных на заданные классы или категории

б) Анализ данных с помощью статистических методов

в) Обработка больших объемов данных

г) Определение кластеров в данных

2. Какая задача решается при классификации данных?

а) Прогнозирование значений зависимой переменной

б) Поиск аномалий в данных

в) Выделение групп схожих объектов

г) Определение вероятности принадлежности объекта к определенному классу

3. Какие методы классификации данных вы знаете?

а) Логистическая регрессия

б) Дерево решений

в) Метод опорных векторов

г) Все вышеперечисленные методы

4. Что такое кластеризация данных?

а) Процесс разделения данных на группы схожих объектов

б) Анализ данных с помощью математических методов

в) Фильтрация данных для улучшения качества информации

г) Разработка алгоритмов для сжатия данных

5. Для каких целей применяется кластеризация данных?

а) Выявление скрытых закономерностей в данных

б) Классификация объектов по заданным признакам

в) Минимизация размерности данных

г) Построение математических моделей для прогнозирования

6. Какие методы кластеризации данных вы знаете?

а) K-средних

б) Иерархическая кластеризация

в) DBSCAN

г) Все вышеперечисленные методы

7. Что означает понятие "K" в методе K-средних?

 а) Количество групп, на которое разбиваются данные

б) Количество признаков, по которым проводится кластеризация

в) Координаты центроидов, определяющие кластеры

г) Коэффициент корреляции между переменными

8. Как выбрать оптимальное количество кластеров при кластеризации данных методом локтя?

а) Построить график зависимости внутрикластерного расстояния от числа кластеров

б) Построить график зависимости межкластерного расстояния от числа кластеров

в) Выбрать количество кластеров, при котором график "перегибается" или "ломается"

г) Нет определенного метода выбора оптимального количества кластеров

9. Какие метрики можно использовать для оценки качества кластеризации?

 а) Силуэт

б) Давидовского-Болдуина

в) Коэффициент Жаккара

г) Все вышеперечисленные метрики

10. Какой метод классификации можно использовать для предсказания категориальной переменной?

а) Логистическая регрессия

б) K-ближайших соседей

в) Решающее дерево

г) Все вышеперечисленные методы

11. Что такое переобучение модели в контексте классификации и кластеризации данных?

а) Ситуация, когда модель слишком точно подстраивается под обучающую выборку и плохо обобщает новые данные

б) Процесс тренировки модели на больших объемах данных

в) Ситуация, когда модель неправильно классифицирует объекты

г) Ситуация, когда модель не может обработать большие объемы данных

12. Что такое регуляризация в контексте классификации и кластеризации данных?

а) Метод, уменьшающий влияние некоторых признаков на модель для предотвращения переобучения

б) Процесс выделения групп схожих объектов в данных

в) Построение алгоритма на основе предобученной модели

г) Вычисление стандартных ошибок оценок параметров модели

13. Что такое метод опорных векторов (SVM)?

а) Метод классификации, использующий гиперплоскости для разделения данных на классы

б) Метод поиска аномалий и выбросов в данных

в) Метод сжатия и упорядочивания данных

г) Метод, основанный на использовании ближайших соседей для классификации объектов

14. Что такое решающее дерево?

а) Метод классификации, основанный на построении древовидной структуры

б) Метод описания и аппроксимации данных с помощью дерева

в) Метод выделения групп схожих объектов на основе расстояний между ними

г) Метод оптимизации критерия информативности при разделении данных

15. Что такое случайный лес?

а) Ансамбль решающих деревьев

б) Метод оценки вероятности принадлежности объекта к определенному классу

в) Графическое представление структуры данных

г) Метод, использующий случайные числа для выделения групп схожих объектов

16. Что такое метод K-средних?

а) Метод кластеризации, основанный на поиске центроидов для групп объектов

б) Метод классификации, основанный на K ближайших соседях

в) Метод построения линейной модели для предсказания зависимой переменной

 г) Метод, использующий ближайших соседей для выделения групп схожих объектов

17. Что такое метод DBSCAN?

а) Метод кластеризации, основанный на плотности и соседстве объектов

б) Метод классификации, основанный на определении кратчайшего пути между объектами

в) Метод, использующий нейронные сети для анализа данных

г) Метод, основанный на принципе максимума правдоподобия

18. Как выбрать оптимальные гиперпараметры для модели машинного обучения?

а) Использовать метод перебора всех возможных значений гиперпараметров

б) Применить кросс-валидацию для оценки качества модели с различными значениями гиперпараметров

в) Оценить модель на отложенной выборке и выбрать те гиперпараметры, которые дают наилучший результат

г) Все вышеперечисленные методы могут быть использованы

19. Что такое выбросы в данных?

а) Аномальные значения, отклоняющиеся от остальных

б) Пропущенные значения в данных

в) Завышенная размерность данных

г) Коррелирующие переменные

20. Какая функция используется для измерения близости объектов при кластеризации?

а) Евклидово расстояние

б) Манхэттенское расстояние

в) Косинусное расстояние

г) Все вышеперечисленные функции

Тема: Подготовка данных

1. Что такое подготовка данных?

а) Процесс очистки, преобразования и объединения данных перед анализом

б) Статистический анализ данных

в) Хранение и сохранение данных

г) Визуализация данных

2. Какой известный принцип подготовки данных гласит "Мусор в, мусор вы"?

а) Принцип GIGO (Garbage In, Garbage Out)

б) Принцип ETL (Extract, Transform, Load)

 в) Принцип CRUD (Create, Read, Update, Delete)

г) Принцип MVP (Minimum Viable Product)

3. Какой этап подготовки данных отвечает за удаление несущественных или поврежденных данных?

а) Очистка данных

б) Интеграция данных

в) Трансформация данных

 г) Загрузка данных

4. Что такое пропущенные значения в данных?

а) Отсутствующие значения или пустые ячейки в таблице данных

б) Значения, которые выбиваются за пределы диапазона

в) Значения, которые сильно коррелируют между собой

г) Значения, которые имеют нестандартный формат

5. Какой метод можно использовать для заполнения пропущенных значений в данных?

а) Удаление строк или столбцов с пропущенными значениями

б) Замена пропущенных значений средним или медианой

в) Создание нового признака на основе имеющихся

г) Все вышеперечисленные методы

6. Что такое выбросы в данных?

а) Экстремальные значения, которые сильно отклоняются от остальных

б) Значения, которые несут вред при анализе данных

в) Значения, которые имеют низкую корреляцию с другими переменными

г) Ошибки в данных, которые мешают проведению анализа

7. Какой метод можно использовать для обработки выбросов в данных?

а) Удаление строк или столбцов с выбросами

б) Замена выбросов на среднее или медиану

в) Логарифмическое преобразование данных

г) Все вышеперечисленные методы

8. Что такое масштабирование данных?

а) Преобразование данных в заданный диапазон значений

б) Проверка данных на наличие ошибок

в) Создание новых переменных на основе имеющихся

г) Объединение нескольких таблиц данных

9. Какой метод масштабирования можно использовать для непрерывных числовых данных?

а) Min-Max масштабирование

б) Стандартизация

в) Нормализация

г) Все вышеперечисленные методы

10. Что такое кодирование категориальных переменных?

а) Преобразование категориальных переменных в числовой формат

б) Проверка наличия выбросов в категориальных переменных

в) Заполнение пропущенных значений в категориальных переменных

г) Удаление категориальных переменных из данных

11. Какой метод кодирования категориальных переменных можно использовать для переменных без порядка?

а) One-Hot Encoding

б) Label Encoding

в) Ordinal Encoding

г) Все вышеперечисленные методы

12. Что такое объединение данных?

а) Процесс слияния нескольких таблиц данных в одну общую

б) Преобразование данных в стандартный формат

в) Удаление повторяющихся строк или столбцов в данных

г) Проверка наличия ошибок и пропущенных значений в данных

13. Что такое агрегация данных?

а) Процесс группировки данных по определенным признакам и вычисления агрегированных статистик

б) Проверка соответствия данных заданным критериям

в) Преобразование текстовых данных в числовой формат

г) Разделение данных на обучающую и тестовую выборки

14. Что такое отбор признаков?

а) Процесс выбора наиболее значимых признаков для анализа или моделирования

б) Преобразование данных в бинарный формат

в) Деление данных на кластеры в зависимости от признаков

г) Создание новых признаков на основе имеющихся

15. Какой метод можно использовать для отбора признаков?

а) Корреляционный анализ

б) Важность признаков в модели машинного обучения

в) Рекурсивное устранение признаков

г) Все вышеперечисленные методы

16. Что такое шумовые признаки в данных?

а) Признаки, которые не несут полезной информации для анализа

б) Признаки, которые имеют большую корреляцию с зависимой переменной

в) Признаки, которые имеют большую дисперсию по сравнению с другими признаками

г) Признаки, которые не могут быть преобразованы в числовой формат

17. Какой метод можно использовать для обработки шумовых признаков в данных?

а) Удаление шумовых признаков из данных

б) Преобразование шумовых признаков в бинарный формат

в) Объединение шумовых признаков в один признак

г) Пропуск шумовых признаков при анализе данных

18. Что такое дубликаты данных?

а) Повторяющиеся строки или столбцы в таблице данных

б) Значения, которые выбиваются из общего тренда данных

в) Данные, которые состоят из одного и того же значения

г) Ошибки в данных, вызванные некорректным вводом информации

19. Какой метод можно использовать для обнаружения и удаления дубликатов данных?

а) Проверка уникальности значений в таблице данных

б) Удаление повторяющихся строк в таблице данных

в) Объединение дубликатов в одну строку или столбец

г) Все вышеперечисленные методы

20. Что означает термин "feature scaling" в контексте подготовки данных?

а) Масштабирование признаков для обеспечения одинаковой важности в модели

б) Отбор наиболее важных признаков для анализа

в) Генерация новых признаков на основе имеющихся

г) Удаление шумовых признаков из данных

Тема: Парадигма Map Reduce

1. Что такое парадигма MapReduce?

а) Методология обработки больших объемов данных, основанная на параллельном выполнении операций map и reduce

б) Алгоритм для поиска кратчайшего пути в графе

в) Техника преобразования данных в графическое представление

г) Распределенная система хранения данных

2. Какие основные операции включает в себя парадигма MapReduce?

а) Операцию map и операцию reduce

б) Операцию select и операцию join

в) Операцию sort и операцию aggregate

г) Операцию insert и операцию update

3. Что происходит на этапе операции map в парадигме MapReduce?

а) Преобразование и фильтрация данных

б) Соединение данных из разных источников

в) Группировка данных по заданному признаку

г) Вычисление агрегированных статистик

4. Что происходит на этапе операции reduce в парадигме MapReduce?

а) Объединение и сжатие данных

б) Сортировка данных по заданному значению

в) Вычисление окончательного результата на основе промежуточных значений

г) Преобразование данных из структуры key-value в таблицу

5. Какую роль играют "ключи" и "значения" в механизме MapReduce?

а) Ключи - категории, по которым группируются данные, значения - данные, которые сопоставлены с каждым ключом

б) Ключи - идентификаторы, по которым происходит сортировка данных, значения - промежуточные результаты вычислений

в) Ключи - переменные, которые используются в вычислениях, значения - выходные результаты операций map и reduce

г) Ключи и значения не играют особой роли в механизме MapReduce

6. Какие преимущества имеет парадигма MapReduce при обработке больших объемов данных?

а) Параллельная обработка данных, масштабируемость, обработка отказов

б) Высокая точность и надежность вычислений

в) Простота использования и низкая стоимость реализации

г) Возможность обработки неструктурированных данных

7. Какую структуру данных обычно используют для промежуточных результатов операции map в парадигме MapReduce?

а) Хэш-таблицу

б) Граф

в) Список

г) Бинарное дерево

8. Какую структуру данных обычно используют для промежуточных результатов операции reduce в парадигме MapReduce?

а) Хэш-таблицу

б) Граф

в) Список

г) Бинарное дерево

9. Какой язык программирования чаще всего используется для реализации парадигмы MapReduce?

а) Java

б) Python

в) C++

г) R

10. Какой фреймворк/систему чаще всего используют при работе с парадигмой MapReduce?

а) Apache Hadoop

б) Apache Spark

в) MongoDB

г) MySQL

11. Что такое "делитель" (partitioner) в контексте парадигмы MapReduce?

а) Компонент, определяющий, к какому reducer'у отправлять промежуточные результаты операции map

б) Алгоритм, определяющий кратчайший путь в графе

в) Функция, применяемая к значениям на этапе операции reduce

г) Компонент, выполняющий операцию reduce на основе промежуточных результатов операции map

12. Какие типы задач хорошо подходят для решения с помощью парадигмы MapReduce?

а) Агрегирование данных, сортировка, фильтрация, группировка

б) Машинное обучение, прогнозирование, оптимизация

в) Графовые алгоритмы, поиск путей, кластеризация

г) Кодирование, декодирование, обработка изображений и звука

13. Каким образом парадигма MapReduce обеспечивает параллельность обработки данных?

а) Разбивает данные на фрагменты (chunks) и выполняет операции map и reduce параллельно на этих фрагментах

б) Использует многопоточность для ускорения выполнения операций map и reduce

в) Использует распределенные вычисления на кластере, где каждый узел обрабатывает свою часть данных

г) Комбинирует результаты из разных map и reduce операций на разных узлах кластера

14. Каким образом парадигма MapReduce обеспечивает отказоустойчивость в случае сбоев узлов кластера?

а) Репликация данных на нескольких узлах кластера

б) Создание резервных копий данных перед выполнением операций map и reduce

в) Использование транзакций при выполнении операций map и reduce

 г) Масштабирование кластера при обнаружении сбоев

15. Какой этап парадигмы MapReduce является более трудоемким: операция map или операция reduce?

а) Это зависит от конкретной задачи и объемов данных

б) Операция map обычно более трудоемкая

 в) Операция reduce обычно более трудоемкая

 г) Оба этапа имеют примерно одинаковую трудоемкость

16. Каким образом парадигма MapReduce обрабатывает большие объемы данных, не помещающиеся в оперативную память одного узла?

а) Разбивает данные на фрагменты (chunks) и обрабатывает их по частям

б) Загружает данные на диск и обрабатывает их, минимизируя использование оперативной памяти

в) Использует специальные алгоритмы сжатия данных перед обработкой

г) Использует инфраструктуру распределенного файла для хранения и обработки данных

17. Каким образом реализуются операции map и reduce в парадигме MapReduce на кластере?

а) Каждый узел кластера выполняет операции map и reduce на своих локальных данных

б) Координирующий узел кластера выполняет операции map и reduce на собранных данных

в) Операции map и reduce могут выполняться параллельно на разных узлах кластера

г) Операции map и reduce выполняются на каждом узле кластера, с последующим слиянием результатов

18. Какой этап парадигмы MapReduce может быть выполняем несколько раз для работы с несколькими операциями reduce?

а) Операция map

б) Операция reduce

в) Оба этапа могут быть выполняемы несколько раз

г) Промежуточный этап - операция sort

19. Какой из следующих элементов есть в каждой задаче MapReduce?

 а) MapReduce Job (задача)

б) Mapper (отображение)

в) Reducer (сокращение)

г) Все вышеперечисленные элементы

20. Какие возможности предоставляет парадигма MapReduce для реализации алгоритмов обработки данных?

а) Возможность параллельного выполнения операций на большом количестве узлов

б) Легкость масштабирования обработки данных при увеличении количества узлов кластера

в) Возможность обработки больших объемов данных, не помещающихся в память одного узла

г) Все вышеперечисленные возможности

# 5 Титульный лист

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Бугульминский филиал федерального государственного бюджетного образовательного учреждения высшего образования

«Казанский национальный исследовательский технологический университет»

**Контрольная работа**

|  |  |
| --- | --- |
| по дисциплине | Большие данные |
| Группа | 2115 |  |
| Обучающийся |  |  |  |  |
|  | (подпись) |  |  Фамилия И.О. |  |
| Преподаватель |  |  | Хакимова А.А. |  |
|  | (подпись) |  |  Фамилия И.О. |  |
| Оценка |  | Дата |  |  |
| Регистрационный № |  | дата регистрации |  |
| Адрес электронной почты студента |  |  |

Бугульма – 2024