Вопросы к экзамену по дисциплине «Технологии обработки больших данных»

1. Опишите задачи, приводящие к необходимости использования технологии MapReduce. Раскройте смысл этой технологии. Приведите пример модели использования MapReduce.
2. Укажите область применения технологии PageRank. Определите модель потока для этой технологии, укажите способы анализа модели. Приведите пример использования.
3. Укажите проблемы, возникающие при использовании технологии PageRank в web-сетях. Укажите способы их устранения. Приведите соответствующие модели потока. Сформулируйте теорему Брина-Пейджа.
4. Укажите основные технологии поиска похожих объектов, их содержание и области применения. Дайте определение шинглов, приведите примеры. Опишите технологию minhash.
5. Дайте определение коэффициента схожести Жаккара. Опишите технологию хэширования с учетом близости. Укажите, как вычисляются вероятности совпадения/различия сигнатур и как можно повысить/уменьшить эти вероятности.
6. Дайте определение метрики множества. Укажите основные используемые метрики, области их применения, преимущества и недостатки.
7. Укажите практические области применения рекомендательных систем, основные проблемы с которыми сталкиваются при использовании систем и возможные пути решения. Опишите, как функционируют рекомендации на основе содержания, их плюсы и минусы.
8. Изложите алгоритм работы коллаборативной фильтрации, способы нахождения похожих пользователей. Приведите формулы расчета рейтингов. Укажите плюсы и минусы такой фильтрации.
9. Дайте определение кластеризации, укажите использующие её практические области. Раскройте метод иерархической кластеризации, укажите его модификации. Приведите критерии остановки алгоритма.
10. Дайте определение кластеризации, укажите использующие её практические области. Раскройте метод кластеризации k-means, укажите его модификации. Приведите критерии остановки алгоритма.
11. Укажите преимущества и недостатки методов классификации и регрессии, основанных на деревьях решений. Опишите пошаговый процесс построения регрессионного дерева решений.
12. Объясните, зачем проводится обрезка деревьев решений. Укажите возможные стратегии. Приведите алгоритм поиска наилучшего поддерева регрессионного дерева.
13. Опишите процесс построения дерева классификации. Приведите формулы разбиения. Укажите преимущества и недостатки деревьев классификации и регрессии. Опишите методы улучшения прогноза.
14. Опишите процедуру аналитического поиска собственных чисел и собственных векторов матрицы. Приведите и обоснуйте алгоритм поиска собственных чисел и собственных векторов матрицы степенным методом.
15. Опишите метод главных компонент понижения размерности. Приведите пример. Выведите соответствие между собственными числами матриц M^TM и MM^T.
16. Дайте определение сингулярного разложения матрицы и его интерпретацию. Объясните, как можно использовать сингулярное разложение для понижения размерности.
17. Укажите способы снижения размерности с помощью отбора признаков. Опишите, как осуществляется пороговая обработка дисперсии признаков. Укажите критерии удаления высоко коррелированных и нерелевантных признаков. Опишите алгоритм рекурсивного устранения признаков.
18. Приведите особенности социальных графов. Опишите недостатки применения к социальным графам стандартных методов кластеризации. Приведите пошаговый алгоритм Гирвана-Ньюмана кластеризации.
19. Дайте определение разреза графа и нормализованного разреза. Укажите способ поиска разрезов графа с использованием матрицы Лапласа.