Список вопросов к экзамену по Анализу Данных

Вопросы с черточкой обязательные. Вопросы с кружочком для тех, кому слишком просто и кто хочет произвести неизгладимое приятное впечатление на экзаменатора.

Big Data

- Необходимость распределенных вычислений в современной аналитике.
- Три особенности больших данных: Volume, Variety, Velocity.
- Распределенное хранение данных (DFS).
- Парадигма MapReduce (общее понимание).
- Написать WordCount на MapReduce.

◉ Reduce combiner. Какие задачи помогает решать, какие не помогает.

Нейронные сети

- Общее представление. Какие задачи можно решать? Какие нельзя?
- Градиентный спуск. Стохастический градиентный спуск.
- Мотивация создания глубоких сетей. Особенности глубоких сетей.
- Сверточные нейронные сети.
- Переобучение сетей. Регуляризация и dropout.

◉ Рассказать про AlphaGo (или ваш любимый проект на нейронных сетях).

Анализ текстов и NLP

- Матрица термы-на-документы. Предпосылки. Почему может оказаться ну очень большой? Как её лучше хранить в памяти?
- Нормализация слов. Стеминг и лематизация.
- Борьба со стоп-словами и с редкими словами.
- Мера tf-idf.
- Что такое word2vec? Что он позволяет делать?

◉ Как обучить word2vec?

◉ Подсчитать tf-idf через MapReduce.