Вопросы к экзамену по курсу "Основы обработки текстов". 2019 г.

1. Задачи обработки текста. Многозначность при обработке текста. Проблема понимания. Тест Тьюринга. Китайская комната.
2. Типичная схема решения задач анализа данных.
3. Задача определения и классификации именованных сущностей. Оценка качества. Методы решения.
4. Машинное обучение. Перекрестная проверка. Переобучение.
5. Методы классификации. Линейные классификаторы. Метод опорных векторов.
6. Методы классификации. Мультиномиальная логистическая регрессия.
7. Методы классификации. Скрытая марковская модель. Алгоритм Витерби.
8. Модели классификации. Марковская модель максимальной энтропии.
9. Модель нейрона МакКаллока-Питтса. Персептрон. Обучение нейронной сети с учителем. Метод стохастического градиентного спуска.
10. Модель нейрона МакКаллока-Питтса. Персептрон. Граф вычислений. Алгоритм обратного распространения ошибки (backpropagation). Проблема затухающих градиентов.
11. Разметка последовательности с помощью нейронных сетей на примере задачи распознавания именованных сущностей. One-hot кодирование. Рекуррентные нейронные сети. LSTM.
12. Проверка статистических гипотез. Нулевая гипотеза. Уровень значимости. P-value. Ошибки 1-го и 2-рода.
13. Проверка статистических гипотез. Т-критерий Стьюдента. Поиск словосочетаний.
14. Проверка статистических гипотез. Критерий Хи-квадрат. Поиск словосочетаний.
15. Проверка статистических гипотез. Критерий отношения правдоподобия. Поиск словосочетаний.
16. Проверка статистических гипотез. Сравнение классификаторов. Критерий Уилкоксона. Проблемы использования статистической проверки гипотез.
17. Векторные представления слов. Дистрибутивная гипотеза. Локальные модели векторов слов: модели skip-gram и continuous bag of words.
18. Векторные представления слов. Дистрибутивная гипотеза. Вычислительная сложность softmax. Иерархический softmax и negative sampling.
19. Векторные представления слов. Дистрибутивная гипотеза. Матрица совместной встречаемости слов. Глобальные модели векторов слов: модель GloVe.
20. Векторные представления слов. Мера близости слов. Способы оценки качества векторов слов.
21. Векторные представления слов. Проблема редких слов в моделях векторных представлений слов. Модель fasttext.
22. Векторные представления слов. Проблема редких слов в моделях векторных представлений слов. Модель CharCNN.
23. Сегментация текста: задачи токенизации и определения границ предложений.
24. Задача определения языка текста. Профили языка. Оценка качества классификации. Проблема коротких и мультиязычных текстов.
25. Задача определения языка текста. Naïve Bayes классификатор. Оценка качества классификации. Проблема коротких и мультиязычных текстов.
26. Задача определения частей речи и морфологического анализа. Разметка последовательности. Multi-label классификация.
27. Задача нормализации слов. Ripple Down Rules и LemmaGen. Грамматическая омонимия и способы борьбы с ней.

28. Задача нормализации слов. Нормализация слов как задача классификации. Грамматическая омонимия и способы борьбы с ней.
29. Синтаксический анализ. Поверхностный синтаксический анализ.
30. Синтаксический анализ. Дерево составляющих. Понятие формальной грамматики, иерархия Хомского. Генерация текста по формальной грамматике.
31. Синтаксический анализ. Дерево составляющих. Разбор предложения по грамматике составляющих. Алгоритм Кока-Янгера-Касами. Неоднозначность разбора, выбор лучшего разбора.
32. Синтаксический анализ. Дерево зависимостей. Разбор предложения на основе переходов: Arc-standard, Arc-eager.
33. Синтаксический анализ. Дерево зависимостей. Проективность деревьев разбора. Методы разбора для непроективных деревьев: Attardi's system, Online reordering.
34. Синтаксический анализ. Дерево зависимостей. Stack LSTM. Синтаксический разбор на основе Stack LSTM.
35. Синтаксический анализ. Графовые методы синтаксического анализа. MST парсер. Алгоритм Эдмондса.
36. Синтаксический анализ. Графовые методы синтаксического анализа. Biaffine graph-based dependency parser.
37. Синтаксический анализ. Дерево составляющих и дерево зависимостей. Оценка качества построенных деревьев разбора.
38. Лексическая многозначность. WordNet. Значения слов.
39. Разрешение лексической многозначности. Алгоритмы классификации. Методы оценки качества.
40. Разрешение лексической многозначности. Методы, основанные на словарях и тезаурусах. Варианты алгоритма Леска. Методы оценки качества.
41. Привязка к базам знаний. Основные проблемы. Алгоритм Milne-Witten.
42. Информационный поиск. Виды поисковых задач: Lookup, Learn, Investigate, особенности процесса поиска.
43. Информационный поиск. Векторное представление документа, инвертированный индекс, булева модель.
44. Информационный поиск. Оценка качества результатов поиска, асессоры, Mean Average Precision.
45. Анализ тональности текстов. Формальная постановка и ее варианты. Подходы к решению задачи.
46. Вопросно-ответные системы. Общая архитектура. Обработка запроса. Извлечение фрагментов текста. Обработка ответа.
47. Автоматическое реферирование. Общая архитектура.Отбор контента. Упорядочение. Переконструирование предложения.
48. Машинный перевод. Различия в языках. Классические подходы к машинному переводу.
49. Статистический машинный перевод. Модель зашумленного канала. Модель перевода на основе фраз. Выравнивание фраз. Декодирование. Методы оценки качества. BLEU.
50. Статистический машинный перевод. Выравнивание слов. Модель IBM Model1. Тренировка моделей выравнивания. EM-алгоритм.
51. Модели кластеризации. Иерархическая кластеризация. Метод К-средних.
52. Тематическое моделирование. Вероятностный латентный семантический анализ.
53. Тематическое моделирование. Скрытое размещение Дирихле.