Вопросы к экзамену по с/к "Основы обработки текстов". 2025 г.

1. Задачи обработки текста. Многозначность при обработке текста. Проблема понимания. Тест Тьюринга. Китайская комната
2. Понятие языковой модели. Модель N-грамм. Сглаживание. Оценка качества. Перплексия.
3. Типичная схема решения задач. Матрица ошибок. Каппа Коэна.
4. Задача определения и классификации именованных сущностей. Оценка качества. Методы решения. Схемы разметки последовательностей IO, BIO, BMEWO
5. Методы классификации. Линейные классификаторы. Мультиномиальная логистическая регрессия.
6. Методы классификации последовательностей. Скрытая марковская модель. Алгоритм Витерби.
7. Модели классификации последовательностей. Марковская модель максимальной энтропии. Условные случайные поля.
8. Искусственный нейрон. Персептрон. Обучение нейронной сети с учителем. Метод стохастического градиентного спуска.
9. Искусственный нейрон. Персептрон. Граф вычислений. Алгоритм обратного распространения ошибки (backpropagation). Проблема затухающих градиентов.
10. Рекуррентные сети. Обратное распространение во времени. Проблема катастрофического забывания. Разметка последовательности с помощью нейронных сетей на примере задачи распознавания именованных сущностей. One-hot кодирование. Рекуррентные нейронные сети. LSTM.
11. Проверка статистических гипотез. Нулевая гипотеза. Уровень значимости. P-value. Ошибки 1-го и 2-рода
12. Проверка статистических гипотез. Т-критерий Стьюдента. Критерий Хи-квадрат. Критерий отношения правдоподобия. Поиск словосочетаний.
13. Проверка статистических гипотез. Сравнение классификаторов. Критерий Уилкоксона. Проблемы использования статистической проверки гипотез
14. Векторные представления слов. Дистрибутивная гипотеза. Локальные модели векторов слов: модели skip-gram и continuousbag of words.
15. Векторные представления слов. Дистрибутивная гипотеза. Вычислительная сложность softmax. Иерархический softmax и negative sampling.
16. Векторные представления слов. Дистрибутивная гипотеза. Матрица совместной встречаемости слов. Глобальные модели векторов слов: модель GloVe.
17. Векторные представления слов. Мера близости слов. Способы оценки качества векторов слов.
18. Векторные представления слов. Проблема редких слов в моделях векторных представлений слов. Модель fasttext.
19. Векторные представления слов. Проблема редких слов в моделях векторных представлений слов. Модель CharCNN.
20. Сегментация текста: задачи токенизации и определения границ предложений.
21. Задача определения языка текста. Профили языка. Оценка качества классификации. Проблема коротких и мультиязычных текстов.
22. Задача определения языка текста. NaïveBayes классификатор. Оценка качества классификации. Проблема коротких и мультиязычных текстов.
23. Задача определения частей речи и морфологического анализа. Разметка последовательности. Multi-label классификация.

24. Задача нормализации слов. Нормализация слов как задача классификации. Грамматическая омонимия и способы борьбы с ней.
25. Синтаксический анализ. Дерево составляющих. Понятие формальной грамматики, иерархия Хомского. Генерация текста по формальной грамматике.
26. Синтаксический анализ. Дерево составляющих. Разбор предложения по грамматике составляющих. Алгоритм Кока-Янгера-Касами. Неоднозначность разбора, выбор лучшего разбора.
27. Синтаксический анализ. Дерево зависимостей. Разбор предложения на основе переходов: Arc-standard, Arc-eager.
28. Синтаксический анализ. Дерево зависимостей. Проективность деревьев разбора. Методы разбора для непроективных деревьев: Attardi's system, Online reordering.
29. Синтаксический анализ. Дерево зависимостей. Stack LSTM. Синтаксический разбор на основе Stack LSTM.
30. Синтаксический анализ. Графовые методы синтаксического анализа. MST парсер. Алгоритм Эдмондса.
31. Синтаксический анализ. Графовые методы синтаксического анализа. Biaffine graph-based dependency parser.
32. Синтаксический анализ. Дерево составляющих и дерево зависимостей. Оценка качества построенных деревьев разбора.
33. Лексическая многозначность. Тезаурус WordNet. Значения слов. Графы знаний.
34. Разрешение лексической многозначности. Алгоритмы классификации. Алгоритма Леска. Методы оценки качества
35. Привязка к базам знаний. Основные проблемы. Алгоритм Milne-Witten
36. Задача языкового моделирования. Нейросетевые языковые модели. Language Model Embedding. ELMo
37. Задача языкового моделирования. Нейросетевые языковые модели. GPT
38. Задача языкового моделирования. Нейросетевые языковые модели. BERT
39. Информационный поиск. Задачи информационного поиска, общая архитектура поисковых систем. Векторное представление документа, векторная модель поиска. Инвертированный индекс. Булева модель. Взвешивание слов TF-IDF.
40. Информационный поиск. Ранжирование. Методы ранжирования. Оценка качества ранжирования. Точность. Полнота. Усредненная средняя точность (MAP).
41. Информационный поиск. Поиск по словарям. Запросы с джокерами. Сжатое префиксное дерево. Перестановочные индексы. K-граммный индекс.
42. Генерация с извлечением информации (RAG). Архитектура и основные элементы.
43. Поиск в Вебе. Алгоритм PageRank.
44. Анализ тональности текстов. Формальная постановка и ее варианты. Подходы к решению задачи
45. Вопросно-ответные системы. Общая архитектура. Обработка запроса. Извлечение фрагментов текста. Обработка ответа.
46. Автоматическое реферирование. Общая архитектура. Отбор контента. Упорядочение. Переконструирование предложения. Методы оценки качества.
47. Машинный перевод. Различия в языках. Классические подходы к машинному переводу
48. Статистический машинный перевод. Модель зашумленного канала. Модель перевода на основе фраз. Выравнивание фраз. Декодирование. Методы оценки качества. BLUE
49. Статистический машинный перевод Выравнивание слов. Модель IBM Model1. Тренировка моделей выравнивания. EM-алгоритм.

50. Модели кластеризации. Иерархическая кластеризация. Метод К-средних
51. Тематическое моделирование. Вероятностный латентный семантический анализ. Скрытое размещение Дирихле. Аддитивная регуляризация, Robust PLSA.
52. Языковое моделирование. Предобучение (pre-training) и дообучение (fine-tuning) больших языковых моделей. Zero-shot, few-shot, in-context learning, chain-of-thought. Как оценивается стоимость обучения языковых моделей, данные vs размер.
53. Эволюция больших языковых моделей: GPT-1, GPT-2, GPT-3. FLAN и инструктивное дообучение (Instruct tuning). Различия между foundation models и instruct models.