Вопросы к экзамену по с/к "Основы обработки текстов". 2015 г.

1. Задачи обработки текста. Многозначность при обработке текста. Проблема понимания. Тест Тьюринга. Китайская комната
2. Регулярные выражения
3. Конечные автоматы, распознавание языка с помощью КА
4. Регулярные языки и конечные автоматы. Построение КА для регулярных выражений
5. Модель N-грамм. Оценка вероятности высказывания
6. Модель N-грамм. Сглаживание (Лапласа, откат, интерполяция)
7. Модель N-грамм. Оценка качества. Тренировочный и проверочный корпуса
8. Задача определения частей речи. Существующие подходы. Алгоритмы, основанные на правилах. Алгоритмы, основанные на трансформации.
9. Методы поиска словосочетаний. Использование мат. ожидания и дисперсии.
10. Методы поиска словосочетаний. Проверка статистических гипотез. Т-критерий Стьюдента.
11. Методы поиска словосочетаний. Проверка статистических гипотез. Критерий Хи-квадрат.
12. Методы поиска словосочетаний. Проверка статистических гипотез. Критерий отношениея правдоподобия
13. Использование скрытой марковской модели для определения частей речи. Алгоритм Витерби
14. Модели классификации. Наивный байесовский классификатор
15. Модели классификации. Логистическая регрессия, модель максимальной энтропии
16. Модели классификации. Марковская модель максимальной энтропии
17. Модели кластеризации. Иерархическая кластеризация
18. Модели кластеризации. Метод К-средних
19. Типы грамматик. Грамматика составляющих. Грамматика зависимостей. Категориальная грамматика
20. Контекстно-свободные грамматики. КС грамматики и регулярные языки. Банк деревьев.
21. Синтаксический разбор. Разбор сверху вниз и снизу вверх
22. Синтаксический разбор. Алгоритм Кока-Янгера-Касами (CKY parsing). Эквивалентность КС грамматик
23. Синтаксический разбор. Группировка (chunking)
24. Стохастические контекстно-свободные грамматики. Разрешение синтаксической многозначности
25. Моделирование языка. Обучение стохастических КС грамматик
26. Вероятностная версия алгоритма Кока-Янгера-Касами. Оценка качества
27. Проблемы стохастический КС грамматик. Алгоритм Коллинза. Оценка качества
28. Лексическая семантика. WordNet. Значения слов
29. Разрешение лексической многозначности. Алгоритмы классификации. Самонастройка. Методы оценки качества
30. Разрешение лексической многозначности. Методы, основанные на словарях и тезаурусах. Варианты алгоритма Леска. Методы оценки качества
31. Семантическая близость слов. Подходы на основе тезаурусов. Методы оценки качества
32. Семантическая близость слов. Подходы на основе статистик. Методы оценки качества
33. Вопросно-ответные системы. Общая архитектура. Обработка запроса
34. Вопросно-ответные системы. Общая архитектура. Извлечение фрагментов текста
35. Вопросно-ответные системы. Общая архитектура. Обработка ответа
36. Автоматическое реферирование. Общая архитектура
37. Машинный перевод. Классические подходы
38. Статистический машинный перевод. Модель зашумленного канала. Модель перевода на основе фраз. Выравнивание фраз. Декодирование
39. Статистический машинный перевод. Выравнивание слов. Модель IBM Model 1
40. Статистический машинный перевод. Выравнивание слов. Тренировка моделей выравнивания
41. Статистический машинный перевод. Методы оценки качества. BLUE
42. Тематическое моделирование. PLSA, LDA