Вопросы осень 2014-1

1. Основные свойства естественного языка
2. Что такое графематический анализ?
3. Что такое лемматизация
4. Как работает словарный морфологический анализ?
5. Как морфологические анализаторы обрабатывают слова, отсутствующие в словаре
6. Что такое постморфологический анализ. Основные методы.
7. Что такое статистическая языковая модель?
8. В чем отличие частотного подхода к вероятности от байесовского подхода. Поясните на примере подбрасывания монетки
9. Что такое сглаживание в языковой модели?
10. Что такое правило Лапласа в языковой модели? Зачем оно нужно?
11. Основные понятия информационного поиска
12. Какие задачи относятся к задачам информационного поиска?
13. Виды поисковых систем по охвату и направленности. Особенности разных типов поисковых систем
14. Особенности научного поиска
15. Основные этапы обработки текстов в поисковой машине
16. Основные этапы обработки запроса в поисковой машине
17. Булевская модель информационного поиска. Преимущества и недостатки булевской модели поиска
18. Как измеряется качество булевского поиска
19. Что такое векторная модель информационного поиска?
20. Поясните смысл показателей *idf* и *tf.idf*.
21. Классическая процедура оценки качества информационно поиска
22. Что такое РОМИП, какие задачи в нем решаются?
23. Что такое кривая полнота-точность? Что такое 11-точечный график TREC?
24. Что такое пулинг в информационном поиске? Сложности, связанные с пулингом
25. Оценка качества в поисковых машинах Интернет.
26. Шкалы оценок. Мера NDCG
27. Что такое информационно-поисковые тезаурусы? Зачем они нужны? Где применяются сейчас
28. Что такое рубрикаторы? Чем они отличаются от информационно-поисковых тезаурусов?
29. Назовите методы расширения запросов пользователей при информационном поиске.
30. Что означает термин *relevance feedback*? Поясните основные принципы работы
31. Алгоритм Роккио для *relevance feedback*
32. Назовите проблемы расширения запроса при помощи обратной связи по релевантности
33. Укажите основные методы автоматической рубрикации (классификации) текстов.
34. Что такое инженерный метод рубрикации текстов?
35. Укажите плюсы и минусы ручного рубрицирования.
36. Укажите плюсы и минусы инженерных методов рубрикации.
37. Метод Байеса для автоматической рубрикации
38. Поясните методы автоматической рубрикации на основе векторного пространства
39. Метод Роккио для автоматической рубрикации
40. Метод Knn для автоматической рубрикации текстов
41. Поясните основной принцип метода SVM для автоматической рубрикации текстов

42. Плюсы и минусы методов машинного обучения для рубрикации текстов
43. Что такое кластеризация текстов? Чем она отличается от классификации (рубрикации) текстов?
44. Метод K-means для кластеризации текстов
45. Аггломеративная кластеризация – основной принцип и подвиды
46. Методы тестирования автоматической кластеризации
47. Особенности кластеризации потока новостей в реальном времени

Задачи на следующие темы:

1. Точность, полнота, F-мера – меры качества
2. Мера качества упорядочения: средняя точность
3. Нахождение близости между запросом и документом по векторной модели
4. Мера упорядочения: NDCG
5. Макро- и микро- усреднение при оценке качества автоматической рубрикации
6. Байесовская модель классификации текстов
7. Правило Лапласа для предсказания последовательности слов в тексте