Утверждаю: Зав. кафедрой 806, Пирумов У.Г.

Информационный поиск и обработка текстов на естественном языке, 2012/2013 учебный год Экзаменационная программа

1. Основные понятия информационного поиска: документ, слово, термин, запрос, релевантность, полнота и точность. Булевский поиск. Организация индекса для булевского поиска, выполнение запросов. Достоинства и недостатки булевской модели.
2. Лингвистика. Рационализм и эмпиризм в изучении языка. Статистическая лингвистика. Основные задачи статистической лингвистики. Законы Ципфа.
3. Выбор единицы индексирования (документа). Предварительная обработка документов, разбиение на слова, выделение терминов, нормализация, стоп-словарь. Характерные особенности текстов, написанных на естественных языках: омонимия, компаунды, морфология. Основные подходы к морфологической обработке.
4. Коллокации, методы статистического поиска коллокаций.
5. Марковские цепи. Скрытые марковские цепи. Использование марковских цепей для обработки текстов. Алгоритмы «вперёд-назад», Витерби, Баума-Велша.
6. Марковские цепи. Скрытые марковские цепи. Определение частей речи при помощи марковских цепей. Исправление опечаток при помощи марковских цепей.
7. Машинный перевод. Параллельные тексты, методы их выравнивания.
8. Структура поискового индекса, координатные блоки, выполнение поисковых запросов. Методы ускорения выполнения многословных поисковых запросов. Виды индекса для цитатного поиска.
9. Организация словаря терминов, основные структуры данных. Выполнение запросов с метасимволами, исправление ошибок в запросах. Расстояние Левенштейна, фонетическая близость.
10. Построение индекса, основные методы сортировки и слияния координатной информации.
11. Распределённое индексирование, понятие о MapReduce. Обновление индексов, индексирование динамических массивов документов.
12. Сжатие поискового индекса. Статистические характеристики словаря терминов и координатных блоков. Сжатие словаря и координатных блоков.
13. Сжатие поискового индекса. Коды Голомба. Методы сжатия семейств PForDelta, Simple.
14. Ранжирование документов. Разбиение документов на зоны, отличие зон от полей метаинформации, использование информации о зонах для вычисления релевантности. Организация индекса с учётом информации о зонах.
15. Ранжирование документов, учёт количества терминов в документе и в массиве документов. Ранжирование tf-idf, его достоинства и недостатки. Модификации tf-idf.

16. Модель векторного пространства, мера близости двух документов. Ранжирование документов в векторном пространстве, выполнение поискового запроса. Эвристики, позволяющие сократить время выполнения запроса. Достоинства и недостатки модели.
17. Различные методы оценки качества поиска, их достоинства и недостатки.
18. Дизайн поисковой выдачи. Построение сниппетов.
19. Статистическая модель, BIM. Ранжирование документов, достоинства и недостатки модели.
20. Построение n-грамм, их применение. Разреженность, сглаживание.
21. Построение моделей языка, ранжирование с учётом моделей языка, порождёнными документами в индексе.
22. Поисковая машина по сети Интернет, основные особенности: документы, ссылочный граф, поисковый спам, контекстная реклама.
23. Определение нечётких дублей, шинглы. LSH, архитектура системы подавления дублей в масштабе сети Интернет.
24. Выкачка документов из сети Интернет, архитектура «паука». Требования и рекомендации к поведению роботов-«пауков». Технические детали реализации.
25. Ссылочное ранжирование, использование текстов ссылок для расчёта релевантности. Алгоритмы PageRank и HITS, их достоинства и недостатки.
26. Методы анализа поведения пользователей для выделения факторов ранжирования и определения качества поиска. Применение марковских цепей для анализа поведения пользователей на поисковой выдаче.
27. Оценка качества веб-поиска.
28. Машинное обучение для задачи ранжирования.

Составил: ст. преп. каф. 806, Калинин А.Л.