Избор теми за реферат по КЛ за редовни студенти (2023-2024г.)

Това са ви темите за реферат. Всеки да коментира тук във форума номера, който си е избрал. Една тема не може да бъде избрана повече от двама човека.

Минимални изисквания за рефератите: 1. Минимум 7 източника; 2. Минимум 10 страници; 3. Заглавна страница по избор.

1. Проблеми при компютърната обработка на естествен език, корпусна лингвистика, статистически подход, примерни приложения.

2. Категоризиране на текст. Основни подходи. Разлики с клъстеризация. Оценка на качеството: матрица на грешките, accuracy, precision, recall, F-measure. Множество категории, йерархично категоризиране, обучение и тестване на класификатор

3. Класификация. Линейни класификатори: персептрон, winnow, SVM. Множествена класификация: метод на най-близкия съсед. Класификационно дърво.

4. Наивен Бейсов класификатор: биномен и мултиномен модел. Особености на реализацията (smoothing).

5. Избор на атрибути: хи-квадрат, информационна печалба, взаимна информация, DF. Претегляне на термини с TF.IDF. Нормализиране, стеминг и лематизация.

6. Векторен пространствен модел. Латентен семантичен анализ.

7. Клъстеризиране на текст. Йерархично клъстеризиране: single-link, complete-link, average-link. Метод на центроида (K-means).

8. Търсене в големи текстови масиви (Information Retrieval). Обратен индекс. Обработка на булеви заявки. Подреждане на резултатите.

9. Устройство и работа на търсещите машини в Уеб. MapReduce. Уеб-спам: видове и борба с него.

10. Устройство и работа на търсещите машини в Уеб. Анализ на връзките между Уеб страниците: PageRank и HITS.

11. Разрешаване на семантична многозначност на ниво дума (word sense disambiguation).

12. Максимална ентропия (maximum entropy).

13. Сравнение на два модела - статистическа значимост. Доверителни интервали за пропорция.

14. Анотиране с част на речта (POS tagging). Скрит марковски модел (HMM). ME и МЕММ. CRF.

15. Анотиране с част на речта (POS tagging). Обучение чрез трансформация и използване на правила: тагер на Ерик Брил.

16. Анотиране с части на речта (POS tagging) и лематизация за български език с използване на рекурентна невронна мрежа (recurrent neural network).

17. Извличане на информация/факти от текст (information extraction). Разпознаване на именовани единици (Named Entity Recognition). Модел B-XXX, I-XXX, O.

18. Разрешаване на синтактична и семантична многозначност с използване на Уеб. Характеризиране на семантични отношения, измерване на близост, автоматично решаване на тестове SAT.

19. Граматика: конституентност, съгласуване, структурна многозначност. Парсиране на текст. Контекстно свободна граматика. Нормална форма на Чомски. Парсер CKY. Chart parser.

20. Вероятностен парсер. Учене на вероятности от PennTreeBank: проблеми и решения. Бинаризация. Анотация с категорията на родителя. Разбиване на символите. Хоризонтална марковизация. Вертикална марковизация. Лексикализация. Депендентно парсиране. Пренареждане на най-добрите кандидати (re-ranking). Вероятност на дърво и вероятност на парсиране.

21. Езикови модели. N-грами. Измерване на качеството на лингвистичен модел: перплексия и ентропия. Закон на Зипф. Условни езикови модели: сравнение с Наивен Бейсов класификатор. Приложения.

22. Отговаряне на въпроси. Обща архитектура. Типове въпроси. Използване на Уеб. AskMSR. Използване на шаблони. Използване на знания.

23. Резюмиране на текст. Резюмиране на един документ. Теория на риторичната структура (RST). Резюмиране на множество документи.

24. Лексикони, семантични мрежи и онтологии. Релации в WordNet. Измерване на семантична близост с използване на WordNet. Автоматично извличане на семантични релации: лексико-семантични шаблони на Хърст.

25. Машинен превод.