Выделение ключевых словосочетаний

Максим Яковлев

НУГ "Методы анализа и визуализации веб-корпусов"

Томита-парсер

Что это?

Утилита, разработанная Яндексом. Выделяет из текста цепочки слов по указанным контекстно-свободным грамматикам.

Зачем?

Будет помогать выделять ключевые словосочетания.

Почему?

Потому что работает.

Грамматика

множество правил на языке КС-грамматик, описывающих синтаксическую структуру выделяемых цепочек

Пример:

```
#GRAMMAR_ROOT S
S -> Noun;
S -> "деньги";
S -> Noun Noun<gram="gen">;
S -> Noun Prep Noun;
S -> Adj<gnc-agr[1]>+ Noun<gnc-agr[1]>;
```

Извлечение словосочетаний

1. Устанавливаем, настраиваем томита-парсер
2. Составляем грамматику
3. Запускаем парсер.

Результат: список словосочетаний из текста, соответствующих указанной грамматике. При этом слова в словосочетании остаются в согласованном виде и, по возможности, все приводятся к нормальной форме.

Выделение ключевых словосочетаний

Имея список словосочетаний, встречающихся в тексте, ключевые словосочетания можно выделить подсчетом частоты вхождения словосочетаний в текст и отбором наиболее часто встречающихся.

Проблема:

подсчитать “в лоб” число вхождений не всегда получится, потому что могут встретиться разные формы слов в словосочетании.

Решение:

1. Провести дополнительную обработку текста и провести подсчет
2. Положиться на результат работы томита-парсера, считая, что он извлекает все вхождения словосочетаний и учитывая, что в приоритете более длинная цепочка. Пример далее

Пример:

Вход: “Поговорим о приватизации жилья. Идёт приватизация.”

Выход: “приватизация жилья”, “приватизация”

Частоты:

приватизация : 2

приватизация жилья : 1

KeyPhraseExtracor

1. Получение статьи по url (newspaper libray)
2. Извлечение словосочетаний (томита-парсер)
3. Выделение ключевых словосочетаний