Семинарское задание №2
Расширенный булев поиск

Мосиенко Константин Викторович

2017

Модифицируйте ваши программы с прошлого семинара с учётом новых требований:

- Интерпретируйте запрос как «документы, содержащие как минимум одно слово из запроса»(вариант «или»).
- Ранжируйте результаты поиска по убыванию метрики схожести запроса и документа.
- Не учитывайте веса слов относительно запроса: $w_{qi} = 1$.
- Положите $p = 1$, в качестве w_{di} воспользуйтесь $tfidf(t_i, d)$. Обратите внимание на то, что необходимо выбрать такие способы нормировки, чтобы $tfidf(t_i, d) \in [0, 1]$.
- На каждый запрос выводите не более 20 результатов поиска. Каждая строчка должна иметь вид:

  [Скор документа]<tab>[Имя файла]<tab>[$tfidf(t_1, d), tfidf(t_2, d), ...$]

  Т.е. помимо имени файла необходимо вывести значение метрики схожести и вес каждого слова из запроса.

Назовите ваши программы make_index и search. Автоматическая проверка не предусмотрена, поэтому не переживайте, из-за того, что вам кажется, что вы как-то не так разбили предложения на слова, за исключением того, что в словах не должно быть посторонних символов, таких как знаки препинания. Также рекомендуется все слова приводить к нижнему регистру. Коллекцию документов можно взять со страницы курса: txt.tar.gz.