Семинарское задание №3
Обобщённая векторная модель

Мосиенко Константин Викторович

2017

В программах, реализованных в предыдущем семинарском задании, необходимо:

- Заменить расширенный булев поиск с весами на основе $tfidf(t_i, d)$ на обобщённую векторную модель со скалярным произведением на основе $NPMI(t_i, t_j)$.
- Для расчёта $NPMI(t_i, t_j)$ пользуйтесь событиями вида «терм встретился в документе» - не считайте количество вхождений. $NPMI(t_i, t_j)$ удобно считать с помощью инвертированного индекса.
- Для расчёта меры близости положите $q_j = 1$ (не учитывайте веса термов относительно запроса) и $a_{\alpha i} = tfidf(t_i, d_\alpha)$.
- Обратите внимание, что мера близости при использовании $NPMI(t_i, t_j)$ может принимать отрицательные значения (слово «мера» мы используем в бытовом контексте).
- Каждая строка выдачи теперь должна иметь вид:

  [Скор документа]<tab>[Имя файла]<tab>[$(t_i, t_j, a_{\alpha j}, \bar{t}_i \cdot \bar{t}_j)$] * 10, где t_i - терм из запроса, а t_j - терм из документа. Обратите внимание на то, что в последнем столбце необходимо вывести не более десяти таплов с максимальным по модулю $\bar{t}_i \cdot \bar{t}_j$.
- В отчёте приведите по 10 самых близких пар термов по $NPMI(t_i, t_j)$ к значениям $-1, 0, 1$.

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.