Московский авиационный институт
Факультет прикладной математики и физики

Лабораторная работа №3

по курсу:
«Обработка естественно-языковых текстов»
по теме:
«Лемматизация»
2 семестр

Студент: Ахмед С. Х.
Преподаватель: Калинин А. Л.
Группа: 8О-106М

Москва, 2019 г

Постановка задачи

Добавить в созданную поисковую систему (ЛР 1-8 по курсу «Информационный поиск») лемматизацию. В простейшем случае, это просто поиск без учёта словоформ. В более сложном случае, можно давать бонус большего размера за точное совпадение слов. Лемматизацию можно добавлять на этапе индексации, можно на этапе выполнения поискового запроса. В отчёте должна быть включена оценка качества поиска, после внедрения лемматизации. Стало ли лучше? Изучите запросы, где качество ухудшилось. Объясните причину ухудшения и как можно было бы улучшить качество поиска по этим запросам, не ухудшая остальные запросы?

Ход работы

Для нахождения словоформ в статьях и запросах используется лемматизатор pymorphy2 с базой OpenCorpora. Лемматизация происходит при парсинге запроса. Полученные постинги словоформ добавляются в список постингов и не включается при поиске цитат. Засчет введения учета морфологии полнота выдачи должна улучшиться, так как теперь в выдаче появляются документы с различными формами слова из запроса:

```
for key in dictionary.keys():
            if morph.parse(key)[0].normal_form == morph.parse(token)[0].normal_form:
               print('morphology')
               result.extend(self.load_postings(dictionary[key][0],dictionary[key][1]))
               queue.append(key)
               idx += 1
               if idx == 5:
                  break
         result = sorted(result
```

Качество поиска

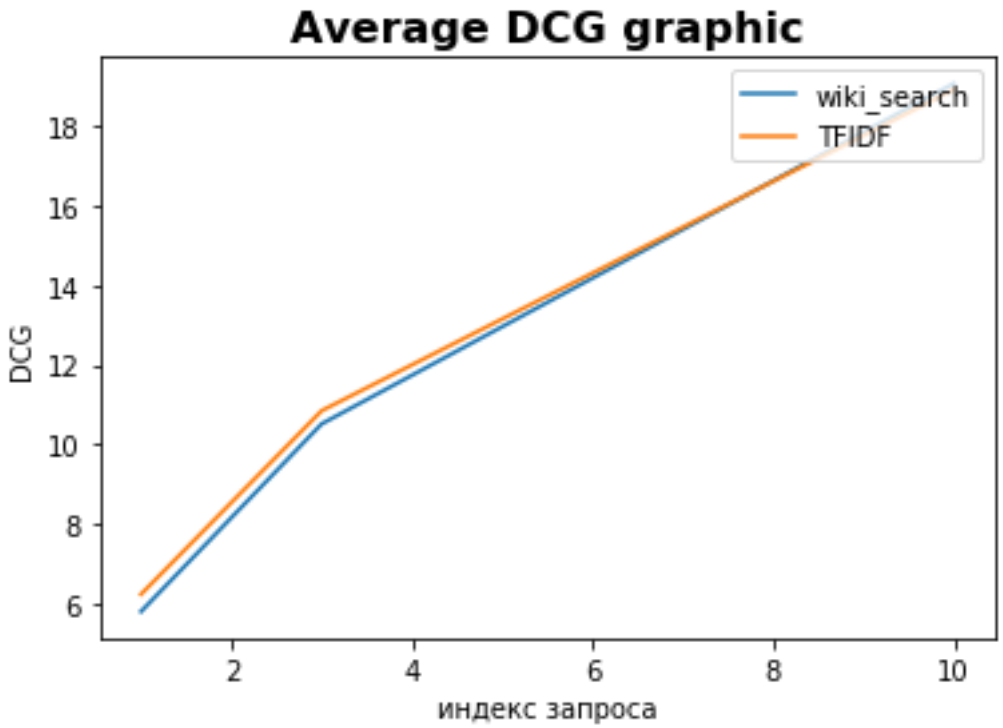
Оценка качества поиска производилась с помощью 30 запросов, которые использовались в лабораторных работах по булевому поиску и TF-IDF ранжированию. Сравнение производилось с TF-Idf ранжированной выдачей (все значение с припиской old)

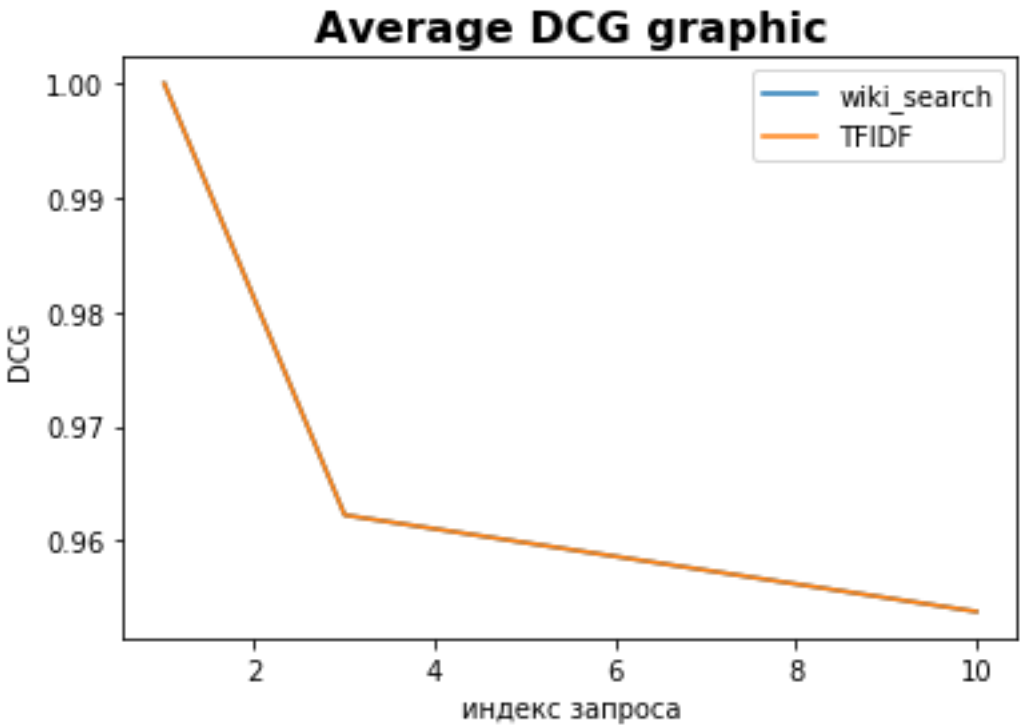
Как можно видеть существенного увеличения качества результата не произошло, некоторые результаты улучшились, некоторые слегка ухудшились. Это может быть объяснено тем, что теперь

в выдачу стали попадать различные формы слов из запроса, что увеличила вариативность выборки. Посмотрим на оценках, верно ли мое утверждение

Out[252]:

	0	1	2	3	4	5	6	7	8	9	P@1	P@3	P@10	P@1Old	P@3Old	P@10Old
служить всю жизнь государству	3	4	4	2	2	2	2	5	2	4	1	1.000000	0.5	0	0.666667	0.3
соглашение о свободной торговле	1	1	4	1	1	4	4	5	3	4	0	0.333333	0.6	0	0.333333	0.4
зимний дворец	5	3	3	4	4	2	1	2	3	4	1	1.000000	0.7	1	1.000000	0.7
российская империя	1	2	5	2	2	1	1	3	2	2	0	0.333333	0.2	0	0.333333	0.2
англо русской войне	2	3	2	2	5	2	2	2	2	2	0	0.333333	0.2	1	0.666667	0.2
всероссийский съезд советов	4	4	4	4	3	2	3	4	2	2	1	1.000000	0.7	1	1.000000	0.7
транссибирская магистраль	5	3	2	2	3	2	2	2	2	2	1	0.666667	0.3	1	0.666667	0.3
свидетели иеговы	5	4	3	3	2	2	3	3	2	5	1	1.000000	0.7	1	1.000000	0.6
владимир ильич ульянов	5	1	1	1	1	5	1	1	1	1	1	0.333333	0.2	1	0.333333	0.1
крымский мост	5	2	2	2	2	2	2	2	2	3	1	0.333333	0.2	1	0.333333	0.2
ионообменные смолы	5	4	2	2	3	3	2	0	5	0	1	0.666667	0.5	1	0.666667	0.2
красный террор	5	5	2	2	2	2	3	2	2	2	1	0.666667	0.3	1	0.666667	0.3
аутсайдер моя жизнь как интрига	5	1	1	1	1	1	1	1	1	1	1	0.333333	0.1	1	0.333333	0.1
день защитника отечества	5	3	2	2	2	2	3	2	2	2	1	0.666667	0.3	1	0.666667	0.3
день пограничника	4	4	5	3	2	2	1	1	1	1	1	1.000000	0.4	1	0.666667	0.4
ударный музыкальный инструмент	5	5	1	5	5	5	5	2	5	5	1	0.666667	0.8	1	0.666667	0.8
генеральный штаб вооружённых сил российской федерации	5	5	2	4	5	2	2	3	2	2	1	0.666667	0.5	1	0.666667	0.5
чирлидинг	5	3	3	4	4	2	1	2	3	4	1	1.000000	0.7	1	1.000000	0.7
голштинских заявлений	0	0	4	0	0	0	0	0	0	0	0	0.333333	0.1	1	0.333333	0.1
быстрая сортировка	2	5	4	2	2	2	2	2	1	1	0	0.666667	0.2	1	1.000000	0.3
умберто боччони	2	2	2	2	2	2	2	2	2	5	0	0.000000	0.1	1	0.333333	0.1
охотники за привидениями	5	4	4	3	3	5	2	4	4	4	1	1.000000	0.9	1	1.000000	0.9
божественный пёс	5	2	2	2	2	3	2	2	2	2	1	0.333333	0.2	1	0.333333	0.2
эдди мерфи	5	4	3	3	3	3	3	3	1	3	1	1.000000	0.9	1	1.000000	0.9
бегущий по льду	5	2	2	5	5	5	4	2	4	4	1	0.333333	0.7	1	1.000000	0.9
человек кусает собаку	5	3	2	3	2	3	2	2	2	2	1	0.666667	0.4	1	0.333333	0.2
военный коммунизм	5	2	4	2	4	3	3	3	2	2	1	0.666667	0.6	1	0.666667	0.6
петропавловская крепость	4	2	3	4	2	2	2	3	3	2	1	0.666667	0.5	0	0.333333	0.4
эрмитаж	3	5	2	4	2	2	2	3	3	2	1	0.666667	0.5	1	0.666667	0.5
ссср	5	2	2	2	3	2	3	2	3	2	1	0.333333	0.4	1	0.333333	0.4





Как видно, все-таки существенного изменения результатов ранжирования не произошло, немного увеличилось значение среднего по DCG, но в незначительной степени.

Выводы

По приведенным оценкам качества поиска, можно заметить прогресс в результатах, однако есть некоторые проблемы связанные с ранжированием (например поменялась позиция текста с полным вхождением). Также стоит заметить ошибки с исчезновением документа по оценке Precision – причина в том, что теперь есть явный перекос по оценке tf, даже с учетом нормализации. То есть мы сталкиваемся с тем, что частотность берет вверх. Это можно исправить путем выделения главных слов, строя например синтаксическое дерево анализа и последующим появлением штрафной функции.