МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)
Факультет информационных технологий и прикладной математики
Кафедра вычислительной математики и программирования

Отчет по лабораторной работе №3
«Лемматизация»
по курсу
«Обработка текстов на естественном языке»

Группа: М80-108М-17
Выполнил: Забарин Н.И.
Преподаватель: А.Л. Калинин

Москва, 2018

Задание

Добавить в созданную поисковую систему 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?

Решение

Для лемматизации был написан отдельный модуль использующий готовые решения по лемматизации/стеммингу для Python3 NLTK. Используются следующие алгоритмы:

- ISRI Stemmer,
- Lancaster Stemmer,
- Porter Stemmer,
- Snowball Stemmer,
- WordNet Lemmatizer.

Выбор алгоритма проходил следующим образом, тестовая выборка из 10000 документов была проиндексирована без лемматизации и с каждым алгоритмом по очереди. Сжатие данных было отключено, для большей наглядности.

- Без лемматизации, 809 Kb/s, 57Mb.
- WordNet, 444 Kb/s, 55 Mb.
- ISRI, 350 Kb/s, 57 Mb.
- Другие алгоритмы давали скорость ниже 200 Kb/s и не давали выигрыша по памяти, то есть почти не снижали размер хедера(число слов).

Что бы ускорить индексацию был написан простенький кэш на основе словаря, результат работы с кэшом:

- WordNet + cache 1e4, 655 Kb/s.
- ISRI + cache 1e4, 601 Kb/s.
- WordNet + cache 1e5, 732 Kb/s.
- ISRI + cache 1e5, 735 Kb/s.

Несмотря на то что оба алгоритма имеют примерно одинаковую скорость WordNet уменьшает размер словаря на 7% против 1% у ISRI, так что я остановился на WordNet.

Качество поиска

Проанализировав работу модуля было замечено что в основном изменения затронули всевозможные словоформы, алгоритм обрезает концы слова -s, -ing, -ed и подобные. При этом разные неправильные глаголы или исключительные случаи превосходных степеней прилагательных остаются без изменения.

Сравнение на наборе запросов из лабораторной работы №8, не дало никаких отличий, минимальные изменения порядка статей в поисковой выдачи, более высокие TF индексы у слов.

Выводы

В целом поиск стал точнее, можно найти запросы в которых присутствуют названия с измененной формой слова, при поиске такие формы будут обрезаны и нужная статья имеет шансы затеряться среди других результатов.

Эту проблему можно решить добавив поиск по префиксу и отключив лемматизацию на этапе индексации и используя цитатный поиск.