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: 80-106М
Выполнил: Демин И.А.
Преподаватель: Калинин А.Л.

Москва, 2019

Задание

Добавить в созданную поисковую систему (ЛР 1-8 по курсу «Информационный поиск») лемматизацию. В простейшем случае, это просто поиск без учёта словоформ. Лемматизацию можно добавлять на этапе индексации, можно на этапе выполнения поискового запроса.

Выполнение лабораторной работы

Для выполнения лабораторной работы был использован простой вариант без учета словоформ. Использована Python обертка для Yandex Mystem 3.1. Программа MyStem производит морфологический анализ текста на русском языке. Добавить построение лемм можно на этапе получение хеша от слова.

Использование лемм должно улучшить качество поиска, увеличить количество найденных статей и сократить число слов в обратном индексе, однако увеличить время построенния индекса.

Были испробованы другие библиотеки для лемматизации на Python, но самое быстрое решение предоставило решение от Яндекса.

Примером того, что работает лемматизация, будет одинаковые результаты для запросов, в которых одинаковые слова, но в разных формах.

```
(base) ivan@ivan-G5:~/study/search/search_MAI/l11_lemms$ python l11.py
Запрос лёгкие спортам тренеров
Id: 0.Заголовок: Каратаев,_Николай_Дмитриевич. Url: https://ru.wikipedia.org/wiki/Каратаев,_Николай_Дмитриевич
Id: 2320.Заголовок: Степаненков,_Виктор_Александрович. Url: https://ru.wikipedia.org/wiki/Степаненков,_Виктор_Александрович
Id: 20.Заголовок: Цыплаков,_Александр_Викторович. Url: https://ru.wikipedia.org/wiki/Цыплаков,_Александр_Викторович
Id: 2619.Заголовок: Хайкин,_Владислав_Евгеньевич. Url: https://ru.wikipedia.org/wiki/Хайкин,_Владислав_Евгеньевич
Id: 283.Заголовок: Хозяшева,_Светлана_Анатольевна. Url: https://ru.wikipedia.org/wiki/Хозяшева,_Светлана_Анатольевна
Articles count: 5
get_search_res_for_quotes works 0:00:00.453061
(base) ivan@ivan-G5:~/study/search/search_MAI/l11_lemms$ python l11.py
Запрос: 'лёгкой спорта тренер'
Id: 0.Заголовок: Каратаев,_Николай_Дмитриевич. Url: https://ru.wikipedia.org/wiki/Каратаев,_Николай_Дмитриевич
Id: 2320.Заголовок: Степаненков,_Виктор_Александрович. Url: https://ru.wikipedia.org/wiki/Степаненков,_Виктор_Александрович
Id: 20.Заголовок: Цыплаков,_Александр_Викторович. Url: https://ru.wikipedia.org/wiki/Цыплаков,_Александр_Викторович
Id: 2619.Заголовок: Хайкин,_Владислав_Евгеньевич. Url: https://ru.wikipedia.org/wiki/Хайкин,_Владислав_Евгеньевич
Id: 283.Заголовок: Хозяшева,_Светлана_Анатольевна. Url: https://ru.wikipedia.org/wiki/Хозяшева,_Светлана_Анатольевна
Articles count: 5
get_search_res_for_quotes works 0:00:00.434582
```

Результаты будут сравниваться с цитатным поиском ЛР6. Результат:

"спорт экспресс" — 0.6

"виды спорта" — 1

"активный отдых" — 0.8

"профессиональный бокс" — 1

"боевые искусства" — 1

Средняя точность — 0.88

Результаты поиска с лемматизацией слов:

"спорт экспресс" — 0.6

"виды спорта" — 0.8

"активный отдых" — 0.8

"профессиональный бокс" — 1

"боевые искусства" — 1

Средняя точность — 0.84

Предположение о том, что качество поиска было опровергнуто — оно немного ухудшилось. Объяснение этому может быть в том, что когда пытаются найти цитаты, форма слова все-таки играет роль. Пример тому запрос «виды спорта», он был преобразован в «вид спорт» и тем самым увеличилась поисковая выдача, следовательно результаты стали менее точными и скорее всего они и попали в контрольное множество. Могу предположить, что если тестовые запросы были более разнообразны на формы слов (здесь почти все запросы в начальной форме), то качество поиска могло еще уменьшиться.

Вывод

В ходе лабораторной работы я познакомился с леммизацией и зачем она применяется в построении поисковых систем. Также познакомился с системой леммизации от Яндекса.