### 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

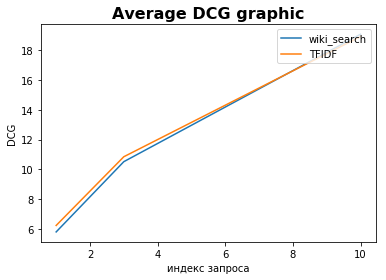
Качество поиска

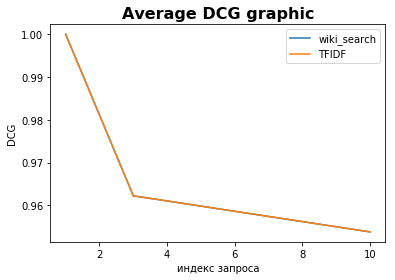
Оценка качества поиска производилась с помощью 30 запросов, которые использовались в лабораторных работах по булевому поиску и TF-IDF ранжированию. Сравнение производилось с TF-Idf ранжированной выдачей (все значение с припиской old)

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

Изображение выглядит как текст

Автоматически созданное описание





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.