### 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

|  |  |  |
| --- | --- | --- |
| ***Запрос*** | ***P@1*** | ***Морфология P@1*** |
| [министерство правды] | 0 | 0 |
| [лунное затмение] | 1 | 1 |
| [чемпионат мира футболу] | 0 | 1 |
| [ракета в космосе] | 1 | 0 |
| [советского государства] | 1 | 1 |
| [сибирской платформы] | 1 | 1(изменился порядок) |
| [Википедия] | 1 | 1 |
| [биткоин] | 0 | 0 |
| [что где когда] | 1 | 1 |
| [товарищ прокурора] | 0 | 1 |

|  |  |  |
| --- | --- | --- |
| ***Запрос*** | ***P@10*** | ***Морфология P@10*** |
| [чемпионат мира футболу] | *0* | *1* |
| [министерство правды] | *0.4* | *0.6* |
| [лунное затмение] | 0,3 | *1* |
| [машина в космосе] | 0.3 | 0 |
| [советского государства] | 1 | 1 |
| [сибирской платформы] | 0.5 | 0.6(появилась ошибка в ранжировании) |
| [Википедия] | 0.8 | 0.8 |
| [биткоин] | 0 | 0 |
| [что где когда] | 0.2 | 0.2 |
| [товарищ прокурора] | 0.3 | 0.8 |

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.