

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

Отчет по лабораторной работе №4
«Построение сниппетов»
по курсу
«Обработка текстов на естественном языке»

Группа: 80-106М

Выполнил: Демин И.А.

Преподаватель: Калинин А.Л.

Москва, 2019

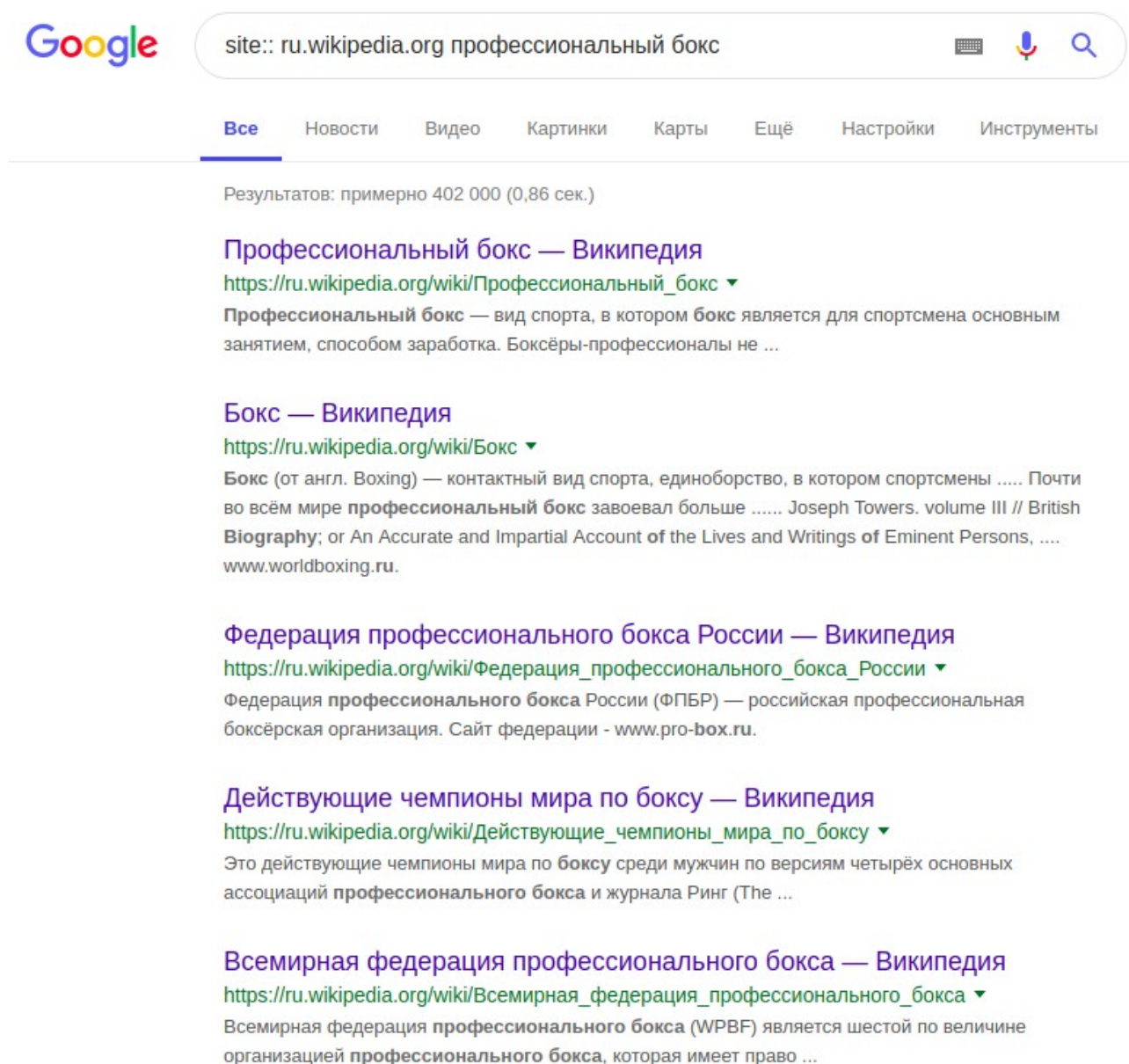
Задание

Необходимо добавить в поисковую систему построение цитат (сниппетов), реферирование документов, найденных по запросу.

Выполнение лабораторной работы

Так как эта работа заключалась исключительно в обработке уже готовой выдачи, то нужно было рассмотреть варианты как преподнести информацию. Были изучены сниппеты, предоставляемые большими поисковиками, с ограничением по википедии.

Вот пример запроса «профессиональный бокс»



The screenshot shows a Google search interface. The search bar contains the text "site:: ru.wikipedia.org профессиональный бокс". Below the search bar, there are tabs for "Все", "Новости", "Видео", "Картинки", "Карты", "Ещё", "Настройки", and "Инструменты". The search results are displayed below the tabs, showing the number of results and the time taken to process the query. The results are listed in a vertical column, each starting with a title in purple, followed by a green URL, and then a brief description in black text.

Google

site:: ru.wikipedia.org профессиональный бокс

Все Новости Видео Картинки Карты Ещё Настройки Инструменты

Результатов: примерно 402 000 (0,86 сек.)

Профессиональный бокс — Википедия
https://ru.wikipedia.org/wiki/Профессиональный_бокс ▼
Профессиональный бокс — вид спорта, в котором бокс является для спортсмена основным занятием, способом заработка. Боксёры-профессионалы не ...

Бокс — Википедия
<https://ru.wikipedia.org/wiki/Бокс> ▼
Бокс (от англ. Boxing) — контактный вид спорта, единоборство, в котором спортсмены Почти во всём мире профессиональный бокс завоевал больше Joseph Towers. volume III // British Biography; or An Accurate and Impartial Account of the Lives and Writings of Eminent Persons, www.worldboxing.ru.

Федерация профессионального бокса России — Википедия
https://ru.wikipedia.org/wiki/Федерация_профессионального_бокса_России ▼
Федерация профессионального бокса России (ФПБР) — российская профессиональная боксёрская организация. Сайт федерации - www.pro-box.ru.

Действующие чемпионы мира по боксу — Википедия
https://ru.wikipedia.org/wiki/Действующие_чемпионы_мира_по_боксу ▼
Это действующие чемпионы мира по боксу среди мужчин по версиям четырёх основных ассоциаций профессионального бокса и журнала Ринг (The ...

Всемирная федерация профессионального бокса — Википедия
https://ru.wikipedia.org/wiki/Всемирная_федерация_профессионального_бокса ▼
Всемирная федерация профессионального бокса (WPBF) является шестой по величине организацией профессионального бокса, которая имеет право ...

Как видно, сниппет представляет из себя заголовок-гиперссылку, ссылку, текст-описание статьи. Если говорить про википедию, то текст-описание всегда соответствует вступлению статьи, которое является неким summary для статьи.

Также рассматривался следующий вариант - рассматривать слова или символы в некотором диапазоне слов относительно нашего целевого слова или слов. Первая проблема заключалась в том, что не ясно, как работать, когда нужных слов несколько, даже если применять зонный поиск и разделять статью хотя бы на заголовок и статью, то опять же, не ясно какой блок текста выводить. Вторая — корректное количество слов или символов, что бы предлагаемый текст-описание был релевантен.

После некоторых раздумий было принято решение брать начало статьи в качестве блока текста. Однако стоит сказать, что данный метод является оптимальным только для википедии из-за структуры статей.

Пример работы

```
(base) ivan@ivan-G5:~/study/search/search_MAI/l12_snippets$ python l12.py
Запрос: мастер спорта по самбо/1
Id: 1440.Заголовок: Амагов,_Адлан_Майрбекович. Url: https://ru.wikipedia.org/wiki/Амагов,\_Адлан\_Майрбекович
Адла'н Майрбе'кович Ама'гов (30 октября 1986 года, Грозный, Чечено-Ингушская АССР, РСФСР, СССР) – боец смешанных единоборств, о
рта по рукопашному бою, самбо и комплексному единоборству, первый чеченец, выступающий в UFC.

Id: 2592.Заголовок: Калмыцкая_борьба. Url: https://ru.wikipedia.org/wiki/Калмыцкая\_борьба
Калмы'цкая борьба' или Беки' барилда'н (калм. Беки барилдан) – национальная калмыцкая борьба, которая является одним из элемент
но проводятся республиканские соревнования по калмыцкой борьбе.

Id: 1795.Заголовок: Козлов,_Геннадий_Андреевич. Url: https://ru.wikipedia.org/wiki/Козлов,\_Геннадий\_Андреевич
Козлов Геннадий Андреевич (1943) – российский спортсмен мастер спорта по самбо (1969), заслуженный тренер России (1996). Среди
В. Пудиков, чемпион Спартакиады народов СССР М. Фурсов, чемпионы РСФСР Н. Задорин и А. Самороков.

Id: 390.Заголовок: Жданов,_Ирик_Гатиятулович. Url: https://ru.wikipedia.org/wiki/Жданов,\_Ирик\_Гатиятулович
Ирик Гатиятулович Жданов (род. 20 октября 1934) – тренер-преподаватель по боксу, почетный гражданин города Оренбурга. Взял себ

Id: 1354.Заголовок: Михайлов,_Владимир_Николаевич. Url: https://ru.wikipedia.org/wiki/Михайлов,\_Владимир\_Николаевич
Владимир Николаевич Михайлов (род. 30 ноября 1954) – советский и российский спортсмен (самбо и дзюдо), тренер. Заслуженный тре

Id: 47.Заголовок: Костылева,_Наталья_Геннадьевна. Url: https://ru.wikipedia.org/wiki/Костылева,\_Наталья\_Геннадьевна
Наталья Геннадьевна Костылева (род. 26 сентября 1969 года) – советская и российская самбистка, чемпионка Мира, Европы и России
борьбе дзюдо, город Краснокамск. Почётный гражданин города Краснокамска.
```

Вывод

В ходе лабораторной работы я познакомился с построением сниппетов и добавил их в свою поисковую систему.