### Московский авиационный институт

**Факультет прикладной математики и физики**

**Лабораторная работа №4**

**по курсу:**

**«Обработка естественно-языковых текстов»**

**по теме:**

**«Построение сниппетов»**

**2 семестр**

Студент: Ахмед С. Х.

Преподаватель: Калинин А. Л.

Группа: 8О-106М

**Москва, 2019 г**

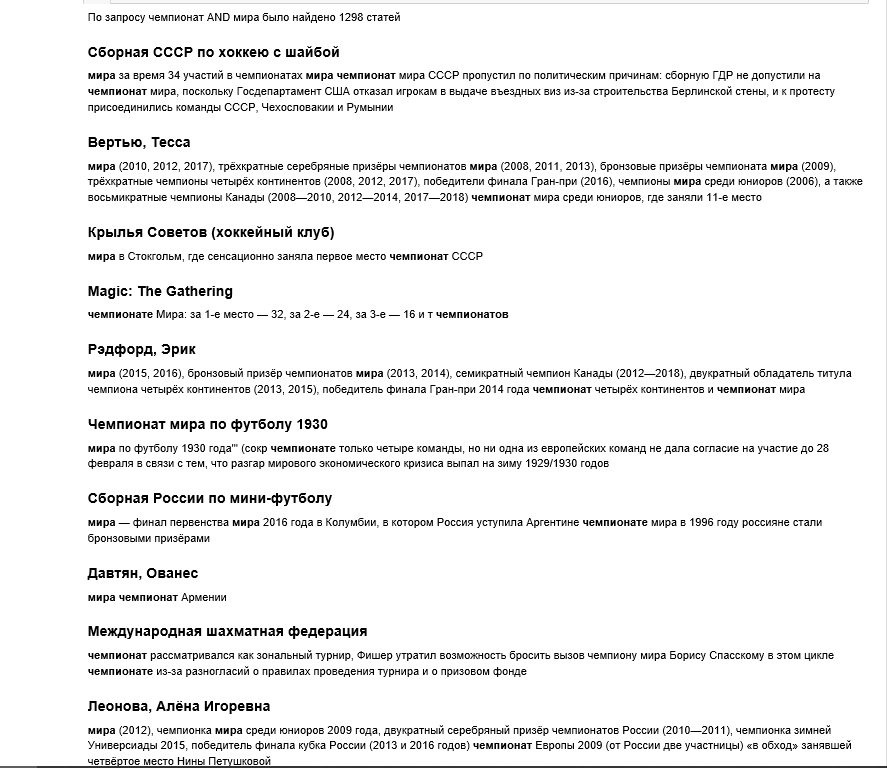
Постановка задачи

Необходимо добавить в поисковую систему построение цитат (сниппетов), реферирование документов, найденных по запросу. Сниппеты должны содержать слова запроса и давать пользователю представление о том, насколько документ отвечает поисковому запросу. Длина сниппета должна быть ограничена двумя-тремя строчками. В отчёте нужно привести описание алгоритма построения сниппетов, примеры.

Методика решения

1. Для построения сниппетов потребовалась служебная информация. Поэтому было решено создать индекс с позициями начала документа
2. Сам алгоритм построения сниппетов следующий:
3. Для каждого термина запроса ищется первое предложение со вхождением данного слова
4. Полученные предложения(их позиции) сортируется по порядку следования предложений в документе
5. Формируем сниппет. Обрезаем его размер 300 символами.
6. Выдаем клиенту

Примеры



Изображение выглядит как снимок экрана

Автоматически созданное описание

Выводы

Примеры выдачи здесь служат для того, чтобы показать саму выдачу (ключевые слова выделены). А также увидеть недочет. Есть вероятность, что длина полученного текста очень мала, а при парсинге дампа википедии я убрал часть служебной информации связанной с со ссылками, то выдача выглядит урезанной

Выявлены следующие недочеты:

1. Длина текста в некоторых моментах мала
2. Ошибки форматирования текста при парсинге(остались некоторые служебные символы)
3. Не учитывается момент связанный с tf оценкой термина.

Построенные сниппеты отображают содержание статьи, однако не учитывают значимость термина для запроса и документа