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:
   a) Для каждого термина запроса ищется первое предложение со вхождением данного слова
   b) Полученные предложения(их позиции) сортируется по порядку следования предложений в документе
   c) Формируем сниппет. Обрезаем его размер 300 символами.
   d) Выдаем клиенту

Примеры

По запросу чемпионат AND мира было найдено 1298 статей

Сборная СССР по хоккею с шайбой

мира за время 34 участий в чемпионатах **мира чемпионат** мира СССР пропустил по политическим причинам: сборную ГДР не допустили на **чемпионат** мира, поскольку Госдепартамент США отказал игрокам в выдаче въездных виз из-за строительства Берлинской стены, и к протесту присоединились команды СССР, Чехословакии и Румынии

Вертью, Тесса

мира (2010, 2012, 2017), трёхкратные серебряные призёры чемпионатов **мира** (2008, 2011, 2013), бронзовые призёры чемпионата **мира** (2009), трёхкратные чемпионы четырёх континентов (2008, 2012, 2017), победители финала Гран-при (2016), чемпионы **мира** среди юниоров (2006), а также восьмикратные чемпионы Канады (2008—2010, 2012—2014, 2017—2018) **чемпионат** мира среди юниоров, где заняли 11-е место

Крылья Советов (хоккейный клуб)

мира в Стокгольм, где сенсационно заняла первое место **чемпионат** СССР

Magic: The Gathering

чемпионате Мира: за 1-е место — 32, за 2-е — 24, за 3-е — 16 и т **чемпионатов**

Рэдфорд, Эрик

мира (2015, 2016), бронзовый призёр чемпионатов **мира** (2013, 2014), семикратный чемпион Канады (2012—2018), двукратный обладатель титула чемпиона четырёх континентов (2013, 2015), победитель финала Гран-при 2014 года **чемпионат** четырёх континентов и **чемпионат** мира

Чемпионат мира по футболу 1930

мира по футболу 1930 года"" (сокр **чемпионате** только четыре команды, но ни одна из европейских команд не дала согласие на участие до 28 февраля в связи с тем, что разгар мирового экономического кризиса выпал на зиму 1929/1930 годов

Сборная России по мини-футболу

мира — финал первенства **мира** 2016 года в Колумбии, в котором Россия уступила Аргентине **чемпионате** мира в 1996 году россияне стали бронзовыми призёрами

Давтян, Ованес

мира чемпионат Армении

Международная шахматная федерация

чемпионат рассматривался как зональный турнир, Фишер утратил возможность бросить вызов чемпиону мира Борису Спасскому в этом цикле **чемпионате** из-за разногласий о правилах проведения турнира и о призовом фонде

Леонова, Алёна Игоревна

мира (2012), чемпионка **мира** среди юниоров 2009 года, двукратный серебряный призёр чемпионатов России (2010—2011), чемпионка зимней Универсиады 2015, победитель финала кубка России (2013 и 2016 годов) **чемпионат** Европы 2009 (от России две участницы) «в обход» занявшей четвёртое место Нины Петушковой

По запросу крымский AND мост было найдено 17 статей

Транспорт Крыма

но вскоре этот **мост** был разрушен транспорт ==== Файл:Skoda **крымский** троллейбусŠkoda 9Tr Крымский троллейбус представляет собой уникальную междугородную систему, связывающую Симферополь с курортами Южного берега Крыма и включает в себя самый длинный в мире междугородный троллейбусный маршрут Симферополь — Алушта — Ялта протяжённостью 96 км

5-й кавалерийский корпус (1-го формирования)

непроезжее состояние железнодорожный **мост** и заминировали **мост** для колёсного транспорта у села Фэлчиу 24:00 наведение понтонного **моста** через Цареградское гирло Днестровского лимана продолжалось при сильном шторме

Смородина (река)

иногда — двух-трех **мостов** или перевоза, у которых богатырь сражается со змеем, чудом-юдом и т

Экономика Крыма

Симферопольской ТЭС и **моста** через Керченский пролив

Добровское сельское поселение

магистраль: автодорога Чонгарский **мост** — Джанкой — Симферополь — Алушта — Ялта (по украинской классификации — автодорога Автодорога М-18)

Волгоград

Этим воспользовался сераскир **крымского** ханства Бахти Герай, организовавший Кубанский погром в августе 1717 г Османской империи

Польско-шведские войны

солдаты быстро навели **мост** через речку к болотистому участку берега, где русские не ждали опасности Россию, украинских казаков, **крымских** татар, Османскую империю и даже протестантских князей Германии

Ерёмин, Владимир Аркадьевич

свете событий вокруг **крымского** кризиса, вместе с рядом других известных деятелей науки и культуры России выразил своё несогласие с политикой российской власти в Крыму

Революция и Гражданская война на Украине

Черноморского флота с **крымского** побережья отплыли в Общая численность покинувших Крым составила около 150 тысяч человек севера путём взрыва **мостов** и разборкой путей и на юго-запад — путём захвата железнодорожных узлов

Каушаны

линии построили 204 **моста** общей длиной 2342 м, проложили 88 труб

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