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: М80-108М-17
Выполнил: Забарин Н.И.
Преподаватель: А.Л. Калинин

Москва, 2018

Задание

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 отчете нужно привести описание алгоритма построения сниппетов, примеры.

Решение

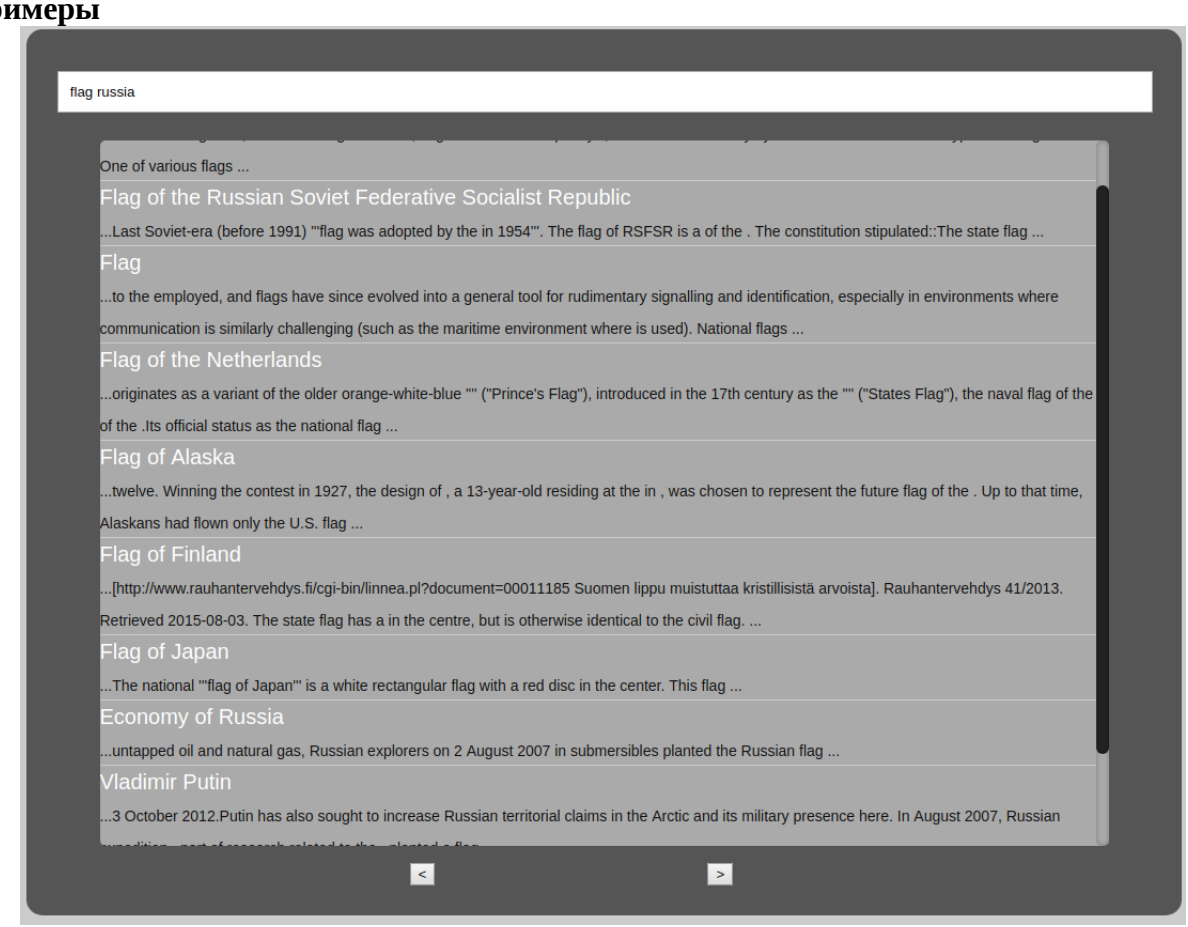
Генерировать сниппеты было решено динамически, так как эта группа алгоритмов лучше отражает причины по которым документ попал в выборку.

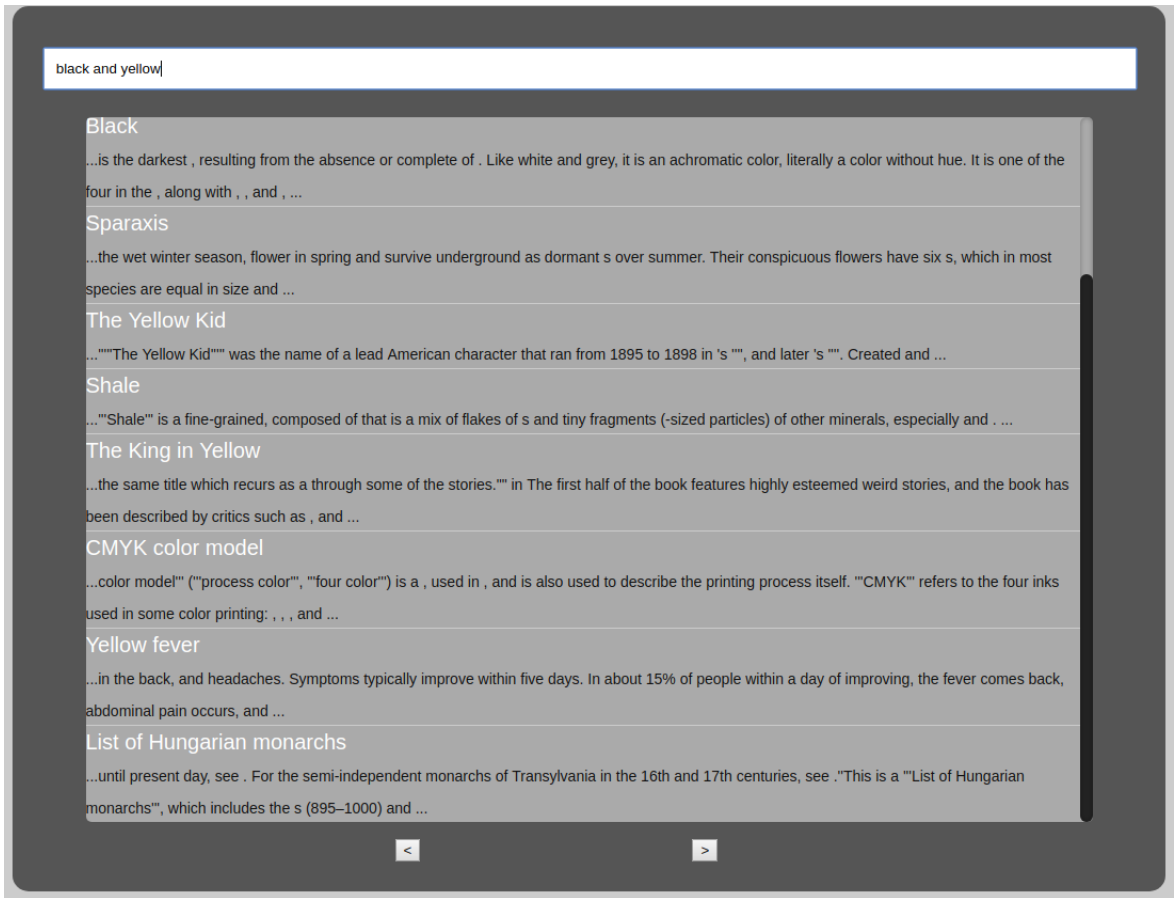
Для реализации было внесено несколько исправлений в итераторы, теперь вместе с найдены документом возвращаются и слова запроса которые присутствуют в документе. Используя эту информацию и текст статьи будем искать окно с ограниченной длиной и максимальным числом слов из запроса. Код ищущий сниппет:

```
def extract_snippet(text, q):
    l, r = 0, 1
    curlen = 0
    curans = 0
    spt = []
    term = ''

    ans = []
    mx = 0

    def check_term(term):
        for i in q:
            if term.startswith(i):
                return True
        return False

    i = 0
    while i < len(text):
        if text[i] in stop_sym:
            while text[i] in stop_sym:
                term += text[i]
                i += 1

            curlen += len(term)
            spt.append(term)
            if curlen > SNIPPET_MAX_LEN:
                if check_term(spt[0]):
                    curans -= 1
                spt = spt[1:]
            if check_term(term):
                curans += 1

            if mx <= curans:
                mx = len(ans)-1
                ans = spt[::]

            term = ''
        else:
            term += text[i]
            i += 1

    spt = ''.join(ans)
    if spt:
        return SNIPPET_FMTSTR.format(spt.sprip())
    return ''
```

Так же была немного переделана страничка отображения результатов, добавлены сниппеты.

Примеры


flag russia
One of various flags ...
Flag of the Russian Soviet Federative Socialist Republic
...Last Soviet-era (before 1991) '"flag was adopted by the in 1954"'. The flag of RSFSR is a of the . The constitution stipulated::The state flag ...
Flag
...to the employed, and flags have since evolved into a general tool for rudimentary signalling and identification, especially in environments where communication is similarly challenging (such as the maritime environment where is used). National flags ...
Flag of the Netherlands
...originates as a variant of the older orange-white-blue "" ("Prince's Flag"), introduced in the 17th century as the "" ("States Flag"), the naval flag of the of the .Its official status as the national flag ...
Flag of Alaska
...twelve. Winning the contest in 1927, the design of , a 13-year-old residing at the in , was chosen to represent the future flag of the . Up to that time, Alaskans had flown only the U.S. flag ...
Flag of Finland
...[http://www.rauhantervehdys.fi/cgi-bin/linnea.pl?document=00011185 Suomen lippu muistuttaa kristillisistä arvoista]. Rauhantervehdys 41/2013. Retrieved 2015-08-03. The state flag has a in the centre, but is otherwise identical to the civil flag. ...
Flag of Japan
...The national '"flag of Japan"' is a white rectangular flag with a red disc in the center. This flag ...
Economy of Russia
...untapped oil and natural gas, Russian explorers on 2 August 2007 in submersibles planted the Russian flag ...
Vladimir Putin
...3 October 2012.Putin has also sought to increase Russian territorial claims in the Arctic and its military presence here. In August 2007, Russian
<
>


black and yellow
Black
...is the darkest , resulting from the absence or complete of . Like white and grey, it is an achromatic color, literally a color without hue. It is one of the four in the , along with , , and , ...
Sparaxis
...the wet winter season, flower in spring and survive underground as dormant s over summer. Their conspicuous flowers have six s, which in most species are equal in size and ...
The Yellow Kid
...'''The Yellow Kid''' was the name of a lead American character that ran from 1895 to 1898 in 's "", and later 's "". Created and ...
Shale
...'"Shale"' is a fine-grained, composed of that is a mix of flakes of s and tiny fragments (-sized particles) of other minerals, especially and . ...
The King in Yellow
...the same title which recurs as a through some of the stories."" in The first half of the book features highly esteemed weird stories, and the book has been described by critics such as , and ...
CMYK color model
...color model"' ("'process color"', "'four color"') is a , used in , and is also used to describe the printing process itself. "'CMYK"' refers to the four inks used in some color printing: , , , and ...
Yellow fever
...in the back, and headaches. Symptoms typically improve within five days. In about 15% of people within a day of improving, the fever comes back, abdominal pain occurs, and ...
List of Hungarian monarchs
...until present day, see . For the semi-independent monarchs of Transylvania in the 16th and 17th centuries, see ."This is a "'List of Hungarian monarchs"', which includes the s (895–1000) and ...
<
>

Выводы

Построение сниппетов не очень сложная, но крайне полезная задача. Они позволяют довольно точно определять релевантен ли результат поиска и нужно ли переходить к самому документу.