Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»
Кафедра ИУ5 «Системы обработки информации и управления»

Курс «Парадигмы и конструкции языков программирования»

Отчет по домашней работе

Выполнил:
студент группы ИУ5-34Б:
Суслов Дмитрий Сергеевич
Подпись и дата:

Проверил:
преподаватель каф. ИУ5
Гапанюк Ю. Е.
Подпись и дата:

Москва, 2023 г.

Задание:

Разработаем поисковую систему. Поиск и ранжирование будет производится на основе массива новостных сводок, найденного в интернете. Обозначим этапы работы:

1. Предобработка массива данных (фильтрация стоп-слов, знаков препинания, приведение к нижнему регистру, стемминг, предварительный расчёт метрик TF-IDF) и сохранение данных в удобном формате .pkl
2. Написание алгоритма ранжирование по запросу
3. Реализация серверной и фронт- частей на Flask и HTML + CSS

Код предобработки хранится в ноутбуке *preproc.ipynb*
Функции расчёта score'а находятся в файле *my_search.py*

```
import pandas as pd
import re
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer
import pymorphy2
import gensim.downloader as download_api
from sklearn.feature_extraction.text import TfidfVectorizer
import itertools
import pickle


class Document:
    def __init__(self, title, text, id):
        self.title = title
        self.text = text
        self.id = id

    def format(self, query):
        return [self.title, self.text + ' ...']



def get_word_info(word):
    p = morph.parse(word)[0]
    normal = p.normal_form
    pos = p.tag.POS
    if pos:
        if 'ADJ' in pos:
            pos1 = 'ADJ'
            return (normal, pos1)
    return (normal, pos)


def upd_str(text, stem=True):
    text = text.lower()
    words = re.findall(r'\w+', text)
    filter_words = list(filter(lambda x: x not in stop_words,
words))
    if filter_words:
```

```
        if stem:
            new = list(map(lambda x: snowball.stem(x),
filter_words))
            return ' '.join(new)
        new = list(map(lambda x: get_word_info(x),
filter_words))
        print(new)
        return new
    return ''


def build_index():
    global index, Y, TF_IDF_vocab, russian_model, documents,
stop_words, snowball, valid_types, morph, k1, k2

    with open('inv_index.pkl', 'rb') as file:
        index = pickle.load(file)

    with open('TF_IDF.pkl', 'rb') as file:
        Y = pickle.load(file)

    with open('TF_IDF_vocab.pkl', 'rb') as file:
        TF_IDF_vocab = pickle.load(file)

    with open('documents.pkl', 'rb') as file:
        documents = pickle.load(file)

    russian_model = download_api.load('word2vec-ruscorpora-300')
    stop_words = stopwords.words('russian')
    snowball = SnowballStemmer(language='russian')
    valid_types = ['NOUN', 'VERB', 'ADJ']
    morph = pymorphy2.MorphAnalyzer()
    k1 = 1
    k2 = 1


def score(query, document, k=(0.5, 0.7)):
    # взвешенная сумма семантического сходства заголовка с
запросом и среднего значения tf_idf встречающихся в тексте слов
из запроса

    k1 = k[0]
    k2 = k[1]
    if not stem_query:
        return 0

    sims = {}
    tfidfs = []
    for word1 in list(filter(lambda x: str(x[1]) in valid_types,
upd_query)):
        sims[word1[0]] = 0
        for word2 in list(filter(lambda x: str(x[1]) in
valid_types, upd_str(document.title, stem=False))):
```

```
            if '_'.join(list(word1)) in russian_model and
'_'.join(list(word2)) in russian_model:
                sims[word1[0]] = max(sims[word1[0]],
russian_model.similarity('_'.join(list(word1)),
'_'.join(list(word2))))

    for word in stem_query:
        if word in TF_IDF_vocab:
            idx = TF_IDF_vocab[word]
            tf_idf = Y[document.id].toarray()[0][idx]
            if tf_idf > 0:
                tfidfs.append(tf_idf)

    return sum(sims.values()) / (len(sims) + 1) * k1 +
sum(tfidfs) * k2 / (len(tfidfs) + 1)



def retrieve(query):
    global upd_query, stem_query

    upd_query = upd_str(query, stem=False)
    stem_query = upd_str(query).split()
    stem_query = list(filter(lambda x: x in index, stem_query))

    if not stem_query:
        return documents[:500]

    most_rel_idxs = []
    f = False

    # перебираем всевозможные комбинации слов в запросе, начиная
с самой длинной, пересекая индексы документов, содержащих слова
в комбинации
    # прекращаем поиск, если нашлось хотя бы 500 документов
    for i in range(len(stem_query), 0, -1):
        for comb in itertools.combinations(stem_query, i):
            valid_idxs = set()
            for word in comb:
                if not valid_idxs:
                    valid_idxs = set(index[word])
                valid_idxs = valid_idxs & set(index[word])
            most_rel_idxs.extend(list(valid_idxs -
set(most_rel_idxs)))

            if len(most_rel_idxs) > 500:
                f = True
                break
        if f:
            break


    candidates = []
```

```
    for idx in most_rel_idxs:
        candidates.append(documents[idx])

    return candidates[:500]
```

Серверная часть описана в файле *server.py*

```
from flask import Flask, render_template, request
from my_search import score, retrieve, build_index
from time import time

app = Flask(__name__, template_folder='.')
build_index()


@app.route('/', methods=['GET'])
def index():
    start_time = time()
    query = request.args.get('query')
    if query is None:
        query = ''
    documents = retrieve(query)
    documents = sorted(documents, key=lambda doc: -score(query, doc))
    results = [doc.format(query)+['%.2f' % score(query, doc)] for doc in documents]
    return render_template(
        'index.html',
        time="%.2f" % (time()-start_time),
        query=query,
        search_engine_name='Гугол',
        results=results
    )


if __name__ == '__main__':
    app.run(debug=True, host='127.0.0.1', port=80)
```

Разметка и подбор коэффициентов для взвешенного score'а находятся в файлах *coefs_setting.ipynb* и *markup.ipynb*

Поисковушка

Как готовить борщ

Search

Found 500 documents in 1.71 seconds.

Фигуристка Медведева рассказала о советующих учиться варить борщи фанатах

0.31

Чемпионка мира по фигурному катанию россиянка Евгения Медведева рассказала о фанатах-недоброжелателях, советующих ей научиться готовить. Ее слова приводит сайт Международного союза конькобежцев (ISU). «Всегда есть те, кто хочет выделиться, не каждому нравится моя программа или то, что я делаю, но доброжелательных людей больше. Иногда я просто смеюсь над глупыми комментариями. Как-то мне написали: почему эта школьница занимается этим, ей нужно пойти и научиться варить борщ», — сказала Медведева. Она добавила, что, несмотря на завоеванные ей титулы, в ее отношениях с друзьями и другими окружающими людьми ничего не изменилось. «Я рада, что все важные для меня люди остались со мной и относятся ко мне просто как к Жене», — отметила 17-летняя спортсменка. Медведева — действующая чемпионка мира и Европы. Она дважды становилась сильнейшей фигуристкой континента. Кроме того россиянка выиграла два последних финала Гран-при. ...

Медведеву взбесили жаждущие борща мужчины

0.29

Российская фигуристка Евгения Медведева высказалась о своем отношении к мужчинам, которые требуют, чтобы женщины постоянно готовили. Слова спортсменки приводит «Спорт-Экспресс» со ссылкой на эфир Love Radio. Медведева призналась, что ее бесит такое отношение к женщине. «Как человек деятельный, я не смогу каждый день сидеть дома и жарить яичницу, борщи варить, хинкали лепить», — заметила фигуристка. Также она сказала, что женщина обязана содержать дом, но в то же время ей нужно иметь возможности для работы. Спортсменка также рассказала, что ей приходилось

Поисковушка

Москва снег

Search

Found 500 documents in 2.61 seconds.

Власти Москвы предрекли Нью-Йорку коллапс из-за снега

0.49

Если бы снегопад, который обрушился на Москву, произошел в Нью-Йорке или Хельсинки, города были бы парализованы. Такое мнение высказал заместитель мэра Москвы по вопросам ЖКХ и благоустройства Петр Бирюков. Его слова во вторник, 6 февраля, приводит агентство «Москва». «Там если выпадет 5 сантиметров снега — чрезвычайная ситуация. Закрываются учреждения, и все ждут, когда снег растает. С любой столицей мира сравнивать: такого отношения к жителям своего города со стороны властей, как в Москве, нет нигде», — заключил чиновник. Бирюков также подчеркнул, что уборка снега без специальных реагентов в настоящее время невозможна. «Они экологически безопасны, и это все подтверждено федеральными нормативами, федеральной технологией, инструкцией», — заверил он. Снегопад, который синоптики назвали самым сильным за историю метеонаблюдений, обрушился на Москву 3 февраля. По нормативам, на уборку выпавшего снега в городе потребуется больше недели. ...

В Москве за сутки выпало девять сантиметров снега

0.48

Прирост снежного покрова в Москве за последние сутки составил девять сантиметров. Об этом сообщает "Интерфакс" со ссылкой на Гидрометебюро Москвы и Московской области. По данным синоптиков, в пятницу, 25 января, к вечеру в Москве выпало семь сантиметров снега, а еще два - в ночь с пятницы на субботу. Всего за последнюю неделю в столице выпало 27 сантиметров снега, что составляет около двух третей от январской нормы. Метеорологи ожидают образование гололедицы, наледи, снежного наката и сосулек. Всего снегоуборочные бригады вывезли с улиц Москвы 330 тысяч