Минобрнауки России

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Волгоградский государственный технический университет»

Факультет\_\_\_Электронной вычислительной техники\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Направление (специальность) \_\_\_Системы автоматизированного\_\_\_\_\_\_\_\_\_\_\_\_\_ проектирования и поискового конструирования\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Кафедра\_\_\_\_\_\_\_\_Системы автоматизированного проектирования и поискового конструирования\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дисциплина\_\_\_Компьютерная лингвистика\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |
| --- | --- |
|  | Утверждаю Зав. кафедрой\_\_Щербаков М.В.\_\_\_\_\_ |
|  | «\_\_\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20 \_\_\_ г. |

##### **Задание**

##### **на курсовую работу (проект)**

Студент\_\_\_\_\_\_\_\_\_\_\_\_\_Манукян Арсен Ваганович\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

##### (фамилия, имя, отчество)

Группа\_\_\_\_\_\_\_\_ИВТ-365\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Тема: \_\_\_\_\_\_\_\_\_\_\_Создание модуля кроулер новостного сайта\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Утверждена приказом от «\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_\_ г. № \_\_\_\_\_\_\_\_\_

2. Срок представления работы (проекта) к защите «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_ г.

3. Содержание расчетно-пояснительной записки: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4. Перечень графического материала: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5. Дата выдачи задания «\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20 \_\_\_ г.

Руководитель работы (проекта)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись, дата инициалы и фамилия

Задание принял к исполнению\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись, дата инициалы и фамилия

Минобрнауки России

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Волгоградский государственный технический университет»

Факультет \_\_\_\_\_\_\_\_\_\_\_\_\_ Электронной вычислительной техники\_\_\_\_\_\_\_\_\_\_\_\_

Кафедра\_\_\_\_\_\_\_\_Системы автоматизированного проектирования и поискового конструирования\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

##### **ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

**к курсовой работе (проекту)**

по дисциплине \_\_\_\_\_\_\_ Компьютерная лингвистика\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

на тему\_\_\_\_\_\_\_\_ Создание модуля кроулер новостного сайта\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент\_\_\_\_\_\_\_\_\_\_\_\_\_\_Манукян Арсен Ваганович\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

##### (фамилия, имя, отчество)

Группа\_\_\_\_\_\_\_\_\_\_\_ ИВТ-365\_\_\_\_\_\_\_\_\_\_\_\_\_

Руководитель работы (проекта) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(подпись и дата подписания) (инициалы и фамилия)

Члены комиссии:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(подпись и дата подписания) (инициалы и фамилия)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(подпись и дата подписания) (инициалы и фамилия)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(подпись и дата подписания) (инициалы и фамилия)

Нормоконтролер \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(подпись, дата подписания) (инициалы и фамилия)

### Волгоград 20\_\_ г.

Содержание

[1 Описание таблиц 4](#__RefHeading___Toc41001462)

[2 Пример заполнения бд записями 5](#__RefHeading___Toc41001462)

[3 Код и его описание 6](#__RefHeading___Toc41001462)

1. Описание таблиц:

header – название новости

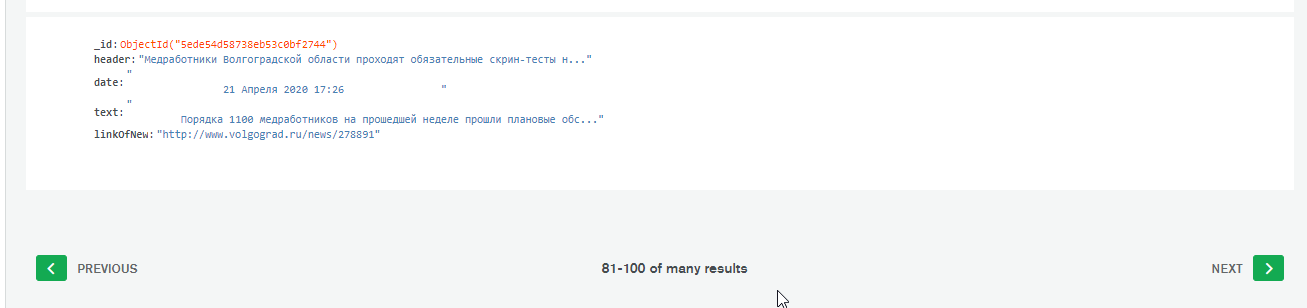
date – дата и время новости

text – основной текст новости

linkOfNew – ссылка на новость

2. Пример заполнения БД записями:





3. Код и его описание:

Подключение к библиотекам

import requests

from pymongo import MongoClient

from bs4 import BeautifulSoup

from multiprocessing import Pool

from datetime import datetime

from tqdm import trange

import sys

from pymongo import MongoClient

Подключение к клиенту mongodb

client = MongoClient("mongodb+srv://root:root@cluster0-xuev7.mongodb.net/sema?retryWrites=true&w=majority")

db=client.sema

news = db.sema

Ловим путь к новостям с помощью самой первой свежей ссылки

mainurl = 'http://www.volgograd.ru/news'

mainpage = requests.get(mainurl).text

mainsoup = BeautifulSoup(mainpage, 'html.parser')

newssoup = mainsoup.find("div", {"class":"col-md-12 news-item"}).find('a').get('href')

Находим число в ссылке, создаем массив, в который поместим все ссылки с 0 по текущую пойманную с помощью цикла

numarticle=""

for char in newssoup:

  if char.isdigit():

    numarticle+=char

all\_links = []

while int(numarticle)>0:

  all\_links.append(mainurl+"/"+str(numarticle))

  numarticle = int(numarticle)-1

Функция main- создаем многопоточность и реализуем ее с созданной функцией make\_all с помощью map

def main():

  start = datetime.now()

  n = len(all\_links)

  listOfList = []

  count\_of\_thread = 10

  for i in range(0, count\_of\_thread):

    listOfList.append(all\_links[int(i\*n/count\_of\_thread):int((i+1)\*n/count\_of\_thread)])

  for i in trange(count\_of\_thread, file=sys.stdout, desc='outer loop'):

    with Pool(count\_of\_thread) as p:

      p.map(make\_all, listOfList[i])

  end = datetime.now()

  total = end - start

  print(total)

if \_\_name\_\_ == '\_\_main\_\_':

  main()

Функция make\_all- получаем переменную soap типа BeautifulSoap и кидаем ее с ссылкой в созданную функцию findcontent

def make\_all(url):

  page = requests.get(url).text

  soup = BeautifulSoup(page, 'html.parser')

  findcontent(url, soup)

Функция findcontent- основная функция с парсингом данных. Заходим в основную ветку, где будем парсить (maincontent), после делаем проверку на наличие новостей (пустых новостей огромное количество и нужно их пропускать), создаем словарь и записываем в него полученные при парсинге данные, после с помощью функции insert\_one вставляем наши словари в mongodb

def findcontent(url, soup):

  maincontent = soup.find("div", {"class":"col-md-9 main-content"})

  if maincontent.find("div", {"class":"news-detail"}) !=None :

    header = maincontent.find('h1').text

    date = maincontent.find("div", {"class":"date"}).text

    texts = maincontent.find("div", {"class":"news-detail"}).find('p').text + maincontent.find("div", {"class":"news-detail"}).find("div", {"id":"full\_text"}).text

    new = {

      "header": header,

      "date": date,

      "text": texts,

      "linkOfNew": url

    }

    news.insert\_one(new)

  else:

    pass