

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Волгоградский государственный технический университет»

Дисциплина Компьютерная лингвистика

« » _____ 20 ____ г.

Задание принял к исполнению _____

Минобрнауки России
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Волгоградский государственный технический университет»

Факультет Электроники и вычислительной техники
Кафедра Системы автоматизированного проектирования и поискового
конструирования

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
к курсовой работе (проекту)

по дисциплине Компьютерная лингвистика
на тему Создание приложения по парсингу сайта novostivolgograda.ru и
выявлению тональности предложений с упоминанием персон и
достопримечательностей Волгоградской области

Студент Заборонок Сергей Валентинович
(фамилия, имя, отчество)

Группа ИВТ-363

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 2020г.

Содержание

Первая часть.....	3
База данных.....	3
Парсер.....	3
Web-интерфейс	6
Вторая часть.....	6
Томида-парсер	6
Word2Vec	7
Третья часть	9
Анализ тональности	9
Руководство системного администратора	10
Установка библиотек для Python.....	10
Установка и настройка MongoDB	11
Установка томида-парсера	12
Руководство пользователя.....	12
Работа с web-интерфейсом:.....	12
Использование парсера:.....	13
Использование модуля томида-парсера:.....	13
Использование word2vec:	13
Использование анализатора тональности:.....	13

Первая часть

База данных

Используется СУБД MongoDB.

Описание коллекций базы данных:

data – все найденные и скаченные новости

analysis - все найденные предложения с упоминанием персон и достопримечательностей.

tonality – тональность найденных предложений с упоминанием.

synonyms – найденные синонимы.

person – фамилии и имена персон.

attractions – название достопримечательностей.

Парсер

Парсер новостного сайта (parser/scrapper.py) — проходит по новостям сайта novostivolgograda.ru и собирает ссылки на статьи. Затем заходит по всем собранным ссылкам и получает дату новости, её текст и название. Все данные записываются в коллекцию data.

Скриншот заполненной коллекции data парсером:

```
_id: ObjectId("5ee164c31dd8a693e3f6fed2")
newsDate: 2020-06-10T00:00:00.000+00:00
newsLink: "https://novostivolgograda.ru/news/society/10-06-2020/v-volgogradskoy-o..."
newsName: " В Волгоградской области выявили 116 новых случаев коронавируса"
newsText: "Что доказывают обновленные данные федерального оперативного штаба. За ..."
forAnalysis: false
```

```
_id: ObjectId("5ee164c31dd8a693e3f6fed3")
newsDate: 2020-06-10T00:00:00.000+00:00
newsLink: "https://novostivolgograda.ru/news/society/10-06-2020/koronavirus-stal-..."
newsName: "Коронавирус стал главной темой обращений от волгоградцев в Роспотребна..."
newsText: "Как сообщает «НовостиВолгограда.ру» со ссылкой на пресс-службу областн..."
forAnalysis: false
```

```
_id: ObjectId("5ee164c31dd8a693e3f6fed4")
newsDate: 2020-06-10T00:00:00.000+00:00
newsLink: "https://novostivolgograda.ru/news/economy/10-06-2020/ksp-volgogradskiy-..."
newsName: "КСП: волгоградский Облтуризм неэффективно расходует средства и погряз ..."
newsText: "Как сообщает «НовостиВолгограда.ру» со ссылкой на документ, КСП озвучи..."
forAnalysis: false
```

```
_id: ObjectId("5ee164c31dd8a693e3f6fed5")
newsDate: 2020-06-10T00:00:00.000+00:00
newsLink: "https://novostivolgograda.ru/news/economy/10-06-2020/soderzhanie-obldu-..."
newsName: "Содержание облдумы обходится волгоградцам всё дороже"
newsText: "Как сообщает «НовостиВолгограда.ру» со ссылкой на отчёт КСП, на функци..."
forAnalysis: false
```

Рисунок 1. Коллекция data

```
_id: ObjectId("5ee0616835c5cb2ff5dc85d1")
personName: "Андрей БОЧАРОВ"
```

```
_id: ObjectId("5ee0616835c5cb2ff5dc85d2")
personName: "Зина МЕРЖОЕВА"
```

```
_id: ObjectId("5ee0616835c5cb2ff5dc85d3")
personName: "Роман БЕКОВ"
```

```
_id: ObjectId("5ee0616835c5cb2ff5dc85d4")
personName: "Василий ИВАНОВ"
```

```
_id: ObjectId("5ee0616835c5cb2ff5dc85d5")
personName: "Руслан ШАРИФОВ"
```

```
_id: ObjectId("5ee0616835c5cb2ff5dc85d6")
personName: "Александр ДОРЖДЕЕВ"
```

```
_id: ObjectId("5ee0616835c5cb2ff5dc85d7")
personName: "Анатолий ВАСИЛЬЕВ"
```

Рисунок 2. Коллекция person

`_id: ObjectId("5ee0620b6b3b940d1e25d796")`
`attractionsNames: "Авангард"`

`_id: ObjectId("5ee0620b6b3b940d1e25d797")`
`attractionsNames: "Областная универсальная научная библиотека им. М. Горького"`

`_id: ObjectId("5ee0620b6b3b940d1e25d798")`
`attractionsNames: "ОУНБ им. М. Горького"`

`_id: ObjectId("5ee0620b6b3b940d1e25d799")`
`attractionsNames: "Площадь Павших Борцов"`

`_id: ObjectId("5ee0620b6b3b940d1e25d79a")`
`attractionsNames: "Памятник Саше Филиппову"`


`_id: ObjectId("5ee0620b6b3b940d1e25d79b")`
`attractionsNames: "Музей истории Кировского района"`

`_id: ObjectId("5ee0620b6b3b940d1e25d79c")`
`attractionsNames: "Памятник чекистам"`

Рисунок 3. Таблица attractions

Web-интерфейс

Web-интерфейс главной страницы выглядит следующим образом, на нем представлена ссылки на статьи с сайта novostivolgograda.ru, их названия и дата:



[Новости](#) [Факты](#) [Синонимы](#)

Новости с сайта
novostivolgograda.ru

#	Title	Date
1	Бесплатного сыра не бывает: как российские сыровары чувствуют себя в кризис	12-06-2020
2	«Варварское» озеленение добавило жару Волгограду	14-06-2020
3	Качу куда хочу: электросамокаты никак не впишутся в ПДД	14-06-2020
4	Жара +40 °C выгнала волгоградцев на пляжи	12-06-2020
5	В истории с севшим на бутылку контрактником случился неожиданный «проворот»	12-06-2020
6	В Волгограде подорожал проезд в маршрутках	13-06-2020
7	Умерших от коронавируса волгоградцев может быть в два раза больше	13-06-2020
8	У волгоградского депутата Крылова подтвердился COVID-19	12-06-2020
9	В Волгограде сохраняется опасность жары в 40 градусов	14-06-2020
10	В Волгограде восстановили убивший ребёнка фонтан	14-06-2020

Еще....

Рисунок 4. Интерфейс главной страницы

Вторая часть

Томи́та-парсер

Выделение персон и достопримечательностей и помещение их в таблицу analysis выполняет скрипт tomita.py.

Перед запуском необходимо его поместить в папку с файлами persons.cxx, attractions.cxx, mydic.gz и facttypes.proto.

Пример выделяемых скриптом предложений:

```
[15:06:20 20:03:43] - Start. (Processing files.)
Пока новая методика антивирусной терапии носит единичные случаи .
Только трем инфицированным ввели плазму с антителами к COVID-19 .
У одного из них наступило резкое улучшение , отметил представитель облздрави Николай Алимов .
  Person
  {
    FIO = Николай Алимов
    Text = У одного из них наступило резкое улучшение , отметил представитель облздрави Николай Алимов .
  }
- По наблюдениям клиницистов у одного человека явно положительный эффект .
Это дало в комплексе или конкретно от переливания плазмы , сейчас пока трудно сказать .
Но пациент что называется на глазах вышел из тяжелого состояния .
Другие два человека стабильны , будем дальше вести работу и наблюдения .
Николай Алимов , заместитель председателя комитета здравоохранения Волгоградской области Плазма от выздоровевших пациентов больным вводится однократно , пояснил Николай Алимов .
  Person
  {
    FIO = Николай Алимов
    Text = Николай Алимов, заместитель председателя комитета здравоохранения Волгоградской области Плазма от выздоровевших пациентов больным вводится однократно, пояснил Николай
    Алимов.
  }
Биоматериал записает Волгоградский областной центр крови , рассказывало наше издание .
Пока кровь сдали 15 человек , так как не все перенесшие инфицирование подходят для донорства .
[15:06:20 20:03:44] - End. (Processing files.)
admin@ruvds-tjbt8:~/News-analysis-vlg/tomita$
```

Рисунок 5. Пример выделяемых предложений

Рисунок 6. Интерфейс выделяемых предложений

Word2Vec

Модель word2vec была обучена на новостных статьях из базы данных, объем которой составлял примерно 45000 статей. Модель находится в папке ./word2vec/model/kurs_model/. Запуск модуля осуществляется запуском программы ./word2vec/main.py. Модуль записывает контекстные синонимы в БД и осуществляет следующий вывод:

'Контекстные синонимы слов, полученные из модели, обученной на статьях:'

'-----'

'алимов'

Row(word='зампредседателя', similarity=0.816402792930603)

Row(word='облздрави', similarity=0.781923234462738)

Row(word='себелев', similarity=0.7760810852050781)

Row(word='лукияненко', similarity=0.7573933601379395)

Row(word='семисотов', similarity=0.7363854050636292)

'-----'

'бочаров'

Row(word='губернатор', similarity=0.9528191685676575)


Row(word='андрей', similarity=0.9124425649642944)

Row(word='бочаровглава', similarity=0.910007655620575)

Row(word='поручил', similarity=0.8725244402885437)

Row(word='проинспектировал', similarity=0.8666603565216064)

Модуль имеет следующий web-интерфейс, на нем представлено поле для ввода слово, к которому нужно найти контекстный синоним, после нажатия кнопки «Поиск» происходит вывод контекстных синонимов:

 [Новости](#) [Факты](#) [Синонимы](#)

Поиск контекстных синонимов

Поиск

Синоним

зампредседателя

облздрава

себелев

лукьяненко

семисотов

Рисунок 7. Интерфейс поиска контекстных синонимов

Скриншоты данных в таблицах:

```
> _id: ObjectId("5ee6db443b0b13c080f89247")
  word: "николай"
  > synom: Array

_id: ObjectId("5ee6db443b0b13c080f89248")
  word: "алимов"
  ▼ synom: Array
    0: "зампредседателя"
    1: "облздрава"
    2: "себелев"
    3: "лукьяненко"
    4: "семисотов"

_id: ObjectId("5ee6db443b0b13c080f89249")
  word: "ирина"
  ▼ synom: Array
    0: "татьяна"
    1: "анна"
    2: "наталья"
    3: "попова"
    4: "светлана"
```

Рисунок 8. Таблица synonyms

Третья часть

Анализ тональности

Тональность определяется у предложений, выявленных томита-парсером на втором этапе работы. Для анализа была использована библиотека Dostoevsky. Запуск модуля осуществляется запуском программы tonality/dos.py. В результате работы модуль записывает в БД id предложений и их тональность, также. В web-интерфейсе представлены предложения, выявленные на втором этапе и их тональность, которая указана в скобках:



Новости Факты Синонимы

Волгоградца в тяжелом состоянии спасли переливанием ковидной плазмы([ссылка](#))

Пока новая методика антивирусной терапии носит единичные случаи. Только трем инфицированным ввели плазму с антителами к COVID-19. У одного из них наступило резкое улучшение, отметил представитель облздрава Николай Алимов. – По наблюдениям клиницистов у одного человека явно положительный эффект. Это дало в комплексе или конкретно от переливания плазмы, сейчас пока трудно сказать. Но пациент, что называется на глазах вышел из тяжелого состояния. Другие два человека стабильны, будем дальше вести работу и наблюдения. Николай Алимов, заместитель председателя комитета здравоохранения Волгоградской области Плазма от выздоровевших пациентов больным вводится однократно, пояснил Николай Алимов. Биоматериал запасает Волгоградский областной центр крови, рассказывало наше издание. Пока кровь сдали 15 человек, так как не все перенёвшие инфицирование подходят для донорства.

09-06-2020

1. У одного из них наступило резкое улучшение, отметил представитель облздора Николай Алимов. (neutral)
2. Николай Алимов, заместитель председателя комитета здравоохранения Волгоградской области Плазма от выздоровевших пациентов больным вводится однократно, пояснил Николай Алимов. (neutral)

Рисунок 9. Интерфейс предложений и их тональности

Таблица tonality:

	<pre>_id: ObjectId("5ee164c31dd8a693e3f6fed6") tonality: Array 0: "neutral"</pre>
	<pre>_id: ObjectId("5ee164c31dd8a693e3f6fed7") tonality: Array 0: "neutral"</pre>
	<pre>_id: ObjectId("5ee164c31dd8a693e3f6fedd") tonality: Array 0: "neutral" 1: "neutral"</pre>
>	<pre>_id: ObjectId("5ee164c31dd8a693e3f6fede") tonality: Array 0: "neutral"</pre>
	<pre>_id: ObjectId("5ee164c31dd8a693e3f6fedf") tonality: Array 0: "neutral"</pre>
	<pre>_id: ObjectId("5ee164c31dd8a693e3f6fee0") tonality: Array</pre>

Рисунок 10. tonality

Руководство системного администратора

Установка библиотек для Python

Установить python 3.6 и pip (если их нет)

```
sudo apt install python3.6
```

Установить библиотеки

```
pip3 install pyspark flask dostoevsky pyppeteer pymongo schedule
```

Для корректной работы парсера установите зависимости

```
sudo apt install -y gconf-service libasound2 libatk1.0-0 libc6 libcairo2 libcups2
libdbus-1-3 libexpat1 libfontconfig1 libgcc1 libgconf-2-4 libgdk-pixbuf2.0-0
libglib2.0-0 libgtk-3-0 libnspr4 libpango-1.0-0 libpangocairo-1.0-0 libstdc++6
libx11-6 libx11-xcb1 libxcb1 libxcomposite1 libxcursor1 libxdamage1 libxext6
```

libxfixes3 libxi6 libxrandr2 libxrender1 libxss1 libxtst6 ca-certificates fonts-
liberation libappindicator1 libnss3 lsb-release xdg-utils wget

Установка и настройка MongoDB

```
sudo apt install -y mongodb
```

```
sudo systemctl status mongodb
```

Вы увидите следующий результат:

`mongodb.service` - An object/document-oriented database

Loaded: loaded (/lib/systemd/system/mongodb.service; enabled; vendor preset: enabled)

Active: active (running) since Sat 2018-05-26 07:48:04 UTC; 2min 17s ago

Docs: man:mongod(1)

Main PID: 2312 (mongod)

Tasks: 23 (limit: 1153)

CGroup: /system.slice/mongodb.service

└─2312 /usr/bin/mongod --unixSocketPrefix=/run/mongodb --config
/etc/mongodb.conf

Создаём пользователя:

```
mongo
```

```
use admin
```

Создайте пользователя `mongo-admin`:

```
> db.createUser(  
  {  
    user: "mongo-admin",  
    pwd: "passw0rd",  
    roles: [ { role: "userAdminAnyDatabase", db: "admin" } ]  
  }  
)
```

Создайте пользователя `mongo-root`:

```
> db.createUser(  
  {  
    user: "mongo-root",  
    pwd: "passw0rd",
```

```
roles: [ { role: "root", db: "admin" } ]  
}  
)
```

Установка томиита-парсера

cd ~

git clone <https://github.com/yandex/tomita-parser>

sudo apt-get install build-essential cmake lua5.2

cd tomita-parser && mkdir build && cd build

cmake ../src/ -DCMAKE_BUILD_TYPE=Release

make

Переместить файл libmystem_c_binding.so в ~/tomita-parser/build/bin/ :

wget https://github.com/yandex/tomita-parser/releases/download/v1.0/libmystem_c_binding.so.linux_x64.zip

unzip libmystem_c_binding.so.linux_x64.zip

rm libmystem_c_binding.so.linux_x64.zip

Экспортируем томииту

export PATH="\$HOME/tomita-parser/build/bin:\$PATH"

source ~/.bashrc

Установка приложения

Загрузить с github:

git clone <https://github.com/Sergey1888888/News-analysis-vlg>

Руководство пользователя

Работа с web-интерфейсом:

Работа web-интерфейса осуществляется запуском программы web/app.py. После запуска откроется окно (Рисунок 4), на нем можно просмотреть список названий новостей и дату их публикации с сайта <https://novostivolgograda.ru/> , при нажатии на название новости осуществляется переход на страницу

новости. Пользователь может перейти к поиску синонимов нажав кнопку «Синонимы», откроется окно (Рисунок 9), в котором пользователь может ввести персону или достопримечательность города Волгограда и получить 5 синонимов. Если пользователь перешел на страницу новости, то перед ним откроется окно (Рисунок 9) с текстом новости, ссылкой на её источник, датой и предложениями выделенными томиита-парсером с их тональностью. Если пользователь с главной страницы нажал на кнопку «Факты», то откроется окно с предложениями, выделенными томиита-парсером и ссылками на новости, в которых эти предложения встречались.

Использование парсера:

Для сбора данных необходимо запустить файл `parser/scrapper.py`, который найдет новые новости, запишет их в БД.

Примеры команды запуска:

```
python3 parser/scrapper.py
```

Таблицы `person` и `attractions` берутся из репозитория. (файлы `person.json` и `attractions.json`)

Использование модуля томиита-парсера:

Запустить `tomita.py`, дождаться окончания. `tomita.py` следует запускать из папки, в которой он расположен. В папке со скриптом должны находиться файлы `persons.cxx`, `attraction.cxx`, `mydic.gzt` и `facttypes.proto`.

Пример использования:

```
cd tomita
```

```
python3 tomita.py
```

Использование word2vec:

Если проект скачан из репозитория, то в нем уже имеется обученная на 45000 статей модель (`word2vec/model/kurs_model/`), готовая к выявлению контекстных синонимов. Если же вы хотите получить свою модель на основе данных из БД, то удалите/переименуйте папку исходной модели и запустите программу `word2vec/run.py`, начнется процесс подготовки и обучения модели. Для теста и просмотра контекстных синонимов запустите программу `word2vec/test.py`, в консоле выводиться синонимы.

Использование анализатора тональности:

Чтобы воспользоваться анализатором тональности установить на свой ПК библиотеку `Dostoevsky` (<https://github.com/bureaucratic-labs/dostoevsky>), после чего необходимо запустить код `tonality/dos.py`. Результатом работы

программы будет являться тональность, выявленная у предложений, выделенных томика-парсером.