Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«ВОЛГОГРАДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

(ВолгГТУ)

Факультет Электроники и вычислительной техники\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

Дисциплина Компьютерная лингвистика\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_ г.

**ЗАДАНИЕ**

**на курсовую работу (проект)**

Студент\_\_Самороков Николай Николаевич Группа ИВТ-363

Студент\_\_Новрузов Сергей Романович Группа ИВТ-363

Студент\_\_Григорьев Иван Максимович Группа ИВТ-363

(фамилия, имя, отчество)

1. Тема: Создание приложения по парсингу сайта novostivolgograda.ru и выявлению тональности предложений с упоминанием персон и достопримечательностей\_\_\_\_\_\_ Волгоградской области\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«ВОЛГОГРАДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

(ВолгГТУ)

Факультет Электроники и вычислительной техники\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

Дисциплина Компьютерная лингвистика\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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\_\_Самороков Николай Николаевич Группа ИВТ-363

Студент\_\_Новрузов Сергей Романович Группа ИВТ-363

Студент\_\_Григорьев Иван Максимович Группа ИВТ-363

(фамилия, имя, отчество)

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

подпись, дата инициалы и фамилия

Члены комиссии:

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

подпись, дата инициалы и фамилия

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

подпись, дата инициалы и фамилия

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

подпись, дата инициалы и фамилия

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

подпись, дата инициалы и фамилия

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

подпись, дата инициалы и фамилия

Волгоград 20\_\_ г.

СОДЕРЖАНИЕ

1.Первая часть .................................................................................................................... 3

1.1 База данных……………………………………………………………………..3

1.2 Описание парсера………………………………………………………………5

1.3 Отображение спаршенных новостей на нашем Web-интерфейсе…………..5

2. Вторая часть ....................................................................................................................6

2.1 Томита-парсер .....................................................................................................6

3. Третья часть .....................................................................................................................7

3.1 Тональность ........................................................................................................7

3.2 Word2Vec .............................................................................................................8

4. Руководство пользователю…………………………………………………………….9

1. **Первая часть**

**1.1 База данных:**

В качестве СУБД используется SQLite. Имеется база данных, в ней хранится несколько таблиц под названиями:

«post» - хранит «спаршенные» новости,

Поля таблицы «post»:

1. id\_post – (integer) – индекс новостной статьи,
2. title – (text) – заголовок новостной статьи,
3. date – (text) – дата публикации статьи,
4. link – (text) - ссылка,
5. text – (text) – содержание новостной статьи

«props» - хранит данные для анализа по тональности, содержит ФИО известного человека или название достопримечательности Волгоградской области, предложение, в котором о нем говорится, оценку по смысловой нагрузке (позитивная или негативная)

Поля таблицы «props»:

1. id\_props – (integer) – индекс таблицы,
2. facts – (text) – содержит ФИО известного человека или название достопримечательности Волгоградской области,
3. text – (text) – предложение, в котором упоминалась известная личность или достопримечательность,
4. tonality – (text) – тональность высказывания,
5. dates – (text) – дата публикации новости, в котором упомянута известная личность или достопримечательность Волгоградской области,
6. id\_post – (integer) – индекс новости, в которой упомянута достопримечательность или персона.

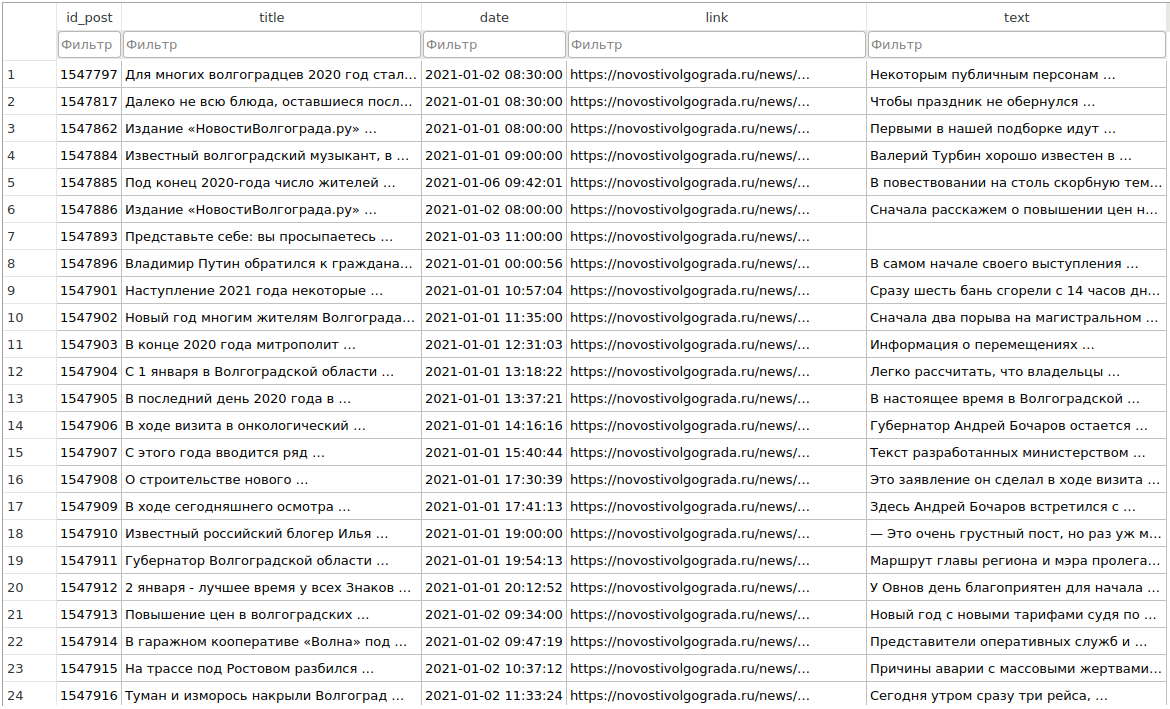


Рисунок 1 – Заполненная таблица post в БД

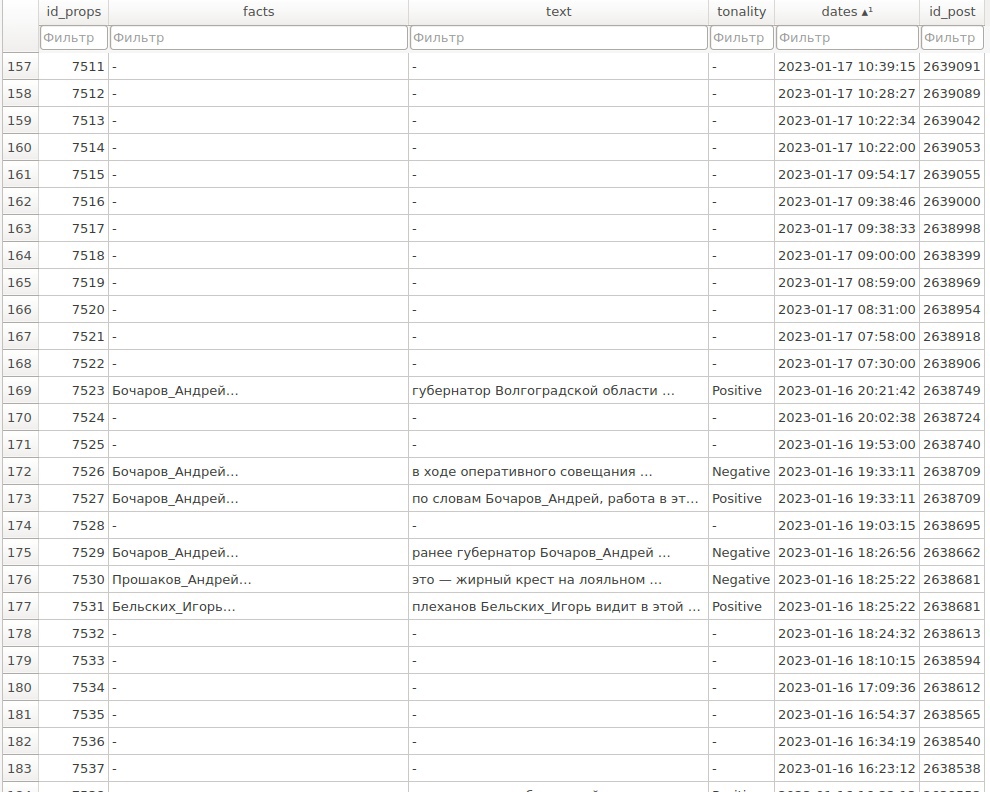


Рисунок 2 – Заполненная таблица props в БД

**1.2 Перейдем к описанию парсера:**

Парсер новостного сайта (располагается по следующему пути: sema/scrapper.py) «выполняет парсинг» новостного сайта novostivolgograda.ru (ссылка: <https://novostivolgograda.ru/>). Преимущественно парсер проходит по сайту и достает оттуда по одной новости и «собирает» с каждой всю необходимую информацию и заносит все поля в БД в таблицу «post».

**1.3 Отображение спаршенных новостей на нашем Web-интерфейсе**

Web-интерфейс страницы парсера выглядит следующим образом. Тут каждая новость отображается отдельно, с отображением всей обязательной информации, т.е. заголовок статьи, дату публикации, текст новости, ссылку на саму новость.

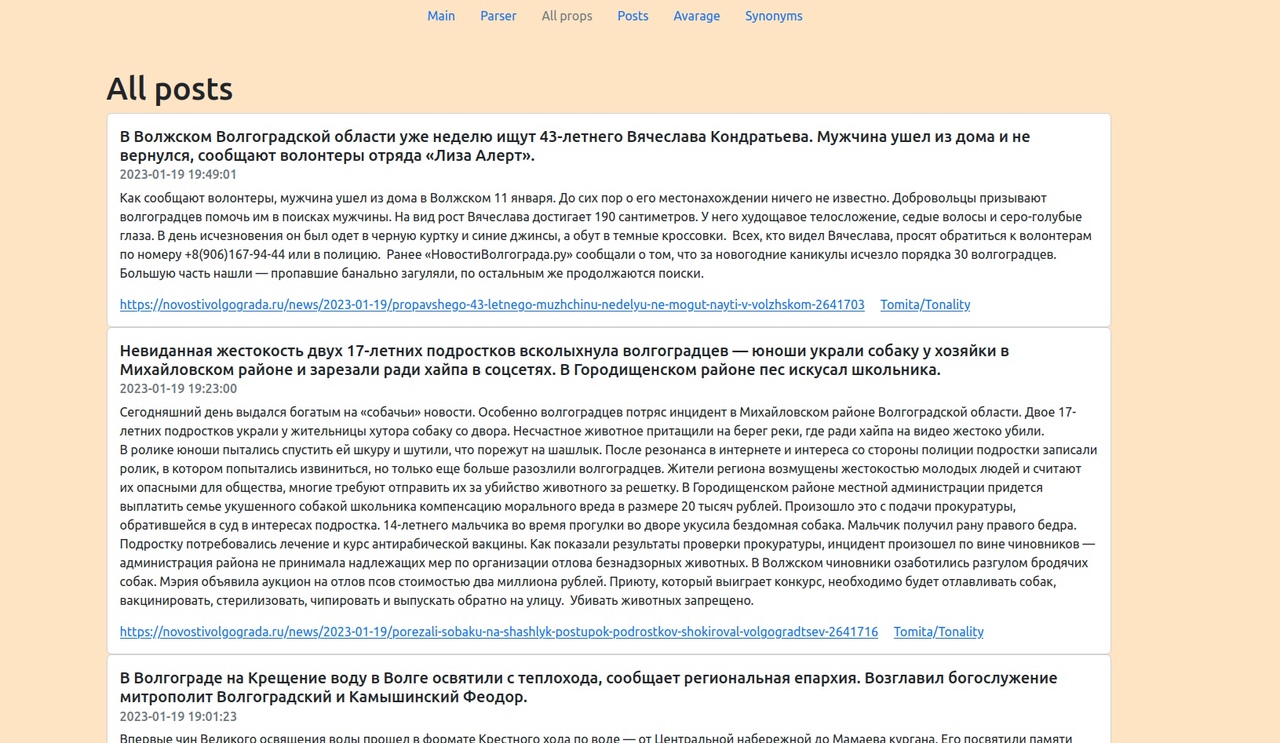


Рисунок 3 – Отображение спаршенных новостей на нашем сайте

1. **Вторая часть**

**2.1 Томита-парсер**

Выделение персон и достопримечательностей Волгоградской области и помещение их в таблицу «props» выполняет скрипт sema/tomita/tomita.py. Перед запуском необходимо его поместить в папку с файлами persons.cxx, attractions.cxx, dic.gzt и facttypes.proto. (По умолчанию эти файлы и сам файл, отвечающий за томиту располагается в папке «sema/tomita/»)

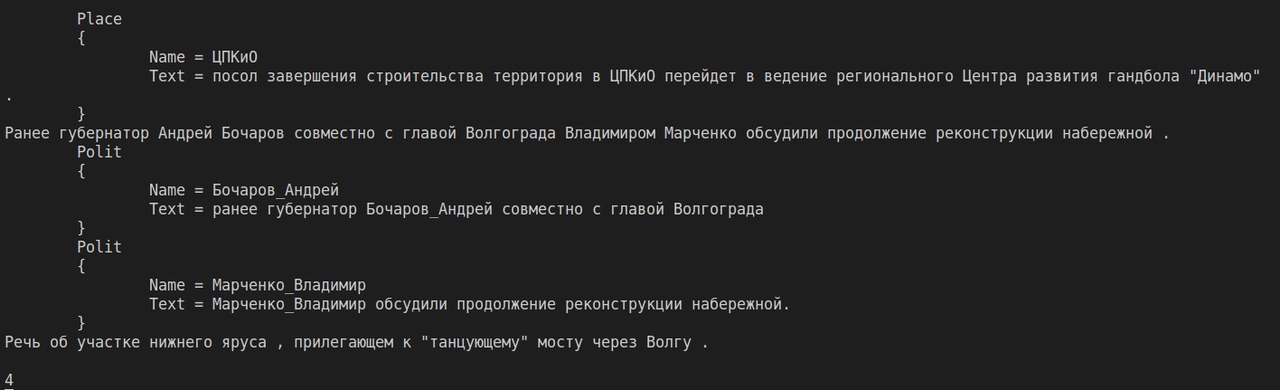


Рисунок 4 – Пример выделяемых предложений томитой-парсером

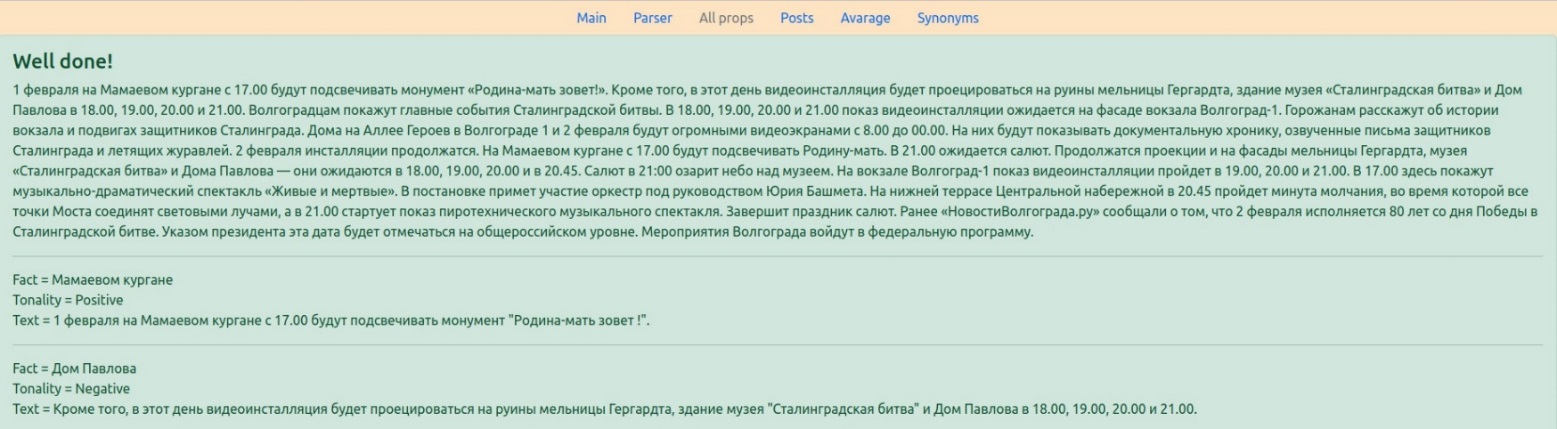
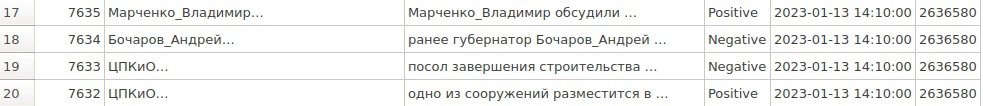


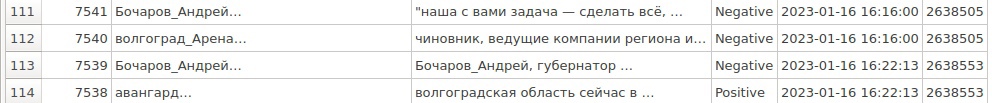
Рисунок 5 – Пример выделяемых предложений томитой-парсером (+ тональность) на сайте

1. **Третья часть**

**3.1 Тональность**

Тональность определяется у предложений, выявленных томита-парсером на втором этапе работы. Запуск модуля, отвечающего за тональность, осуществляется через файл tonality.py, расположенный по пути – sema/tonality/tonality.py. В результате работы модуль записывает в БД в таблицу «props» тональность, по каждому высказыванию (характеризует его смысловой окрас, иными словами – показывает является ли предложение, высказанное в адрес персоны или достопримечательности Волгоградской области положительным или же наоборот – отрицательным.)





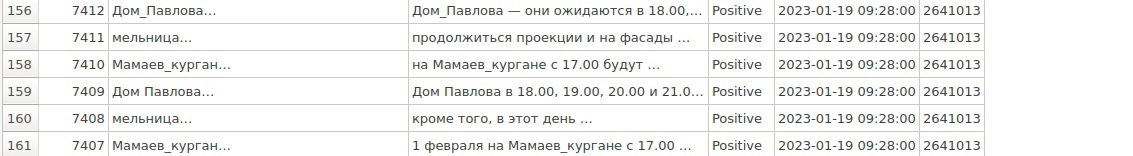


Рисунок 6 – Хранение тональности высказывания в таблице «props»

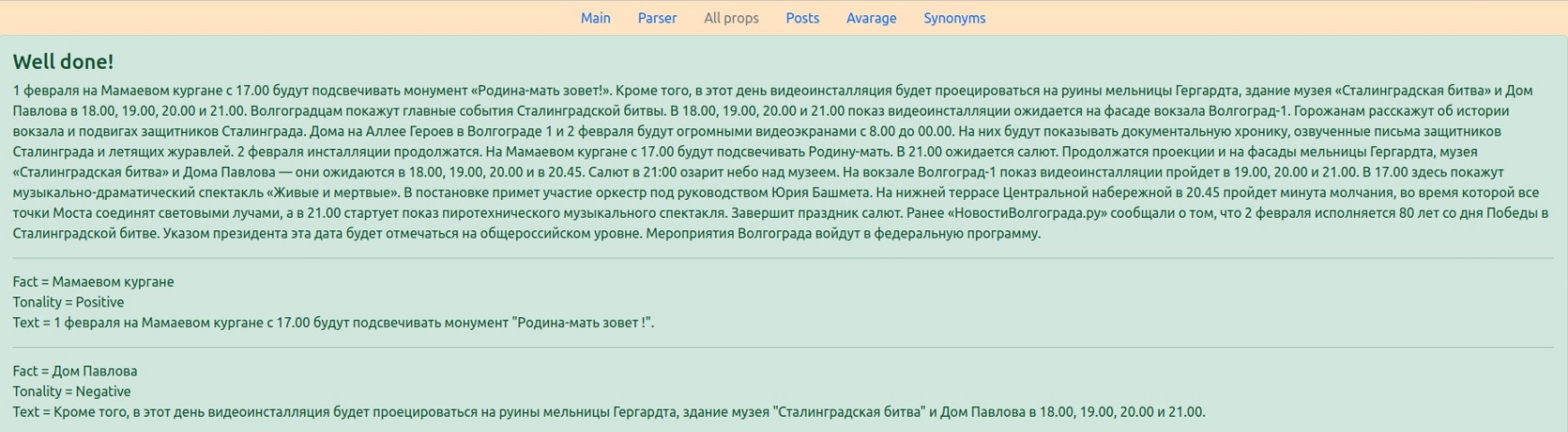


Рисунок 7 – Пример работы тональности на нашем сайте (по конкретной статье)

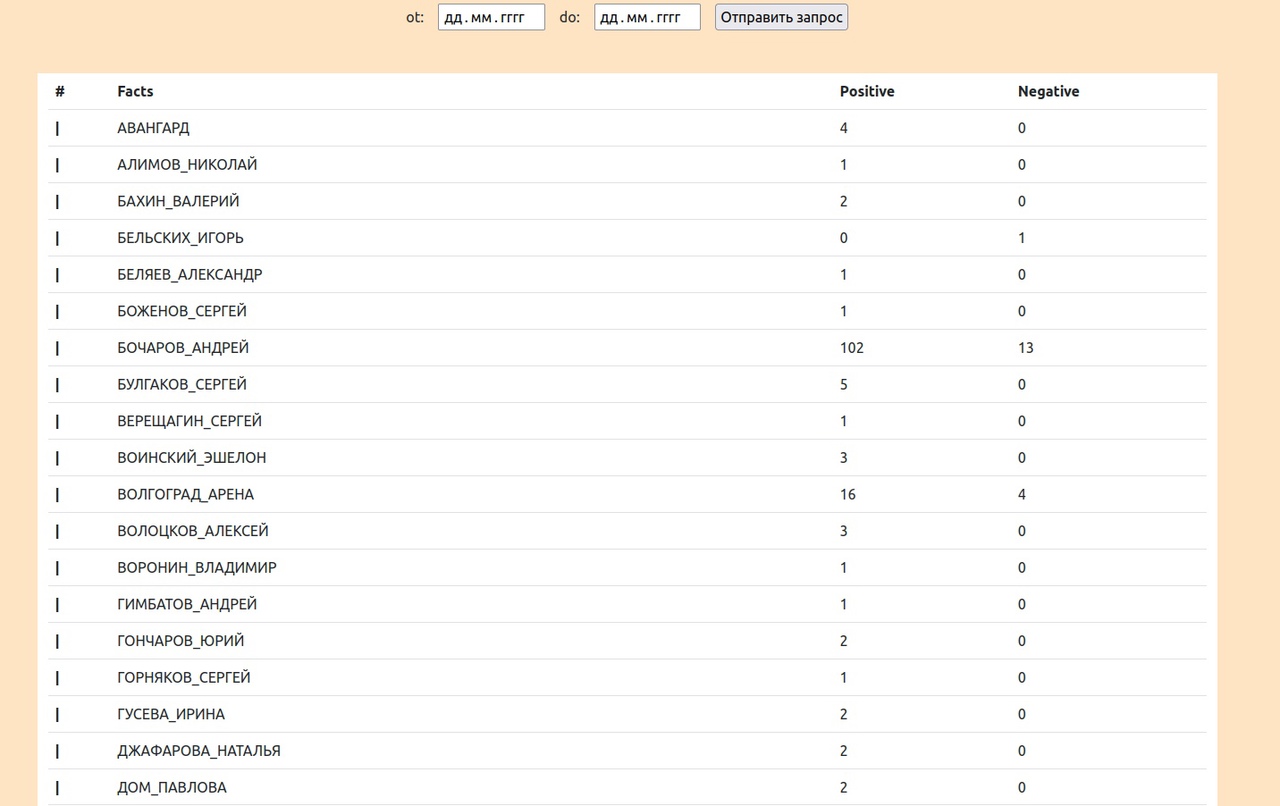


Рисунок 8 – Пример работы тональности на нашем сайте (по всем статьям)

**3.2 Word2Vec**

Модель Word2Vec была обучена на новостных статьях из нашей базы данных из таблицы «post», объем которой составлял примерно 10 000 статей. Модель находится в папке sema/model. Например, если поискать синонимы к слову «Бочаров», которые наиболее часто употреблялись в 10 000 новостях с искомым словом, то выдаст следующий результат:

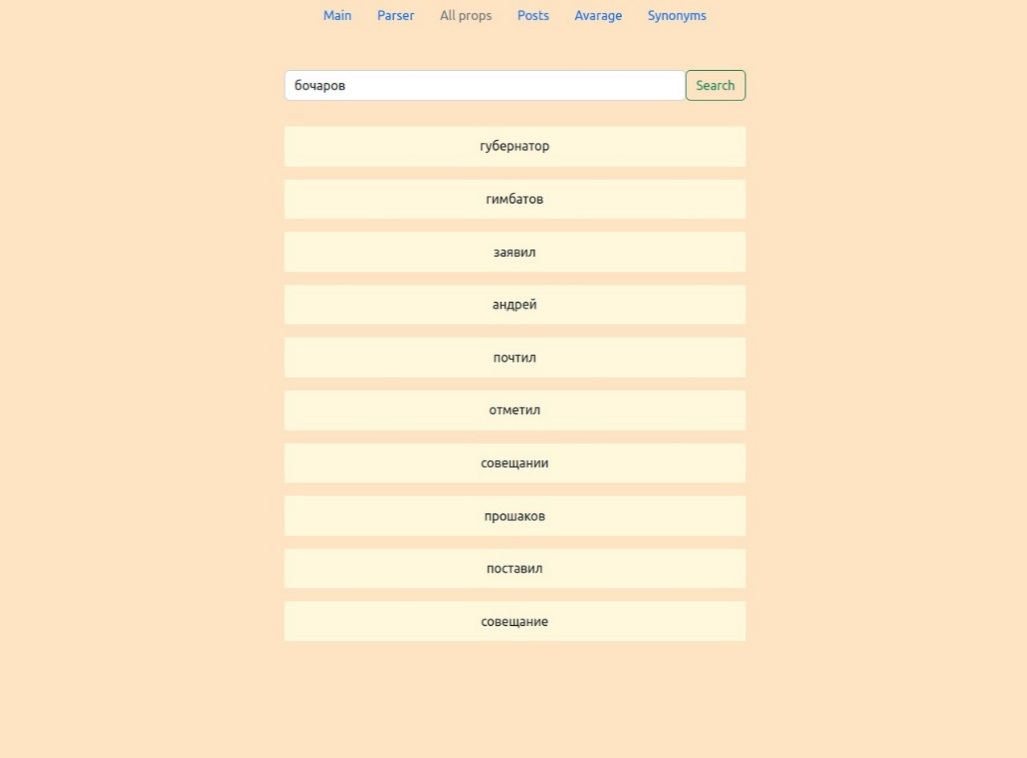


Рисунок 9 – Пример Word2Vec на нашем сайте

1. **Руководство пользователя**

**Работа с web-интерфейсом:**

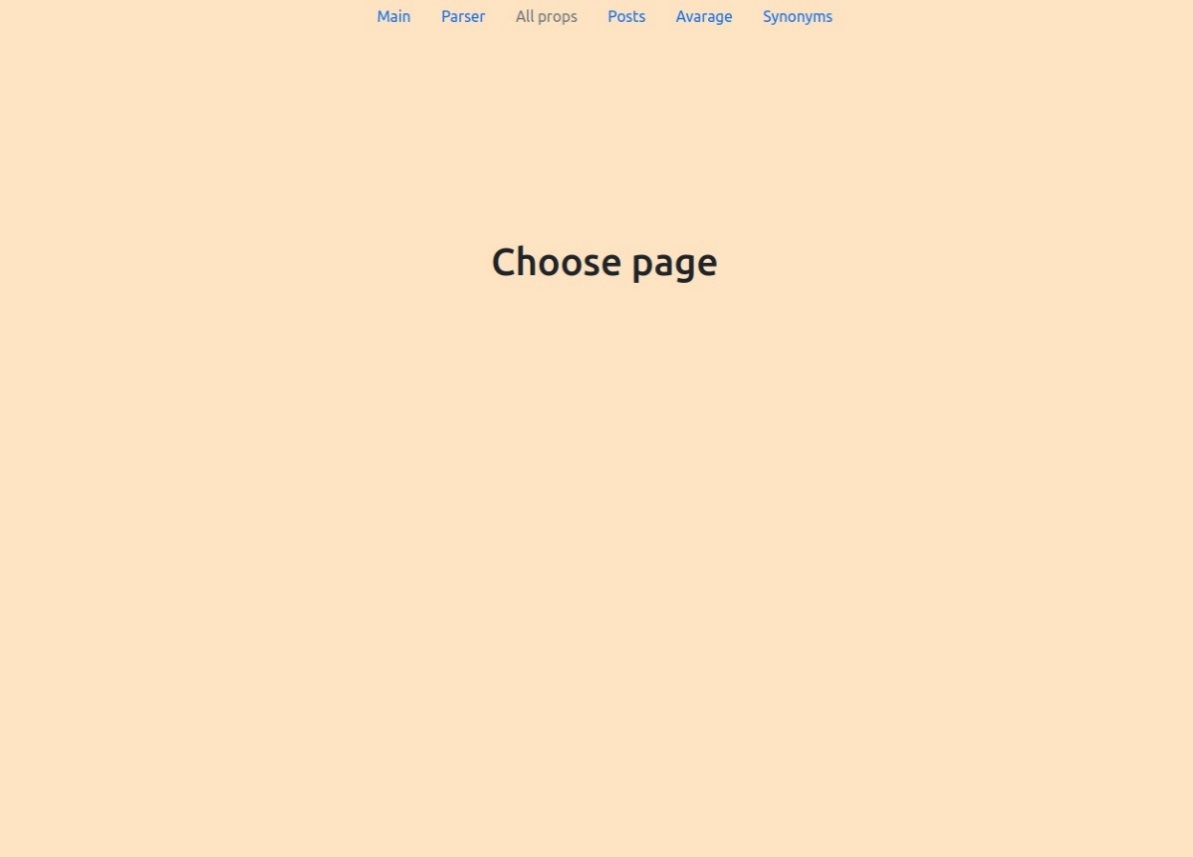


Рисунок 10 – Главная страница сайта

Работа web-интерфейса осуществляется запуском программы sema/app.py. После запуска откроется окно (Рисунок 7), на нем можно выбрать, что вы хотите делать с разработанным нами программным модулем. Например, сверху имеются кнопки, которые отвечают за следующие функции:

* Main – главная / стартовая страница, с нее и начинается приложение,
* Parser – запускает парсинг данных в БД с сайта <https://novostivolgograda.ru/> (при наличии всех записей в БД, парсер завершится достаточно быстро, если же БД будет полностью пустой или почти пустой, то ожидание завершение парсинга может занять продолжительное время, вплоть до нескольких часов),
* All props – запускает томита-парсер,
* Posts - страница отображающая все спаршенные статьи, по ссылке, которая прикреплена к каждой новости, можно перейти на сайт источник новости - novostivolgograda.ru и ознакомиться со всеми деталями подробнее,
* Average – страница, отображающая посты, с эмоциональной оценкой (положительная / отрицательная оценка поста) по количеству упоминаний в спаршенных новостях
* Synonyms – страница, для поиска контекстных синонимов для конкретно заданного пользователем слова.

**Сслыка на GitHub:**

https://github.com/QuanRy/Comp\_Ling