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\_\_Джозеф Прешес Охиани 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\_\_Джозеф Прешес Охиани 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.

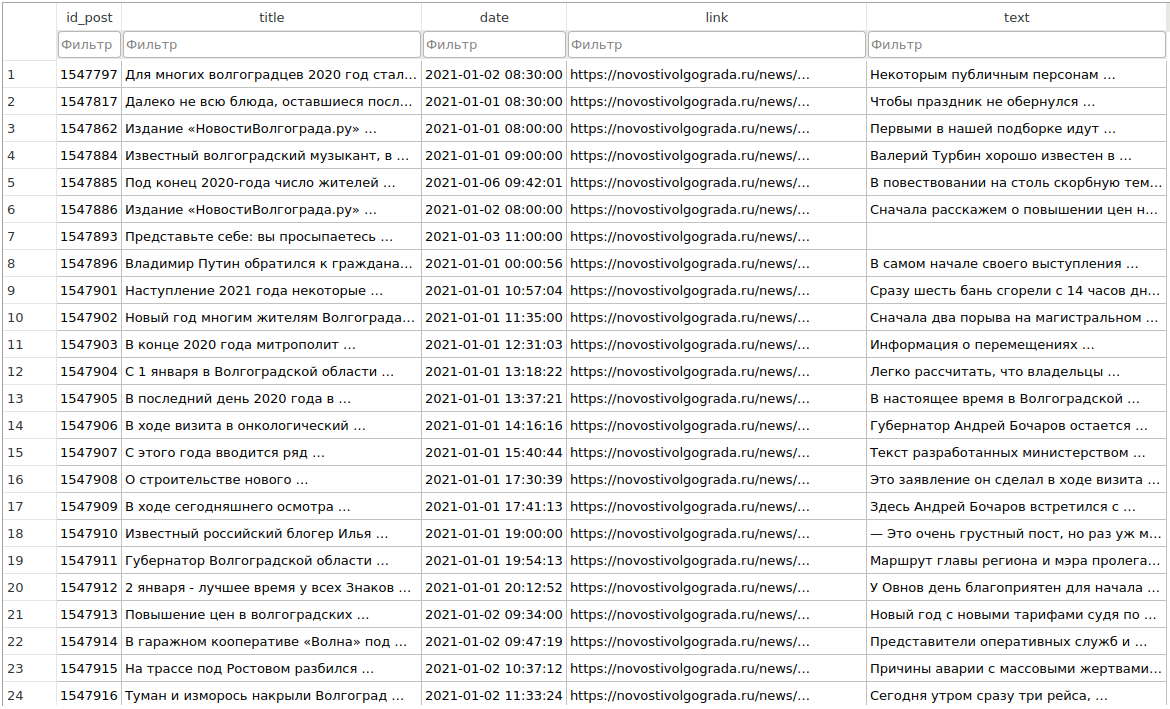


Рисунок 1 – Заполненная таблица post в БД

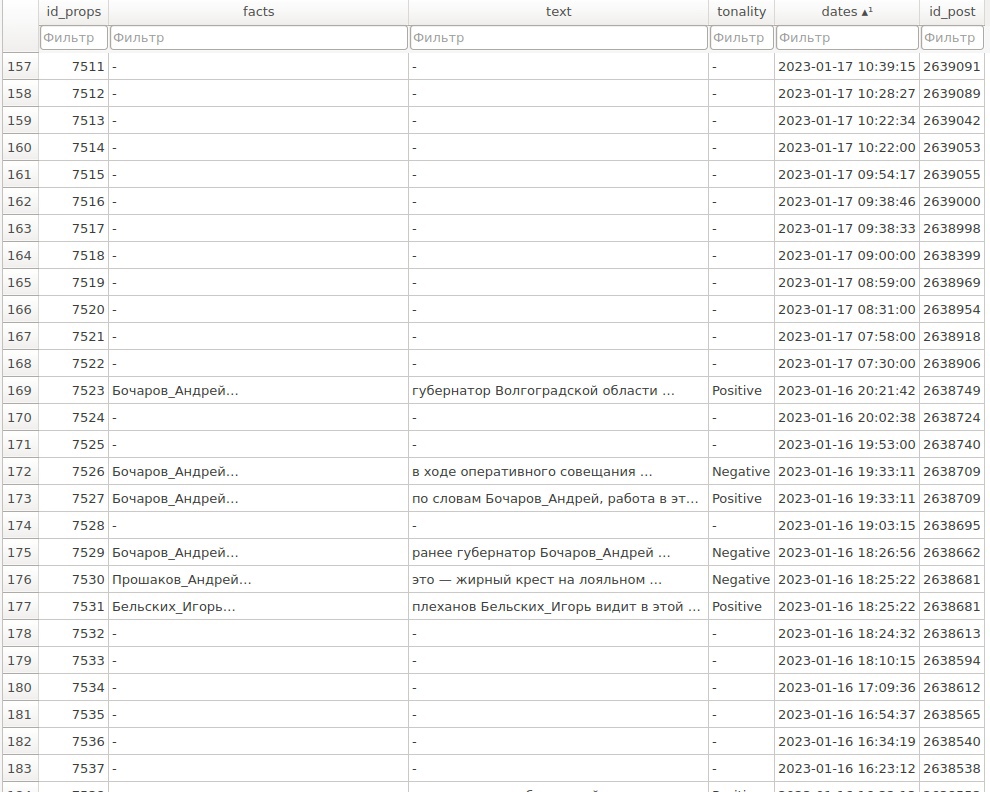


Рисунок 2 – Заполненная таблица props в БД

**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.

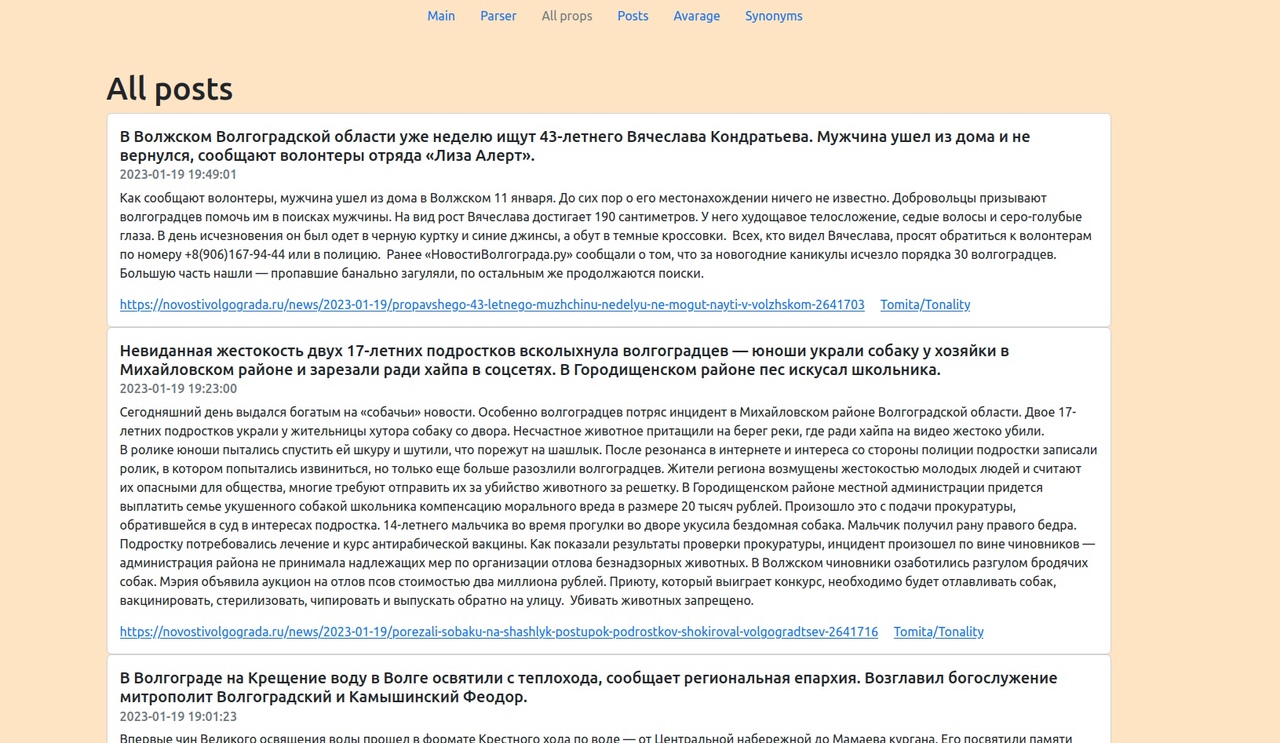


Рисунок 3 – Отображение спаршенных новостей на нашем сайте