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\_\_Бокова Ольга Дмитриевна 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\_\_Бокова Ольга Дмитриевна 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 2023 г.

**Оглавление**

[Первая часть 2](#_Toc125123317)

[Вторая часть 4](#_Toc125123318)

[Третья часть 6](#_Toc125123319)

[Руководство пользователя 7](#_Toc125123320)

[Руководство системного администратора 9](#_Toc125123321)

# Первая часть

**База данных:**

Используется СУБД Mongodb.

Описание базы данных:

1. \_id – идентификатор
2. date\_news – дата новости
3. link\_news – ссылка на новость
4. name\_news – название новости
5. num\_comments – количество комментариев новости
6. sentens - предложения, в которых томита-парсер нашел персон и достопримечательности
7. text\_news - текст новости
8. tonality - предложения, для которых проанализирована тональность

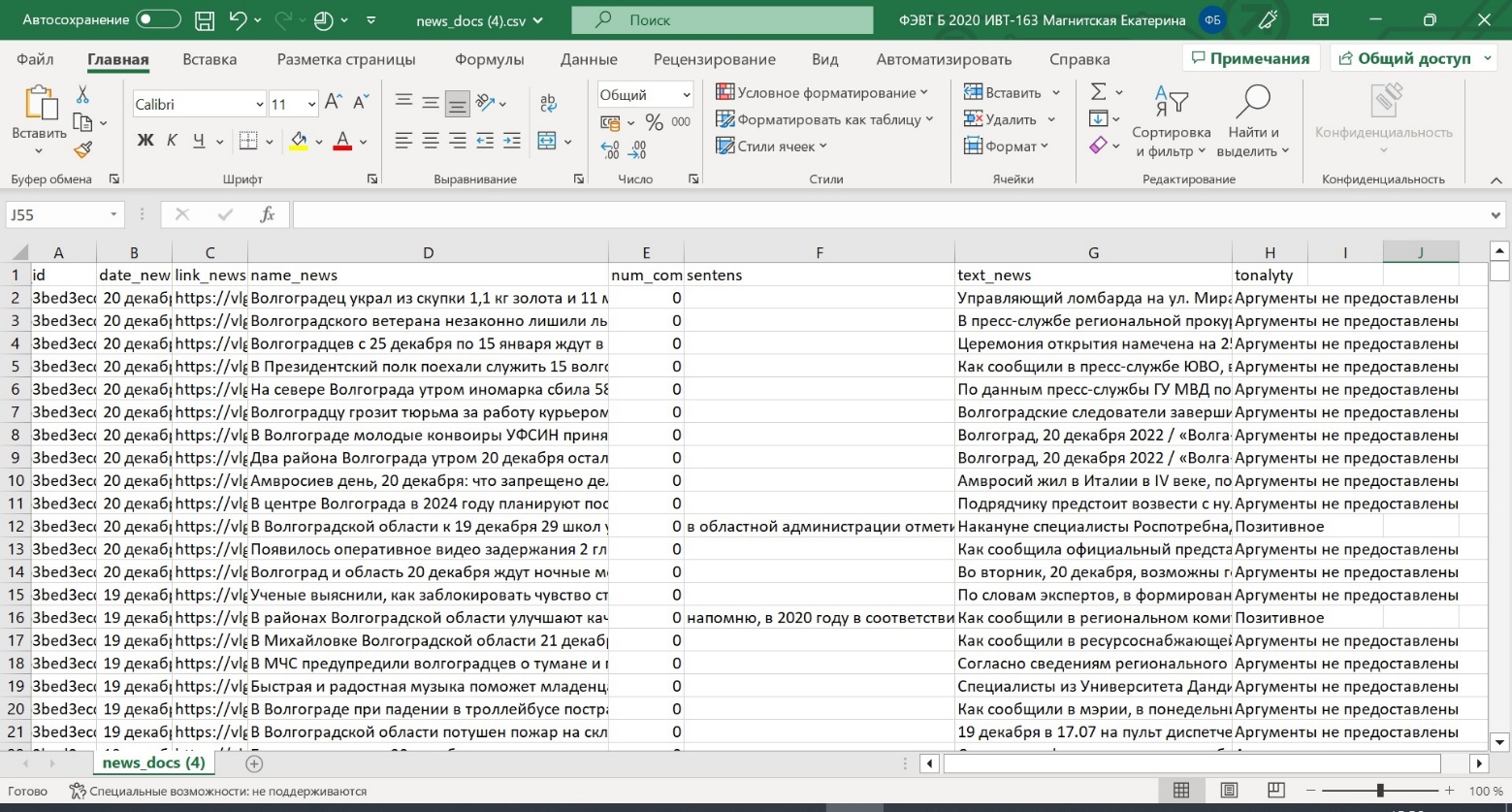


Рисунок 1 – Заполненная таблица с 10 000 новостей

**Перейдем к описанию парсера:**

Парсер новостного сайта (располагается по следующему пути: sema/scrapper.py) «выполняет парсинг» новостного сайта https://vlg-media.ru/. Парсер проходит по сайту и достает оттуда по одной новости и заполняет в БД необходимую информацию и заносит все поля.

**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.

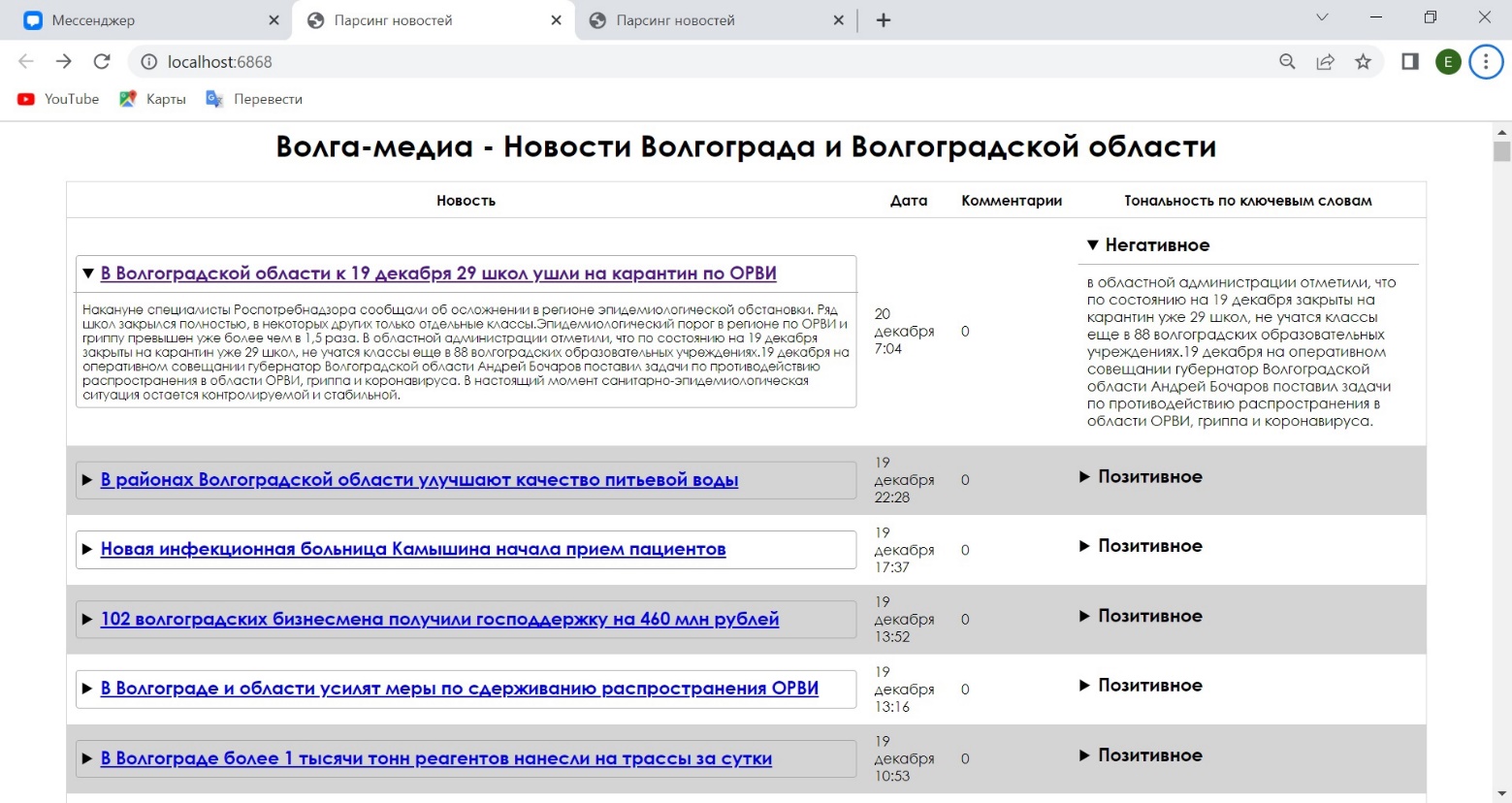


Рисунок 2 – Отображение спаршенных новостей на нашем сайте

# Вторая часть

**Томита-парсер**

Выделение персон и достопримечательностей Волгоградской области и помещение их в таблицу для дальнейшего анализа тональности:

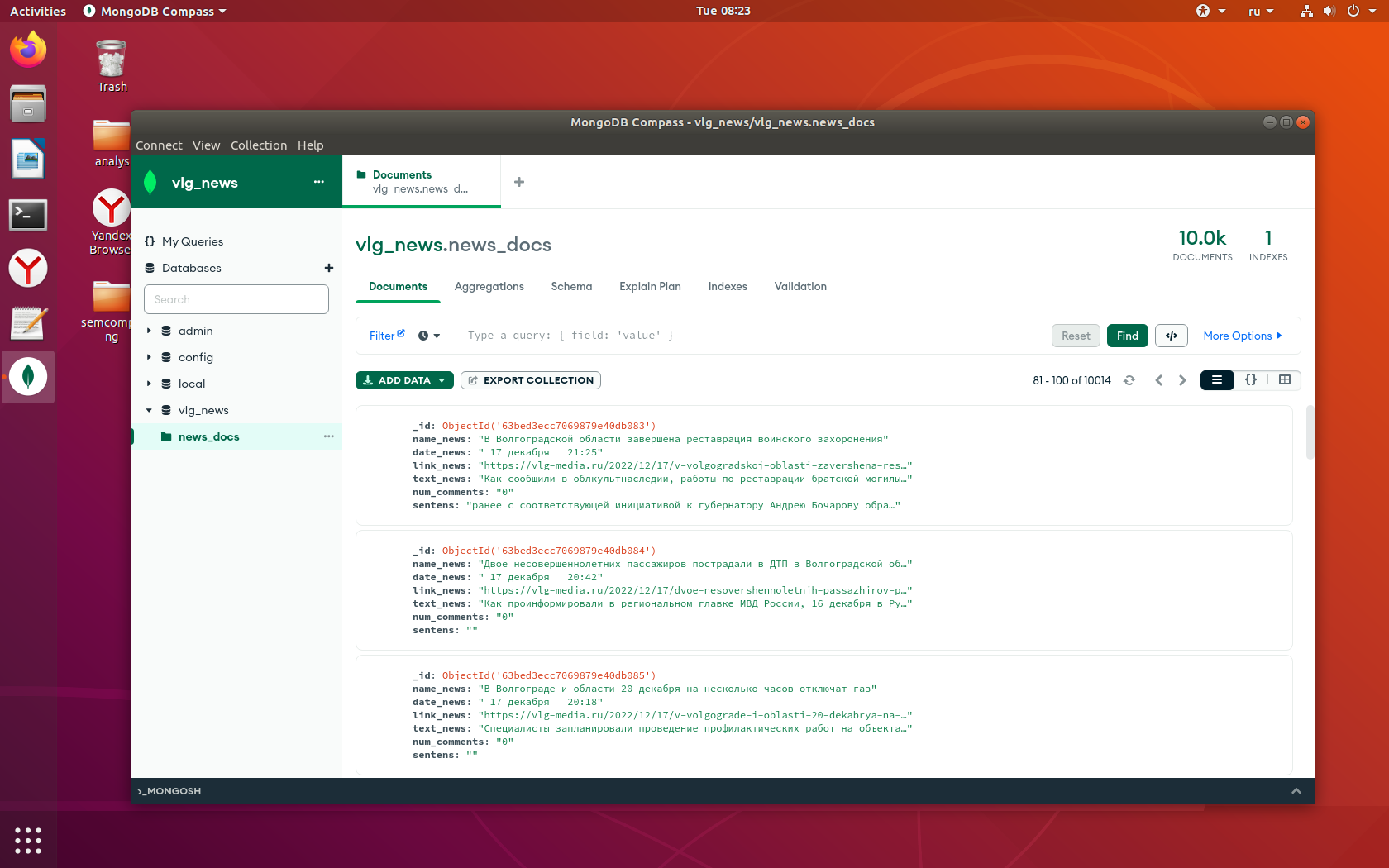


Рисунок 3 – Пример предложений, выделяемых томитой-парсером

**Word2Vec**

Модель Word2Vec была обучена на новостных статьях из нашей базы данных из таблицы, объем которой составлял примерно 10 000 статей. Если поискать синонимы к лисностям и достопремечательностям, которые наиболее часто употреблялись в 10 000 новостях с искомым словом, то выдаст следующий результат:

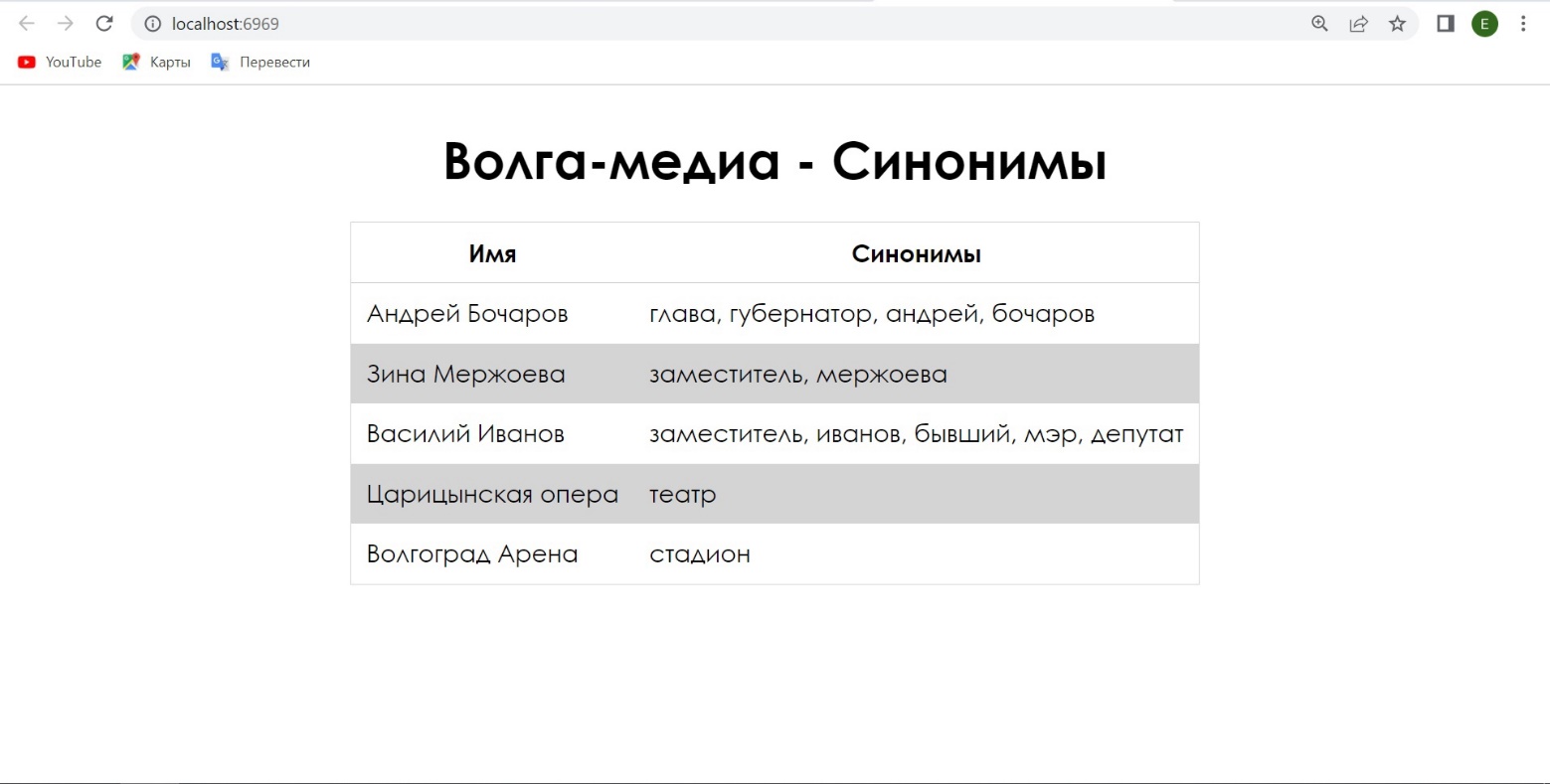


Рисунок 4 – Пример Word2Vec на нашем сайте

# Третья часть

**Тональность**

Тональность определяется у предложений, выявленных томита-парсером на втором этапе работы. В результате работы модуль записывает в БД в таблицу 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.)

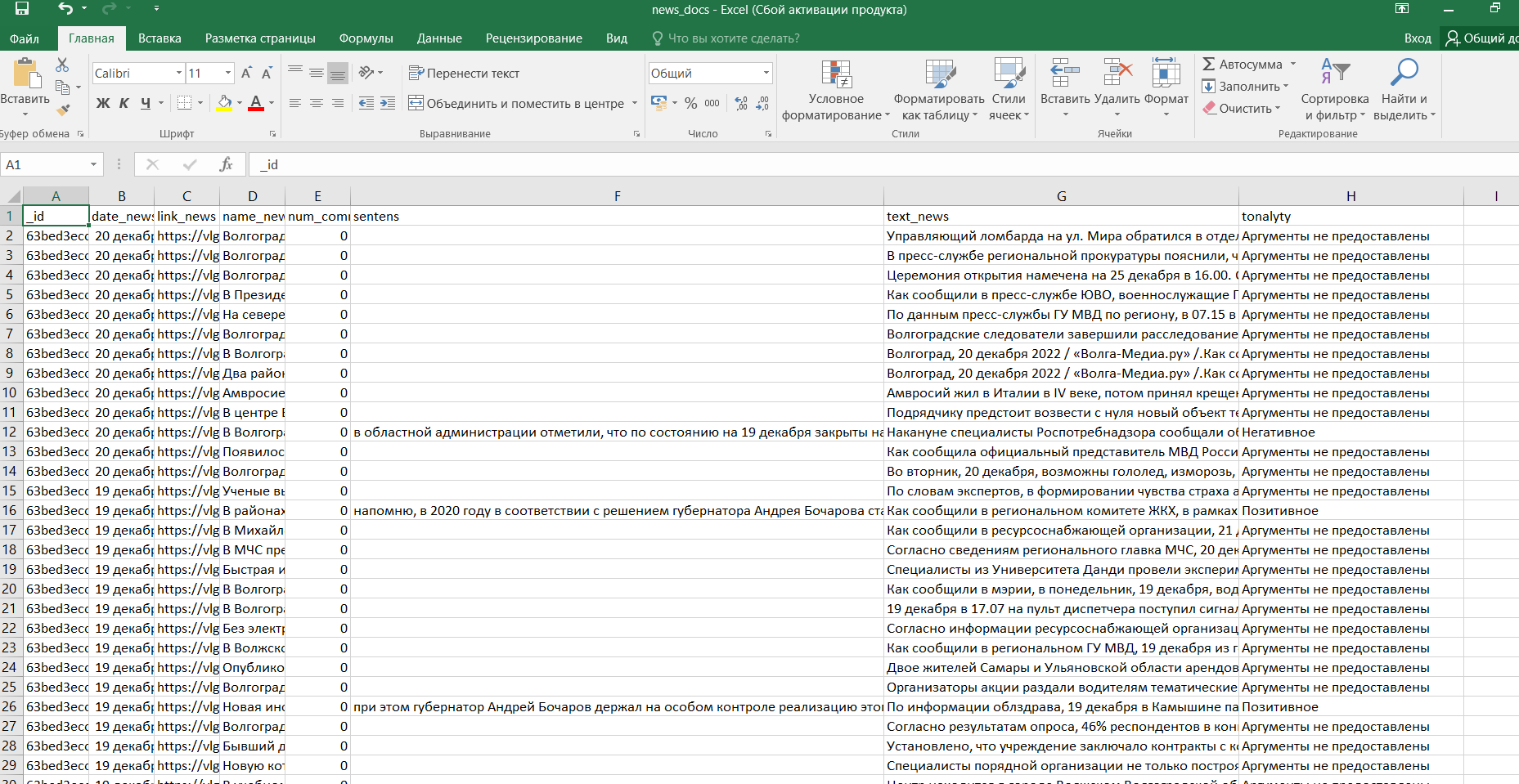


Рисунок 5 – Хранение тональности высказывания в csv-файле

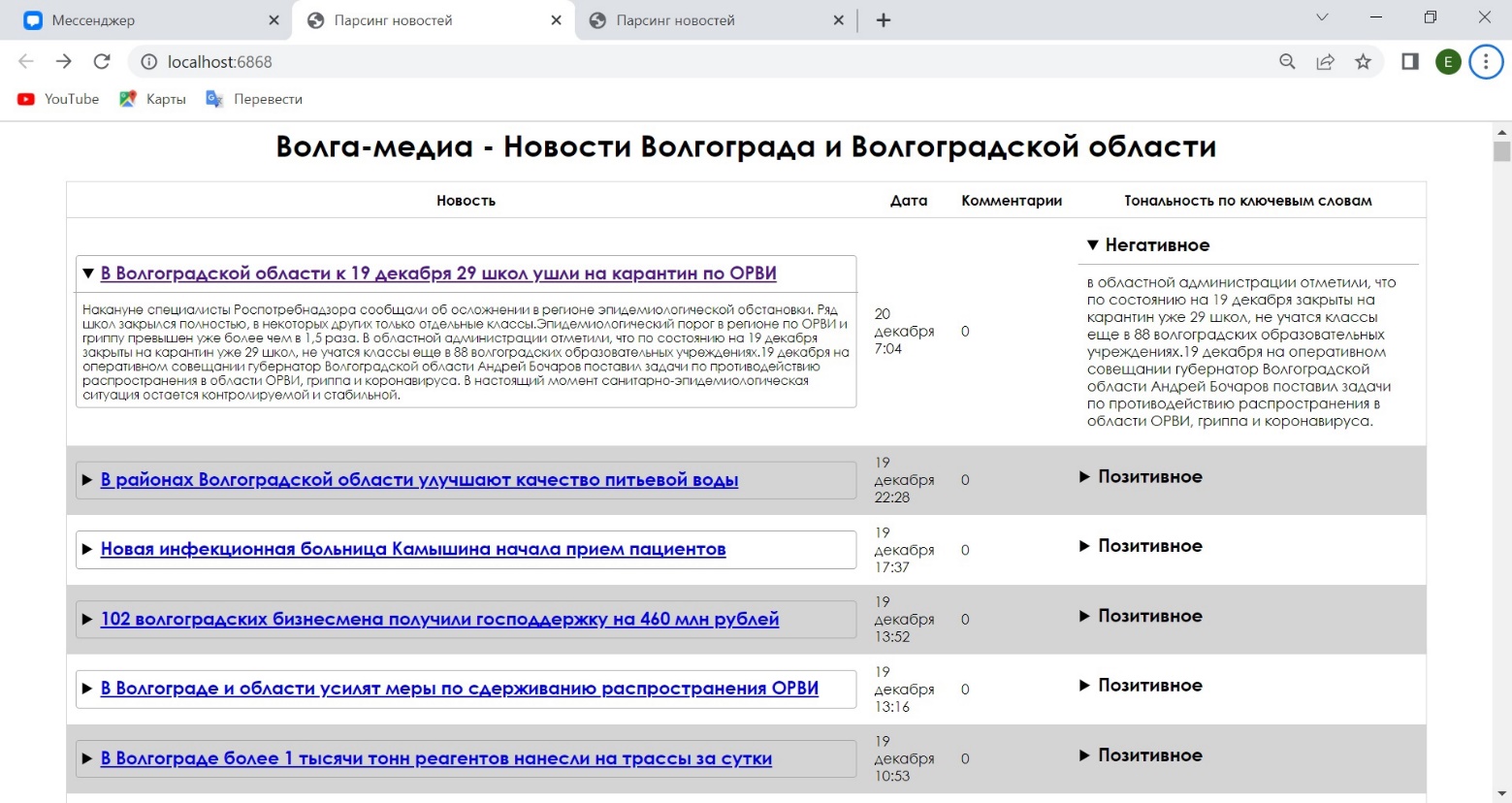


Рисунок 6 – Пример работы тональности на нашем сайте

# Руководство пользователя

**Работа с web-интерфейсом:**

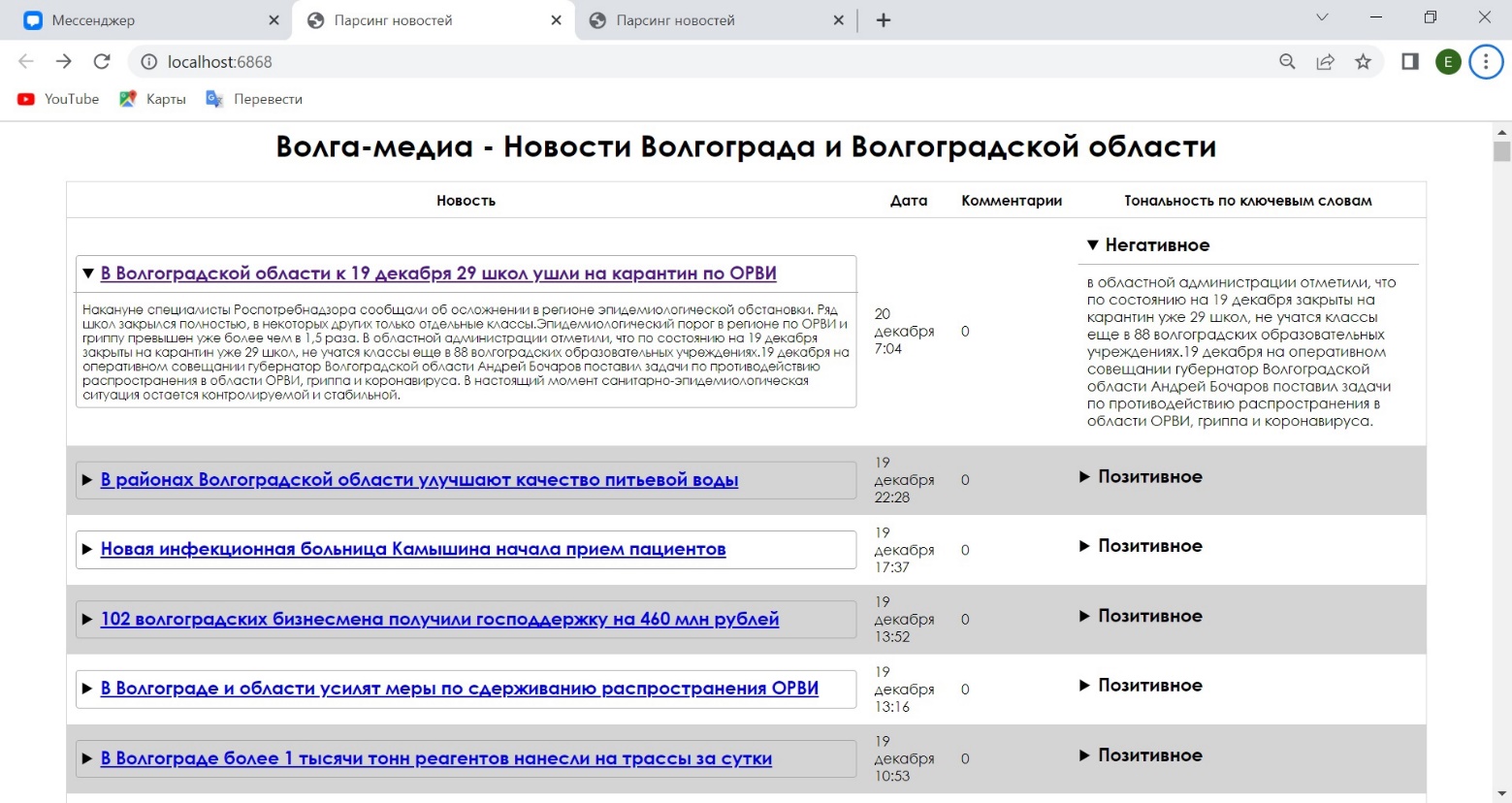


Рисунок 7 – Главная страница сайта

Работа web-интерфейса осуществляется запуском программы index.html. После запуска откроется окно (Рисунок 7), на нем можно просмотреть вывод БД, перейти по ссылкам на страницы со статьями, уточнить предложения с упоминанием персоны или достопримечательности, раскрыть текст самой статьи и увидеть вывод тональности.

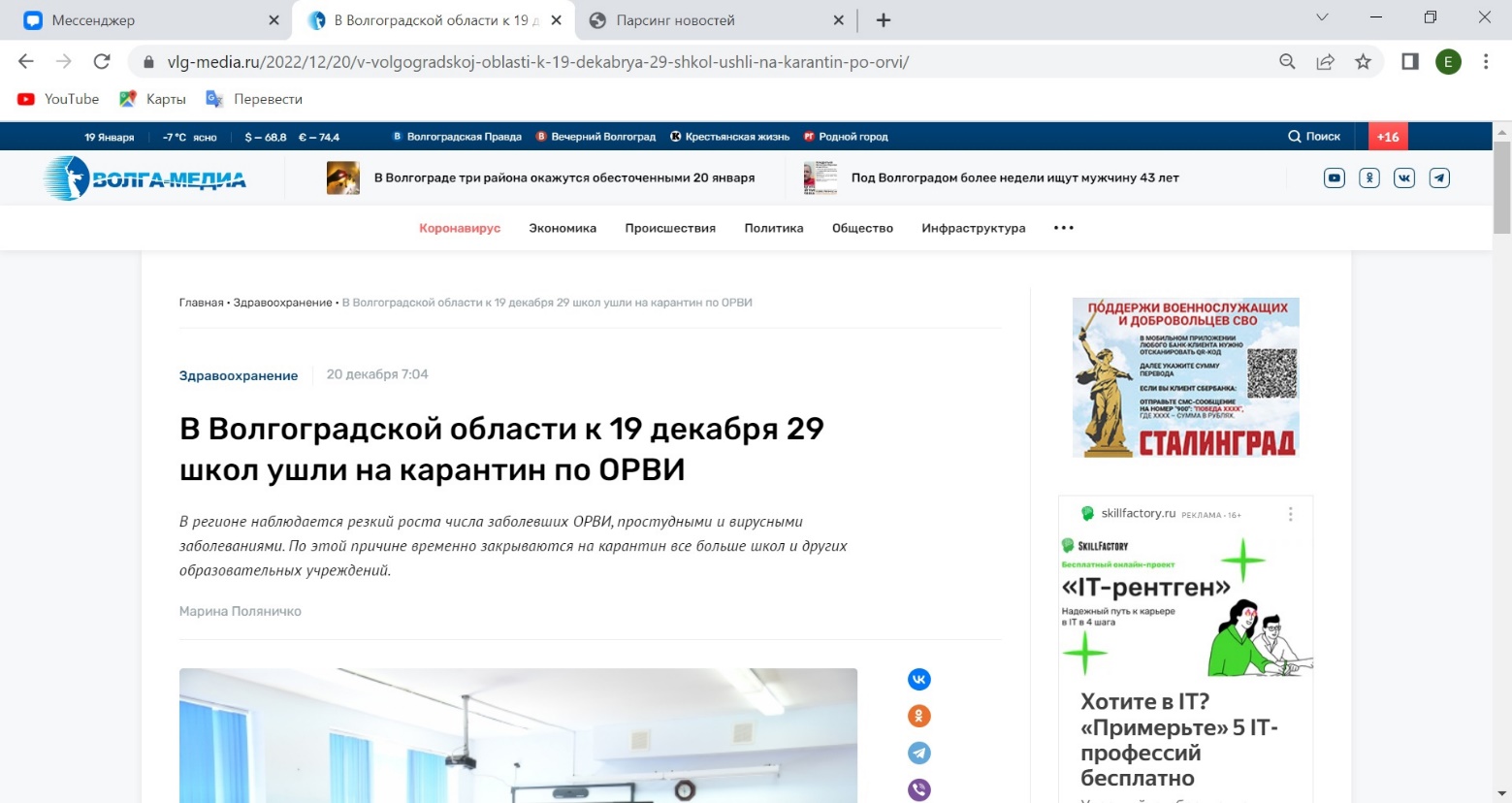


Рисунок 8 – Работа ссылки на страницу статьи

Также на сайте есть вывод word2vec по контекстным синонимам.

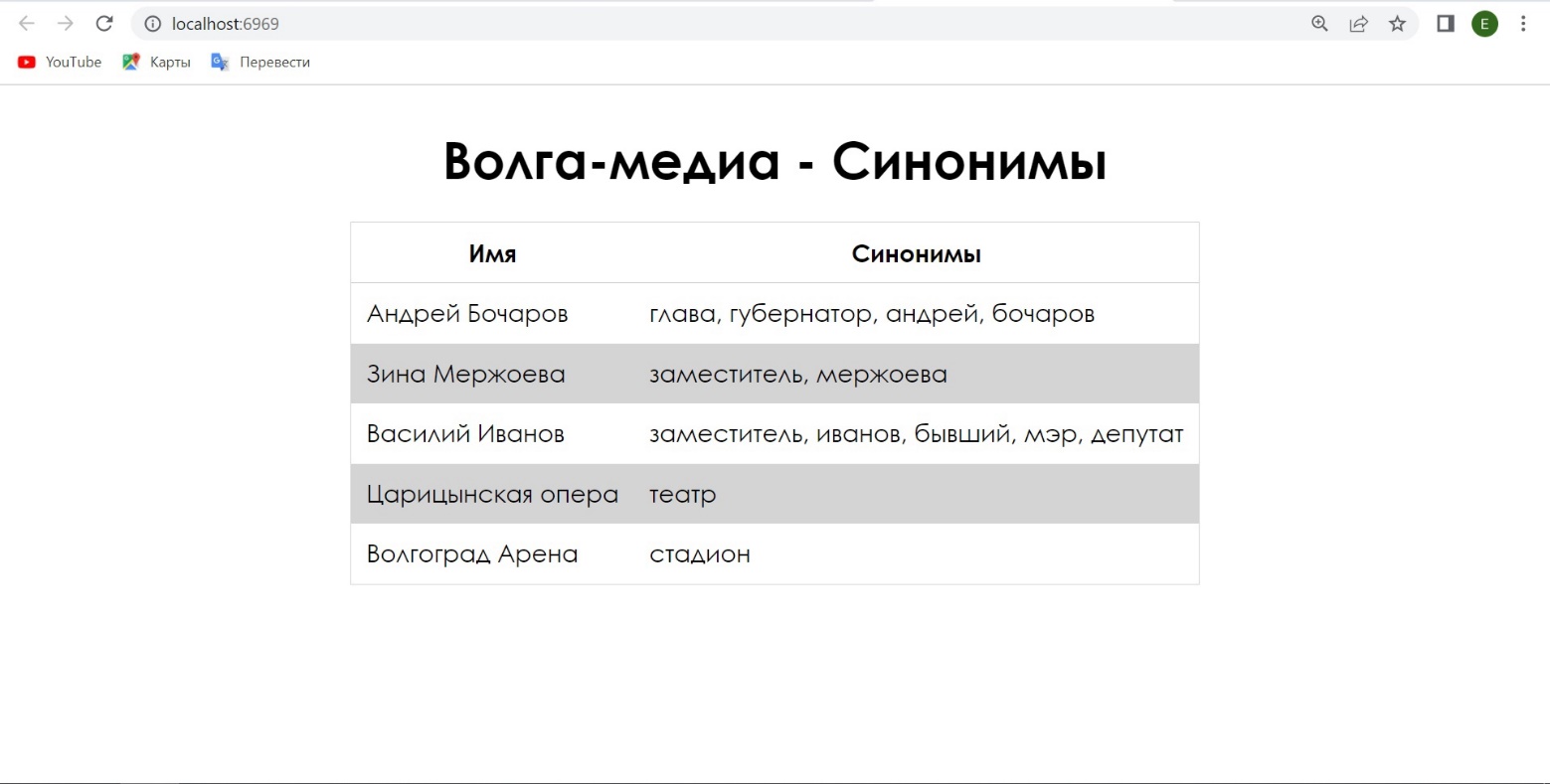


Рисунок 9 – Вывод word2vec на сайт

# Руководство системного администратора

**Установка библиотек для Python:**

sudo apt install python3.7

python3.7 -m pip install -U pip

python3.7 -m pip install -r requrements.txt

**Установка и настройка PyMongo:**

Private Key: 40a23d68-ae98-4243-b8f3-55d89a9e3a8f

**Установка томита-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

**Установка библиотек для анализа тональности:**

python3

import nltk

nltk.download('twitter\_samples')

nltk.download('punkt')

nltk.download('wordnet')

nltk.download('averaged\_perceptron\_tagger\_ru')

nltk.download('stopwords')

pip3 install pymystem3