Министерство науки и высшего образования Российской федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Волгоградский государственный технический университет»

Факультет «Электроники и вычислительной техники»

Кафедра «Системы автоматизированного проектирования и поискового конструирования»

КОНТРОЛЬНАЯ РАБОТА

по дисциплине: «Компьютерная лингвистика»

Выполнили:

студенты 3 курса, гр. ИВТ-365

Меркушов М. С.,

Николенко Д. А.

Проверил:

доц. Коробкин Д.М.

Волгоград, 2022

Содержание

[Первая часть 3](#_heading=h.30j0zll)

[База данных 3](#_heading=h.1fob9te)

[Парсер 4](#_heading=h.3znysh7)

[Вторая часть](#_heading=h.2et92p0) 7

[Томита-парсер](#_heading=h.tyjcwt) 7

[Word2Vec](#_heading=h.3dy6vkm) 8

# Первая часть

## База данных

Используется СУБД MongoDB.

Описание коллекций базы данных:

* news\_vlg – коллекция всех новостей, а также извлеченные из них персоны и достопримечательности.

Значение полей:

1. \_id – id записи в MongoDB;
2. headline – заголовок статьи;
3. text – содержание статьи;
4. url – ссылка на источник;
5. time – время публикации статьи.

* sorted\_news – коллекция с упоминанием персоны или достопримечательности ВО

Значение полей:

1. \_id – id записи в MongoDB;
2. Person/Attraction– имя персоны или название достопримечательности;
3. Sentanse – предложение с упоминанием.

## Парсер

Парсер (main.py) работает в 2 этапа:

Первый этап: парсер проходит по новостной ленте и сохраняет ссылки на новостные статьи в список.

Второй этап: парсер проходит по каждой сохраненной ссылке и получает для каждой новости следующие данные:

1. headline – заголовок статьи;
2. text – содержание статьи;
3. url – ссылка на источник;
4. time – время публикации статьи.

Эти данные составляют новую запись в коллекции MongoDB.

Скриншоты таблиц заполненной парсерами базы данных:

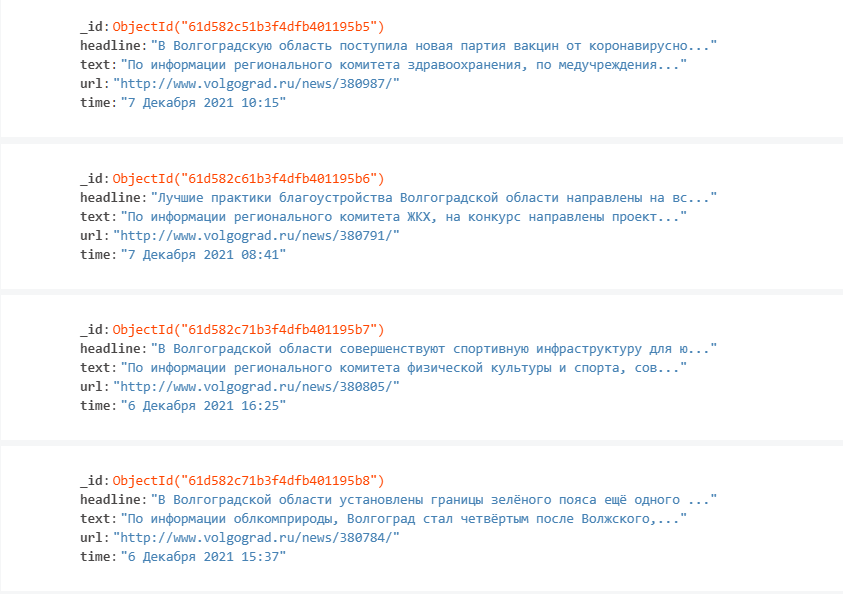


Рисунок 1. Таблица news\_vlg

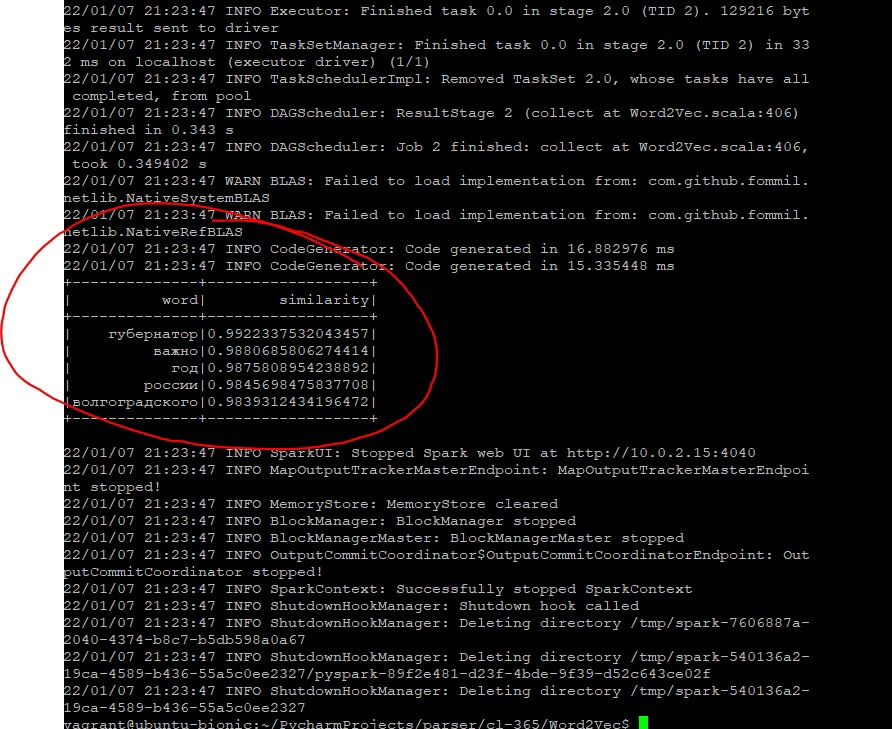


Рисунок 2. Синонимы

# Вторая часть

## Томита-парсер

Грамматики для томита-парсера генерировались на основе данных полученных с сайтов: <https://global-volgograd.ru/person> и <https://avolgograd.com/sights?obl=vgg>

Выделение персон и достопримечательностей и помещение их в таблицу sorted\_news выполняется в скрипте getSentences.py по ходу записи новостей в коллекцию sorted\_news .

Перед запуском скрипта необходимо копировать все файлы в /home/vagrant/tomita-parser/build/bin/.

Для работы томита-парсера необходимо получить из БД текст всех статей и запустить файл для каждого предложения из базы. На выходе получим файл facts.txt, в который будет перезаписываться каждая статья и извлекаются данные. Полученные данные записываются в базу news с полями Person либо Attraction и поле Sentanse

где Person – персоны, Attraction – достопримечательности, Sentanse - предложение с упоминанием.

## Word2Vec

Для построения модели word2vec персоны и достопримечательности были записаны в файлы persons.txt и attaraction.txt. Запускаемый файл useword2vec.py находит синонимы для данных из txt файлов.

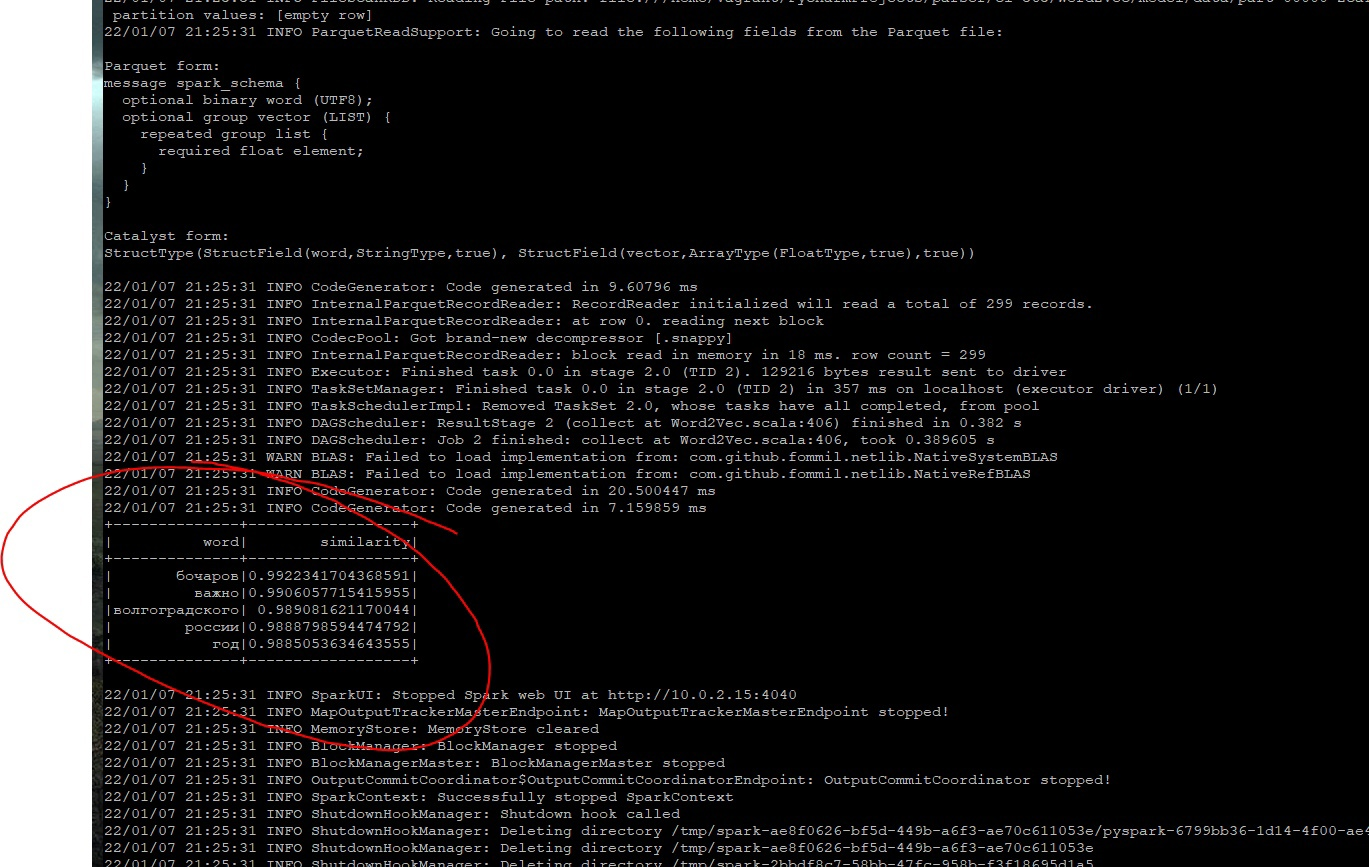


Рисунок 3. Страница контекстных синонимов