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

Линцов И. Н.,

Щербинин В. С.

Проверил:

доц. Коробкин Д.М.

Волгоград, 2022

Содержание

[Первая часть 3](#_Toc43168758)

[База данных 3](#_Toc43168759)

[Парсер 4](#_Toc43168760)

[Вторая часть 9](#_Toc43168761)

[Томита-парсер 9](#_Toc43168762)

[Word2Vec 10](#_Toc43168763)

# Первая часть

## База данных

Используется СУБД MongoDB.

Описание коллекций базы данных:

* news – 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.
6. Person – список персон
7. Object – список достопримечательностей

* synonyms – коллекция контекстных синонимов для каждой персоны и достопримечательности.

Значение полей:

1. \_id – id записи в MongoDB;
2. name – имя персоны или название достопримечательности;
3. synonym – контекстный синоним.

## Парсер

Парсер (parser.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:

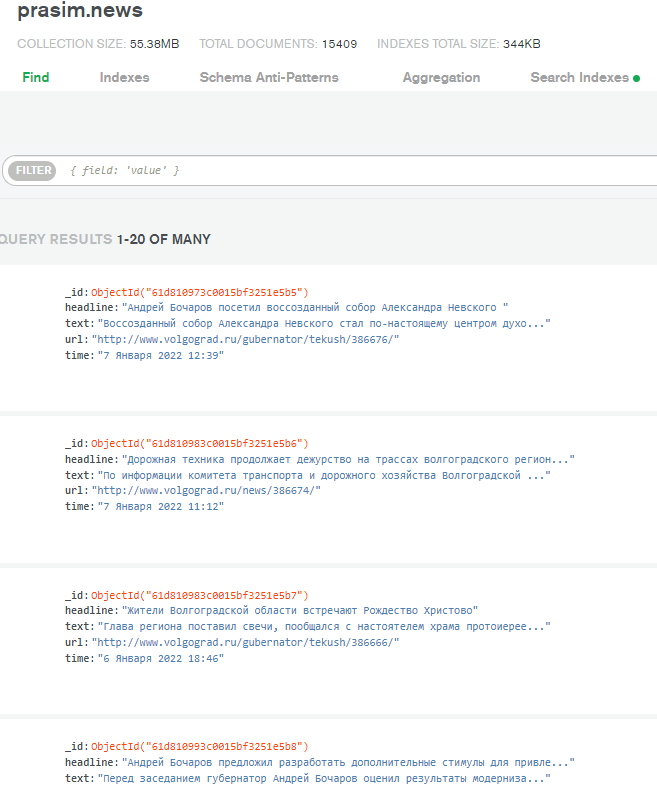


Рисунок 1. Таблица news

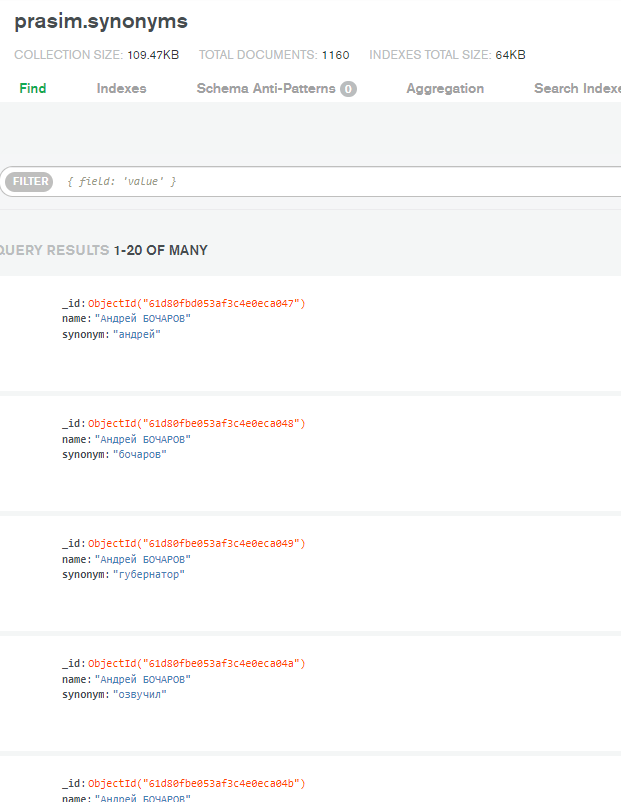


Рисунок 2. Таблица synonymsWeb-интерфейс

Web-интерфейс главной страницы выглядит следующим образом, на нем представлена ссылки на статьи с сайта volgograd.ru, их заголовок и содержание:

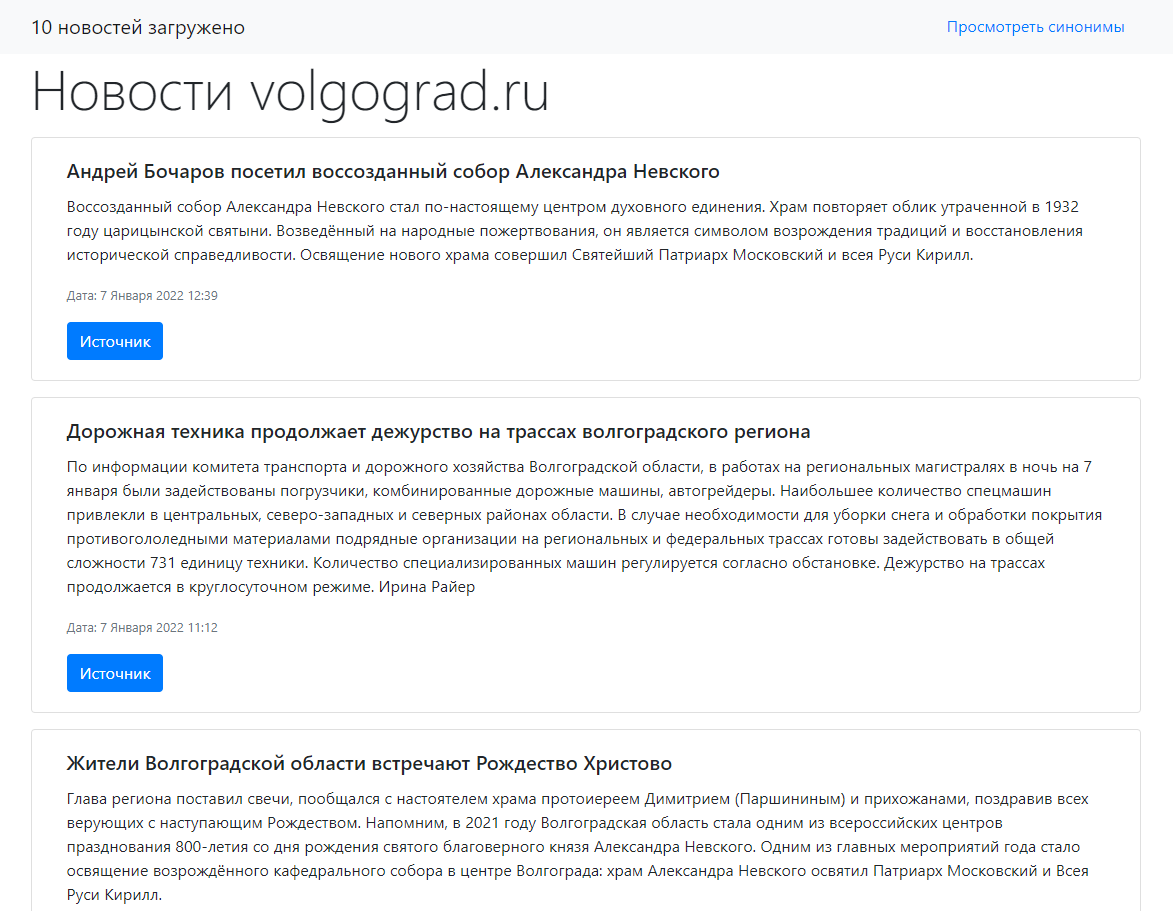


Рисунок 3. Интерфейс главной страницы

# 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 Tomito выполняется в скрипте crawler.py по ходу записи новостей в коллекцию news.

Перед запуском необходимо его поместить в папку с файлами people.txt, places.txt, dic.gzt и config.proto(По умолчанию: лежат вместе в ветке мастер).

## Word2Vec

Модель word2vec была обучена на новостных статьях из базы данных, объем которой составлял примерно ~6000 статей. Модель находится в папке ./models/w2v/. Всю папку нужно перенести в /home/vagrant/. Запуск модуля осуществляется запуском программы /word2vec/[sinonim.py](https://github.com/asuka1211/ComputationalLinguistics/blob/master/word2vec/sinonim.py). Модуль записывает контекстные синонимы в БД и осуществляет следующий вывод:

'Бочаров Андрей'

бочаров

андрей

губернатор

озвучил

глава

…

Модуль имеет следующий web-интерфейс, на нем представлено имя персоны или достопримечательности с контекстным синонимом:

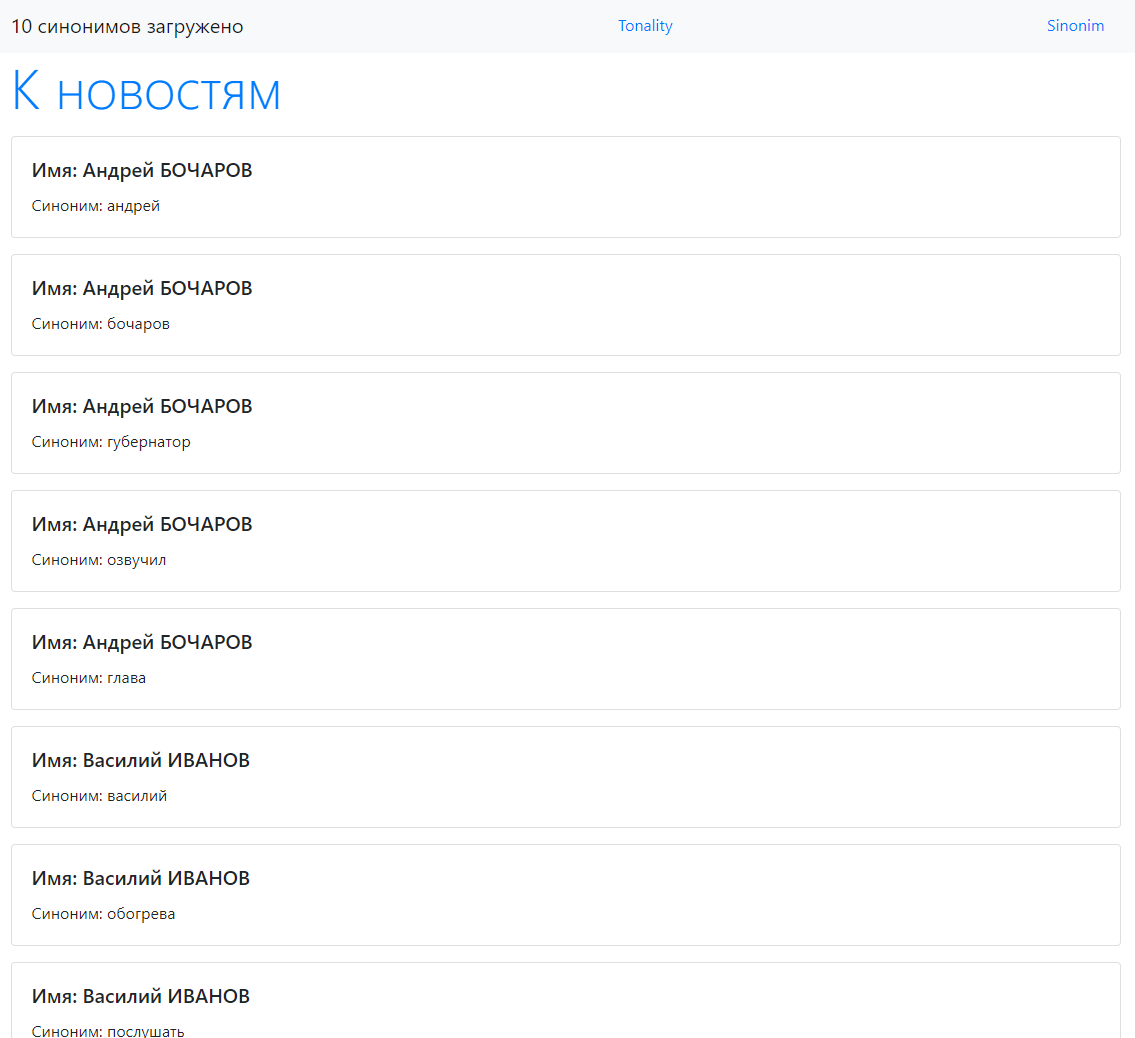


Рисунок 7. Страница контекстных синонимов