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»

Выполнил:

студент гр. ИВТ-360

Донцов С.С.

Проверил:

доц. Коробкин Д.М.

Волгоград, 2022

Содержание

[Парсинг данных и запись в бд 3](#_Toc97971422)

[Томита-парсер 3](#_Toc97971423)

[Word2Vec 4](#_Toc97971424)

## Парсинг данных и запись в бд

## 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 sentence выполняется в скрипте TestSemPart2.2.py

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 TestSemPart2.2.py для каждого предложения из базы. На выходе получим файл output.txt в который будет перезаписываться каждая статья и извлекаются данные. Полученные данные записываются в базу news с полями Politician и Attraction,

где Politician – персоны,

Attraction – достопримечательности.

## Word2Vec

Для построения модели word2vec персоны и достопримечательности были записаны в переменную persons\_and\_sights. Запускаемый файл ModelTest2.3.py находит синонимы для данных.