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»

Дисциплина «Компьютерная лингвистика»

Контрольная работа.

Вариант 2.

Выполнил: студент группы

ИВТ-365 Цыганков Д.С.

Холодов И.В.

Каштанов Р.А.

Проверил: к.т.н., доцент

Коробкин Д.М.

Волгоград, 2024

Оглавление

[1. Введение 4](#_heading=h.gjdgxs)

[2. Первая часть 5](#_heading=h.30j0zll)

[2.1. Структура базы данных 5](#_heading=h.1fob9te)

[2.2. Краулер. 5](#_heading=h.3znysh7)

[3. Вторая часть 6](#_heading=h.2et92p0)

[3.1. Tomita parser 6](#_heading=h.tyjcwt)

[3.2. Spark 6](#_heading=h.3dy6vkm)

[4. Третья часть 7](#_heading=h.1t3h5sf)

[4.1. Сервисы Сбера 7](#_heading=h.4d34og8)

[4.2. Тональность 7](#_heading=h.2s8eyo1)

[5. Бот telegram 8](#_heading=h.17dp8vu)

# Введение

Задача: разработка бота, имеющего функционал для просмотра новостей из базы данных с выводом названия, даты публикации, ссылкой на источник и текста новостей. Содержащий информацию об упоминании VIP-персон и достопримечательностей Волгоградской области, а также тональности новостей, в которых они упоминаются. Имеющий функционал, использующий Суммаризатор и Рерайтер от компании Сбер и вывод контекстных синоним для слов.

# Первая часть

## Структура базы данных

В проекте используется СУБД – SQLite.

Таблица db\_news\_main2 содержит 5 полей:

1. news c полями:

* id – идентификатор новости в БД.
* title – название новости.
* date\_news – дата новости.
* url – ссылка на источник.
* content – текст новости.

## Краулер.

База данных была составлена 17.01.2024 согласно варианту.

Для извлечения информации использовались библиотеки bs4 и selenium для прожимания кнопки “Посмотреть еще” (Загрузить еще новости с ленты)

Сначала файлы записывались в .csv файл, после добавлялись в базу данных.

Краулер бота запускается раз в 2 минуты, чтобы подгрузить новые новости напрямую в базу данных.

# Вторая часть

## Tomita parser

Для функционирования всей программы и подключения томиты все файлы проекта находятся внутри папки build/bin томита-парсера.

## Spark

Бот напрямую инициализирует SparkSession, забирает тексты всех новостей из базы данных, после чего создает Spark DataFrame и затем обучает все требуемые модели.

# Третья часть

## Сервисы Сбера

Для обращения к Суммаризатору и Рераутеру были использованы открытые api.

## Тональность

Для определения тональности был обучен наивный байесовский классификатор из пакета nltk на датасете, представляемый этой библиотекой,

# Бот telegram

Бот написан с использованием библиотеки aiogram.

К сервисам Сбера подключается, отсылая http запросы.

Работает до тех пор, пока не отключится вручную на локальной машине.

Описание сервисов

Для создания бота использовался телеграмм-бот https://telegram.me/BotFather

Ссылка на бота: t.me/NewsSnip\_bot.

Токен бота: 6405501618:AAGPfZz6q5THSrY1BPptvSoy-WfOKK91-L0.

Для написания аннотаций и переписывания текста использовались сервисы Сбера <https://developers.sber.ru/portal/products/summarizer> и <https://developers.sber.ru/portal/products/rewriter>.

Запуск

Запуск осуществляется на локальной машине в ОС Linux 18.04.

Ссылка на проект на github:

https://github.com/rubeusx/CompLing

Все исходники лежат в папке /bin