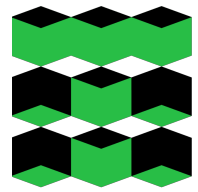
**Министерство науки и высшего образования Российской Федерации**

Федеральное государственное автономное образовательное учреждение

высшего образования



**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ**

**ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Инженерная школа информационных технологий и робототехники

Отделение информационных технологий

Направление подготовки 09.04.04 Программная инженерия

**Отчёт по лабораторной работе №2**

**Веб-скрейпинг**

по дисциплине Представление знаний в системах искусственного интеллекта

Выполнил студент гр. 8ПМ4Л \_\_\_\_\_\_\_ \_\_\_\_\_\_ Сокуров Р.Е.

Подпись Дата Фамилия И.О.

Проверил лаборант ОИТ \_\_\_\_\_\_\_ \_\_\_\_\_\_ Сапегин А.А.

Подпись Дата Фамилия И.О.

Томск 2024 г.

**Цель**

Понимание основных принципов веб-скрейпинга и реализация соответствующей программы (веб-скрейпера, web-scraper).

**Задание**

1.1. Выбрать исполнителя с веб-ресурса: https://www.genius.com.

1.2. Подготовить запрос поиска выбранного исполнителя (по ресурсу).

2. Получить содержимое страницы по URL (ПРИМЕЧАНИЕ: содержимое должность быть представлено в ответе сервера на запрос). Для выполнения данного раздела задания студенту необходимо самостоятельно подобрать требуемые библиотеки.

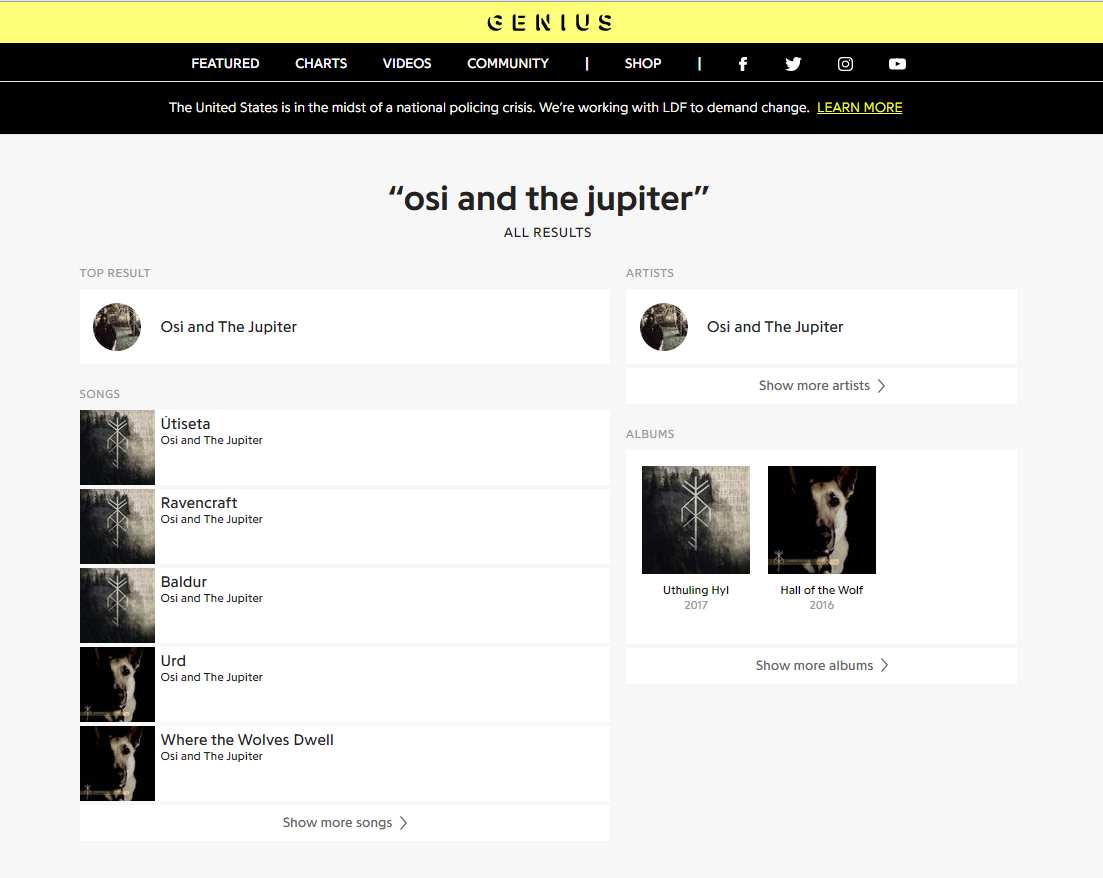


Рисунок 1 – Пример ответа на запрос

3. Получить информацию о “TOP RESULT” и три первых песни (“SONGS”) (ПОДСКАЗКА: для выполнения данного задания вам понадобится дополнительная библиотека).

4. Представить информацию в читаемом виде.

**Ход работы**

На данный момент genius.com заблокировал доступ к получению url страницы поиска выбранного исполнителя. Об этом написано в файле «genius.com/robots.txt»:

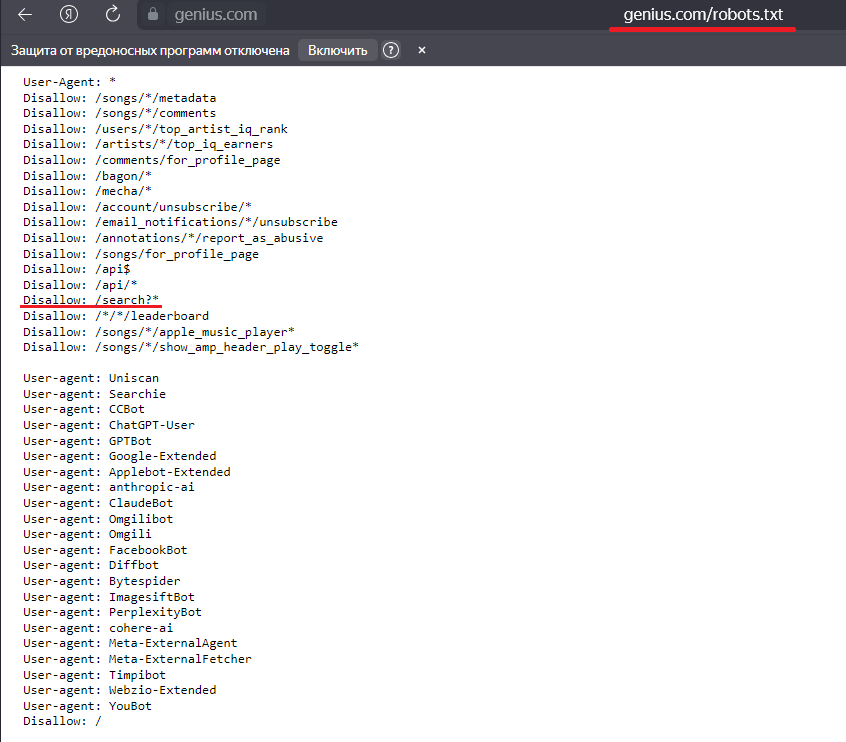


Рисунок 2 – Содержимое файла «robots.txt»

Поэтому для выполнения задания вместо запроса было выполнено получение данных со страницы артиста https://genius.com/artists/«название\_артиста». В качестве артиста была выбрана группа «Самое большое простое число (СБПЧ)». Код и результат его выполнения представлены в таблице 1:

Таблица 1 – Файл «Code.ipynb»

|  |
| --- |
| import requests from bs4 import BeautifulSoup  search\_url = "https://genius.com/artists/Sbpch"  #search\_url = "https://genius.com/artists/Kanye-west" print("Search URL:", search\_url)  Search URL: https://genius.com/artists/Sbpch  # Шаг 2: Отправка GET-запроса и получение содержимого страницы response = requests.get(search\_url)  # Проверка статуса ответа if response.status\_code == 200:  print("Страница успешно загружена")  # Шаг 3: Парсинг содержимого страницы  soup = BeautifulSoup(response.content, 'html.parser')    # Извлечение информации о первом артисте  artist\_link = soup.find('h1', class\_='profile\_identity-name\_iq\_and\_role\_icon u-hair\_bottom\_margin')  if artist\_link:  print("Имя исполнителя:", artist\_link.text)  else:  print("Исполнитель не найден.") else:  print("Ошибка при загрузке страницы:", response.status\_code)  Страница успешно загружена Имя исполнителя: Самое большое простое число (SBPCH)  information\_strs = [] about = soup.find('div', class\_='white\_container') if about:  # Получение текста и фильтрация пустых строк  info = [line.strip() for line in about.text.splitlines() if line.strip()]  for a in info:  if (a != 'Контакты:' and a != 'Паблик в VK'and   a != 'Twitch' and a != 'YouTube' and   a != 'Сообщество ВКонтакте' and a != 'Телеграм-канал' and  a != 'TikTok' and a!= 'Страница ВКонтакте'):  print(a)  information\_strs.append(a)  About Самое большое простое число (SBPCH) Детище поэта Кирилла Иванова и электронного дуэта Ёлочные Игрушки, группа Самое большое простое число началась с вязкого эмбиента, под сопровождение которого Кирилл начитывал необычные болезненно-детские стихи. Затем был нашумевший во всех смыслах этого слова СБПЧ Оркестр, где к Саше, Илье и Кириллу присоединились участники дружественных групп. Всего в записе диска приняли участие 17 разноплановых музыкантов, от человека-горы Стаса Барецкого до 19-летнего рэпера Ларика Сурапова. Одноименный диск стал лучшим русскоязычным альбомом 2008 года по версии журнала «Афиша» Затем это был дуэт, уверенно манипулирующий синтезаторами и драм-машинами, умная лаконичная электроника, движимая живым басом и обогащенная женскими вокализами, разносторонние песни, в которых поэтический речитатив уживается с мелодическими хуками и всё те же необычные стишки. В 2018 году к составу группы присоединились Женя Борзых, чей голос до этого можно было услышать в некоторых песнях, и Стас Астахов — электронный продюсер, известным под псевдонимом Pioneerball. В том же году группу покинул Илья Барамия, переключившись на другой проект — группу АИГЕЛ.  print("Популярные песни этого исполнителя:") artist\_songs = soup.find('div', class\_='mini\_card\_grid') if artist\_songs:  # Получение текста и фильтрация пустых строк  songs = [line.strip() for line in artist\_songs.text.splitlines() if line.strip()]  for song in songs:  if (song != artist\_link.text.strip()):  print(song)  Популярные песни этого исполнителя: Злой (Evil) Комната (Room) Провал (Failure) У нас есть всё (We’ve Got All) Блокада (Blockade) Друг (Friend) Молодость простит (Youth Will Forgive) Вечный взрыв (Eternal Explosion) Нежно (Gently) Всё равно (Whatever) |

# Вывод

В ходе лабораторной работы была изучена технология веб-скрейпинга и был реализован веб-скрейпер для считывания основной информации об артисте с популярного музыкального ресурса genius.com.