

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования

«ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт математики и информационных технологий

Отчет о прохождении производственной технологической практики

Выполнил студент(ка) группы 22307

_____ Меньшиков М.А. _____

Направление подготовки:

___ Программная инженерия ___

Место прохождения практики:

_____ HERE Technologies _____

Сроки прохождения практики:

_____ 30.05.2022 - 09.06.2022 _____

Руководитель практики:

_____ Богоявленская О.Ю. _____

Оценка _____

Дата _____

Петрозаводск 2022 г.

Содержание

Введение	2
Место проведения практики	2
План практики	2
Основная часть	4
Учебные материалы	4
Техническая сторона практики	4
Правила подготовки презентаций	5
Выбранные задачи	6
Tanukifamily	6
Mosoblbank	8
Офисы	8
Банкоматы	11
Заключение	13

1. Введение

1.1. Место проведения практики

Практика проходила в IT-компании **HERE Technologies** в онлайн-формате. **HERE Technologies** — это международная технологическая компания, которая занимается разработкой картографического контента, предоставляет данные о местоположении, а также сопутствующие услуги частным лицам и компаниям. Компания принадлежит консорциуму, в который входят **AUDI AG, BMW Group, Daimler AG, Intel, Navinfo, NVIDIA, Pioneer Corporation, Bosch, Continental** и др. **HERE** является комплексной платформой, которая решает проблему фрагментированных и разрозненных источников данных в совместной среде, позволяя предприятиям монетизировать свои самые ценные активы: а именно данные. Продукты компании используются в многочисленных пакетах программного обеспечения, включая системы навигации; доступен для веб-браузеров и телефонов на различных платформах. Карты включают такие функции, как поиск по карте, спутниковые карты, построение маршрутов, 3D-карты, отображение пробок в реальном времени.

С 2012 года компания активно сотрудничает с университетами по всему миру, создавая совместные проекты, мероприятия и обучающие программы. Приоритетами компании являются:

1. развитие активного сообщества;
2. улучшение профессиональных навыков;
3. постоянная экспертная поддержка.

1.2. План практики

В рамках практики мы проходили урезанную онлайн программу компании по сбору, анализу и визуализации пространственных данных: **HERE Data Engineering**.

Задачами данной программы являются:

- Ознакомиться с обучающими материалами, которые включают: tutoriales по работе с библиотеками python для извлечения гео-данных, мастер-классы по оформлению презентаций, онлайн-курсы с примерами извлечения данных с веб-ресурсов.
- Освоить сервисы **Here Platform, Python, Scrapy**.
- Разработать собственные модели для парсинга сайтов;
- Подготовить отчётные материалы, опубликовать данные на **GitLab**.

Данные, которые требовалось извлекать с веб-ресурсов:

- Адреса точек с геопривязкой (координаты).
- Контактная информация.
- Часы работы.
- Другая сопутствующая информация.

Практика разбита на следующие этапы (рис. 1):

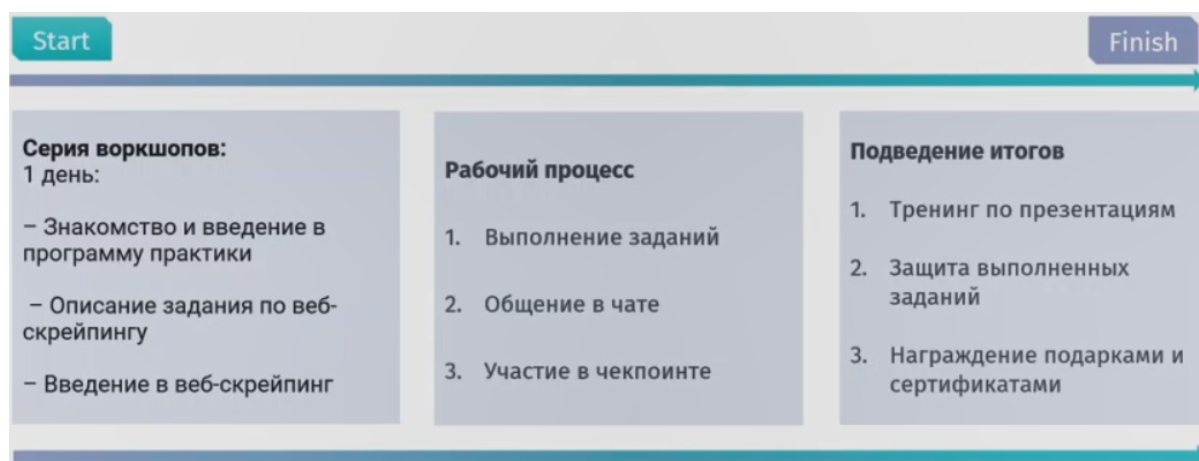


Рис. 1. Этапы практики.

Для успешного прохождения практики было необходимо:

- Извлечь гео-данные минимум с двух сайтов.
- Оформить Python-скрипты извлечения данных по стандарту фреймворка `scrapy`.
- Оформить извлечённые данные в формате **geojson**.
- Визуализировать данные на одной из существующих картографических платформ.
- Загрузить полученные результаты в репозиторий компании на **gitlab**.
- Подготовить презентацию по проделанной работе и выступить с защитой в заключительный день практики.

Предъявлялись следующие критерии оценки итоговых работ:

1. Уровень сложности реализации.
2. Количество распарсенных сайтов.
3. Качество данных.
4. Качество защиты.

2. Основная часть

2.1. Учебные материалы

2.1.1. Техническая сторона практики

Для решения поставленной задачи в рамках практики необходимо было обладать следующим стеком технологий (рис. 2):

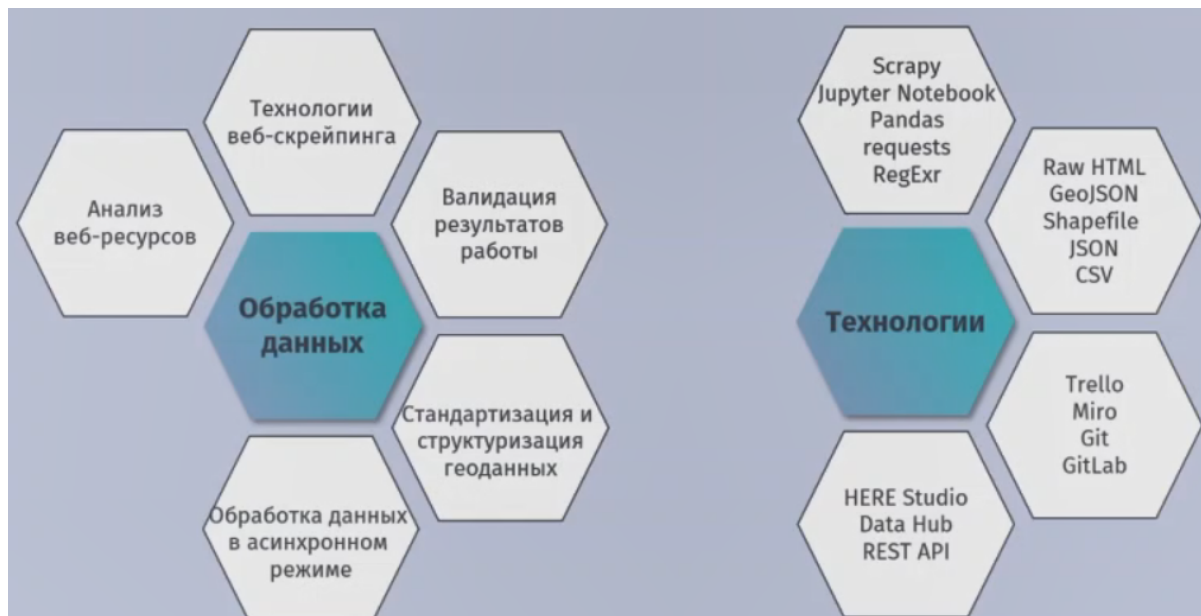


Рис. 2. Необходимый стек технологий.

В рамках одного из проводимых воркшопов были даны базовые навыки по работе с фреймворком **scrapy**, на примере показана настройка виртуального окружения для работы скриптов и показан тестовый вариант скрапинга сайта.

При решении самостоятельных задач были сделаны рекомендации по используемым технологиям для парсинга сайтов:

- Scrapy (скрейпинг).
- BeautifulSoup (библиотека для парсинга).
- Requests (**HTTP**-запросы).
- Pandas (Обработка данных).
- GeoPandas (Обработка пространственных данных).

Для визуализации данных были сделаны следующие рекомендации:

1. **HERE map API for JavaScript.**
2. **HERE Studio & Data Hub.**
3. **HERE Geojson Viewer.**

2.1.2. Правила подготовки презентаций

В рамках практики был проведён мастер-класс по созданию презентаций для выступлений в онлайн/форматах и даны лайфхаки по подготовке к выступлениям.

В начале построения презентации необходимо ответить на следующие вопросы:

- Какая цель презентации?
- Для кого вы выступаете?
- Что вы хотите, чтобы аудитория сделала после презентации?

Рекомендуемый порядок создания презентации:

1. Мозговой штурм.
2. План презентации.
3. История и подача.
4. Репетиция.

Лайфхаки запоминающихся презентаций:

- Контраст и непредсказуемость.
- Иконки это понятно и круто.
- **Google fonts** - найдите свой шрифт; главное без засечек.
- Больше сочных фото (фотостоки не равно google).
- Создайте свой стиль и фишку подачи.

Способы репетиции перед выступлением:

- Репетировать вслух, а не про себя.
- Записывайте себя на видео.
- Научитесь рассказывать речь, а не читать.
- Будьте энергичными, проявляйте эмоциями.
- Говорите простым понятным языком.

Лайфхаки качественного выступления:

1. Помните о целевой аудитории.
2. Сделайте свою речь динамичной.
3. Напишите структуру выступления.
4. Соблюдайте регламент. Больше практики.
5. Всегда выводите в позитив. Выбранные задачи.

2.2. Выбранные задачи

В рамках практических задачи были выбраны следующие сайты для извлечения геоданных:

- <https://tanukifamily.ru/tanuki/restaurants/> (Category: Restaurant);
- <https://mosoblbank.ru/offices/> (Category: Bank, ATM).

2.3. Tanukifamily

Страница сайта выглядела следующим образом (рис. 3):

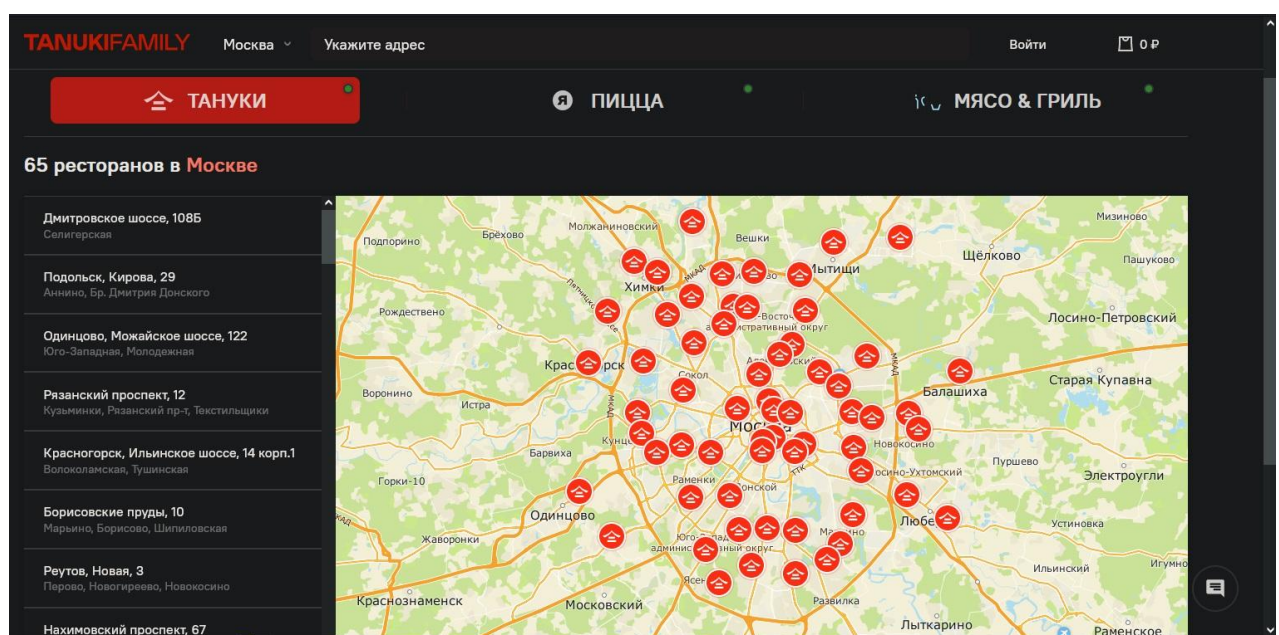


Рис. 3. Страница tanukifamily.ru.

В результате анализа удалось найти **GET**-запрос, который возвращает гео-информацию по отображаемым точкам (рис. 4):

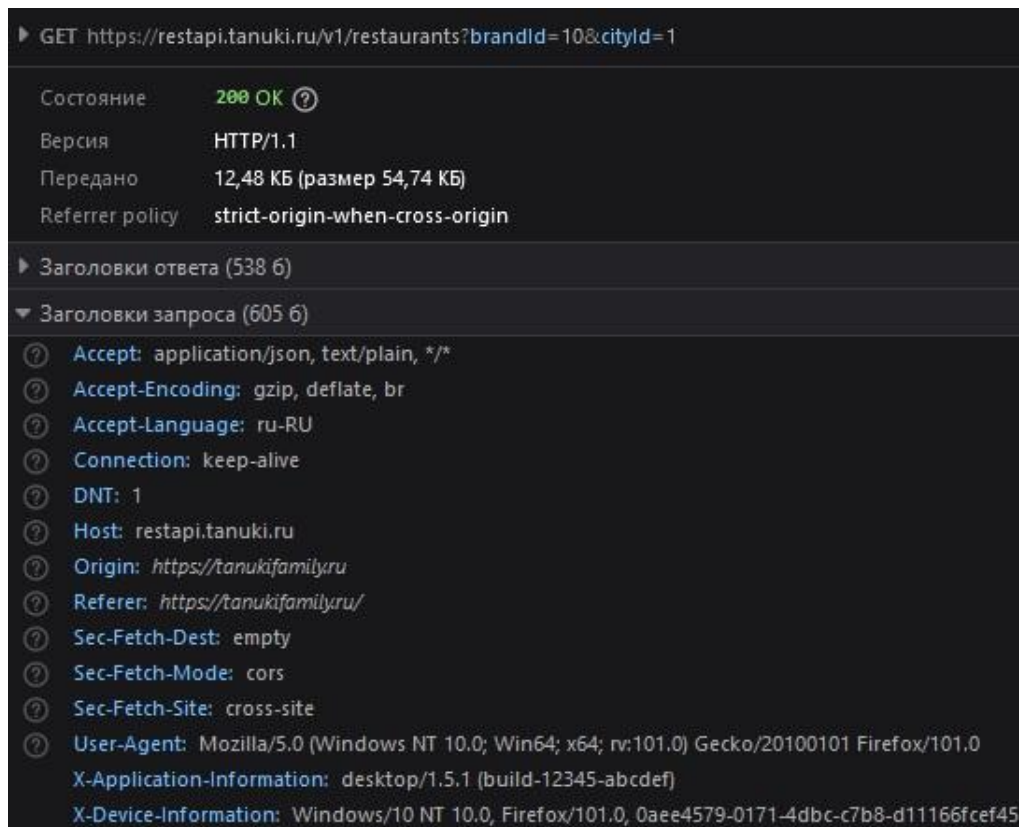


Рис. 4. **GET**-запрос на получение списка с информацией о ресторанах.

Итоговый скрипт работал следующим образом:

1. Построить **GET** запрос для получения информации о ресторанах.
2. Отдельно извлечь город, в котором находится ресторан (лемматизируем).
3. Перевести полученные данные в **geojson**-формат.

Извлечённые данные:

3. координаты ресторана (долгота, широта);
4. телефон ресторана;
5. часы работы ресторана;
6. адрес ресторана (страна, город, улица, номер);
7. ближайшее метро.

После извлечения гео-данных была построена визуализация точек на карте с помощью ресурса **Geojson Visualiser** (рис. 5):

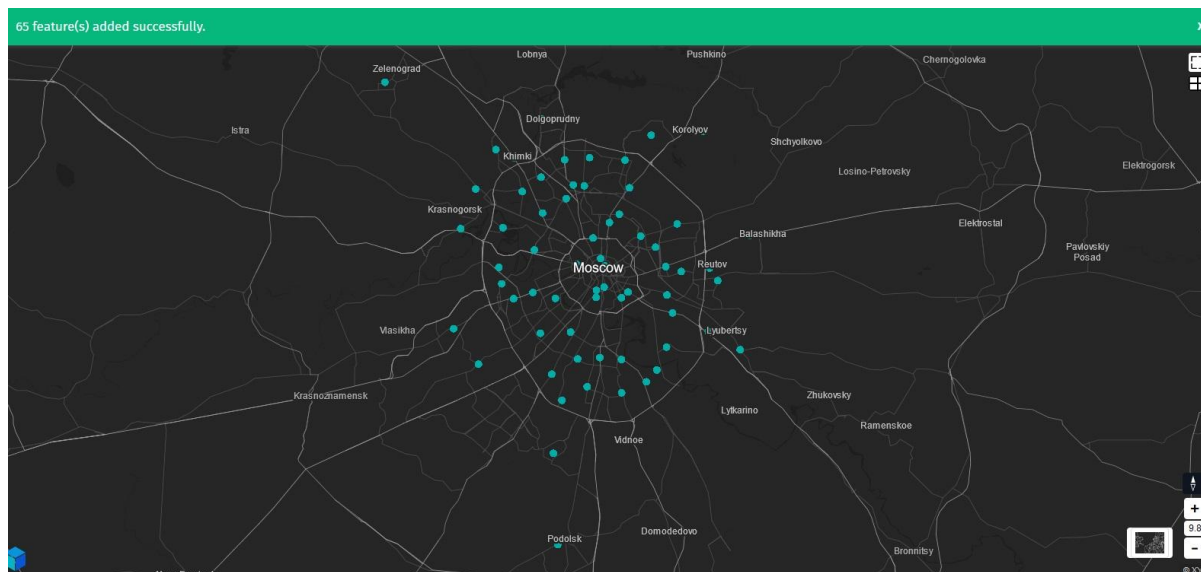


Рис. 5. Визуализация точек с сайта **tanukifamily.ru**.

7.1. Mosoblbank

Для сайта **mosoblbank.ru** по заданию требовалось извлечь гео-данных офисов и банкоматов. Данная задача была разбита на две подзадачи, для решения которых были реализованы отдельные скрипты извлечения данных.

7.1.1. Офисы

Страница сайта со списком офисов выглядит следующим образом (рис. 6):

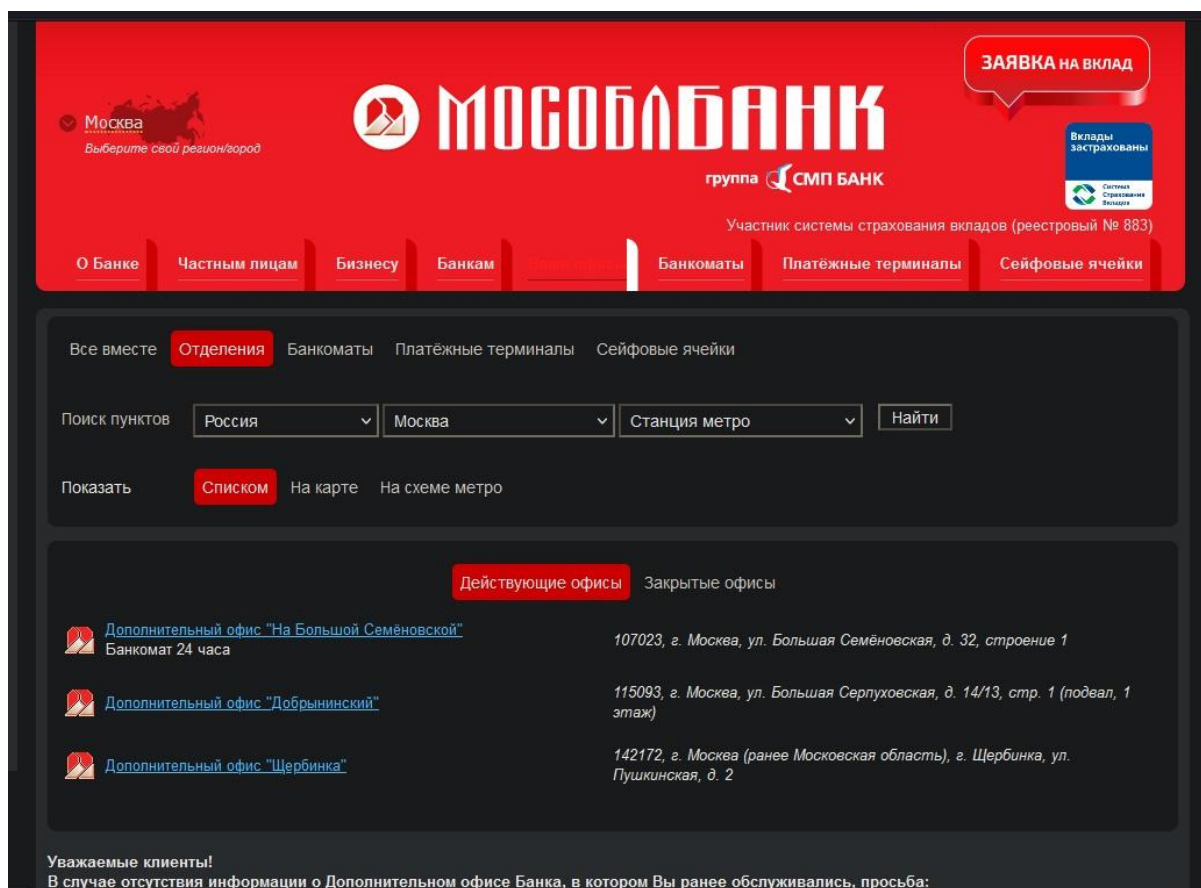


Рис. 6. Страница со списком офисов.

В результате анализа удалось найти **POST**-запрос, который возвращает список с информацией об офисах в зависимости от выбранного региона (рис. 7):

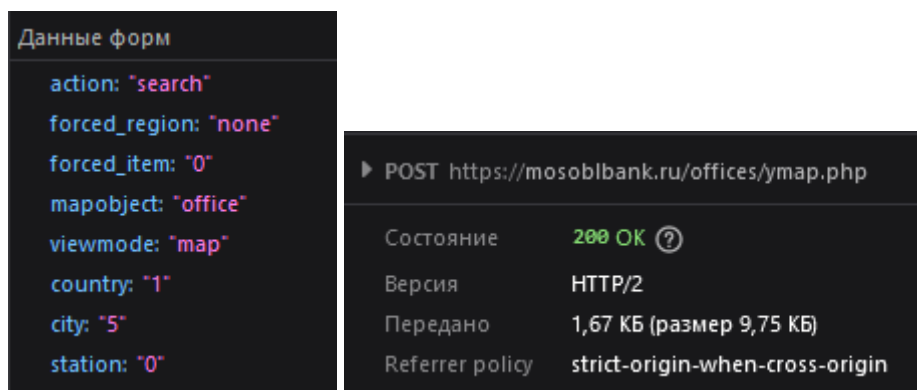


Рис. 7. POST-запрос на получения списка офисов

Идентификаторы регионов были указаны в **HTML**-разметке (рис. 8):

```
<select id="city" class="" name="city">
  <option value="5" selected="">Москва</option>
  <option value="74">Московская область</option>
  <option value="116">Санкт-Петербург</option>
  <option value="95">Ленинградская область</option>
  <option value="94">Архангельская область</option>
  <option value="90">Астраханская область</option>
  <option value="3">Белгородская область</option>
  <option value="83">Владимирская область</option>
```

Рис. 8. Идентификаторы регионов

Итоговый скрипт работал следующим образом:

1. Извлечь идентификаторы доступных регионов.
2. Построить **POST** запрос для получения **json**-списка с информацией об офисах для конкретного региона.
3. Перевести полученные данные в **geojson**-формат.

Извлечённые данные:

- координаты офисы (долгота, широта);
- общие телефоны банка;
- часы работы офиса;
- адрес офиса (postcode, город, улица, номер);
- название офиса.

После извлечения гео-данных была построена визуализация точек на карте с помощью ресурса **Geojson Visualiser** (рис. 9):

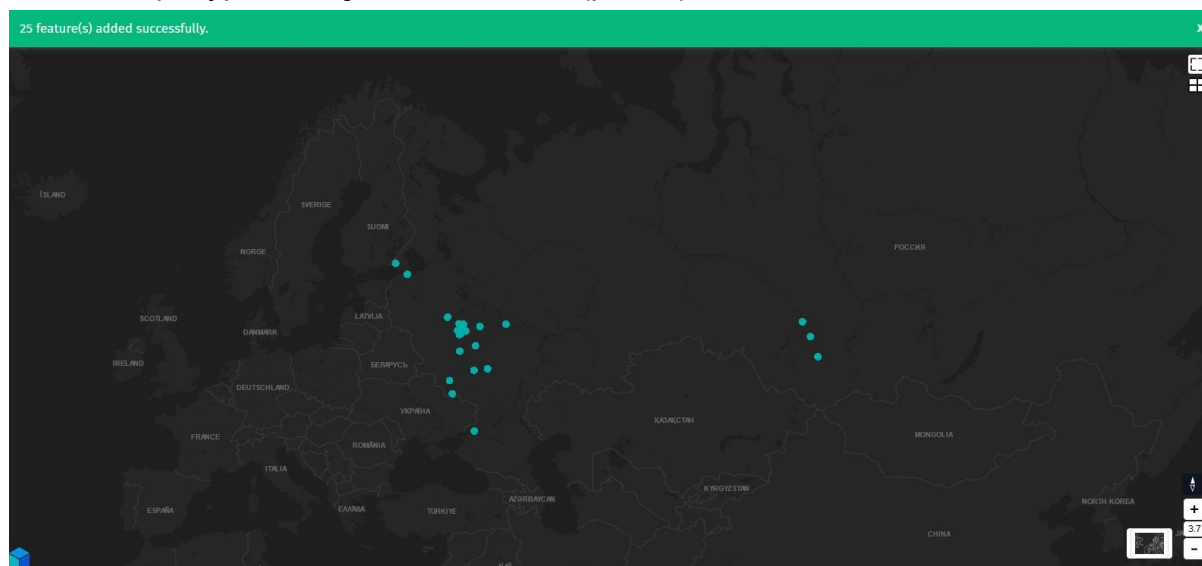


Рис. 9. Визуализация точек-офисов с сайта **mosoblbank.ru**

7.1.2. Банкоматы

Страница сайта со списком офисов выглядит следующим образом (рис. 10):

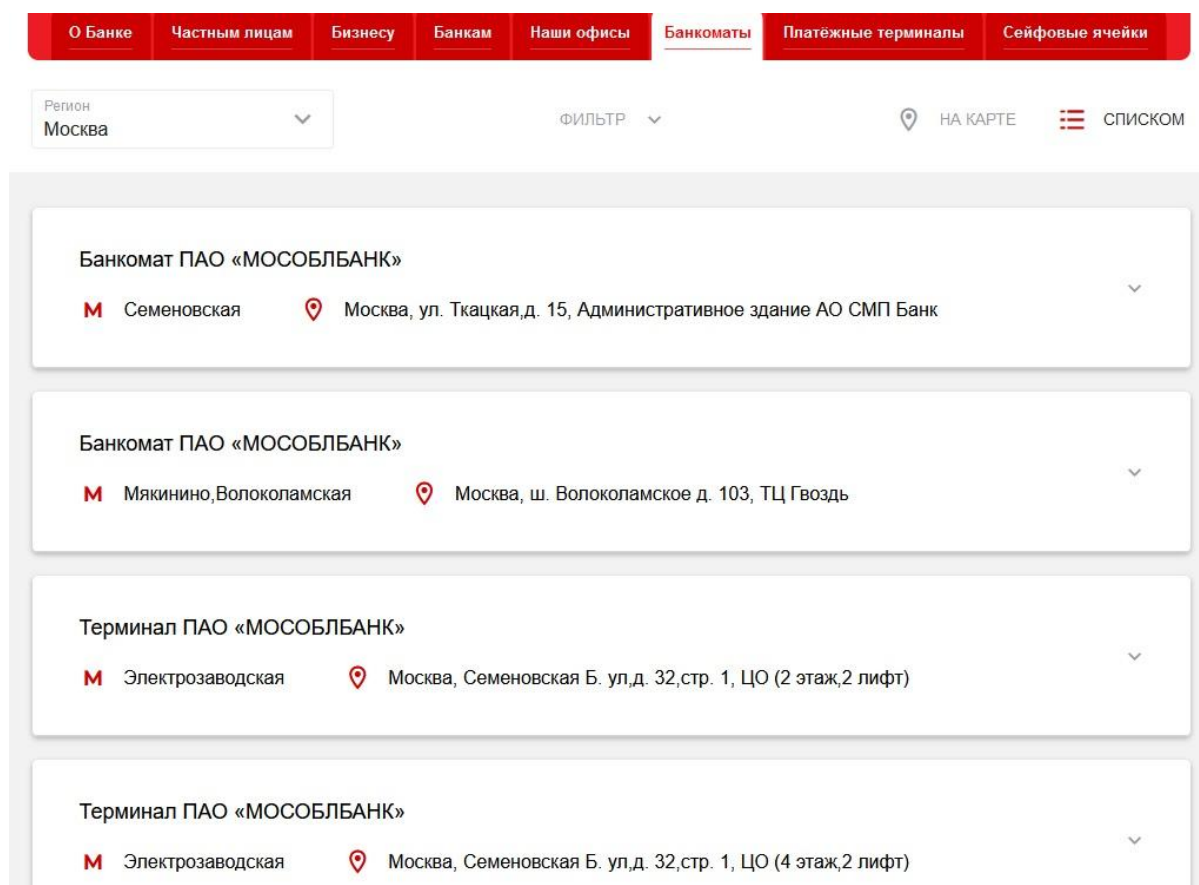


Рис. 10. Страница со списком банкоматов.

В результате анализа оказалось, что информация о банкоматах приходит с использованием **AJAX**-технологии, поэтому было нельзя получить данные, используя простые **HTTP**-запросы. В связи с этим была использована библиотека **selenium** для взаимодействия со страницей и парсинга **HTML**-разметки

Итоговый скрипт работал следующим образом:

1. Воспользоваться библиотекой **selenium** для интерактивного взаимодействия с сайтом.
2. Отобразить список доступных банкоматов для каждого региона.
3. Нажать на область вывода подробной информации для каждого банкомата и извлечь данные.
4. Перевести полученные данные в **geojson**-формат.

Извлечённые данные:

- координаты банкомата (долгота, широта);
- общие телефоны банка;

- часы работы банкомата;
- адрес банкомата (город, улица, номер, название здания);
- ближайшие метро;
- доступные транзакции.

После извлечения гео-данных была построена визуализация точек на карте с помощью ресурса **Geojson Visualiser** (рис. 11):

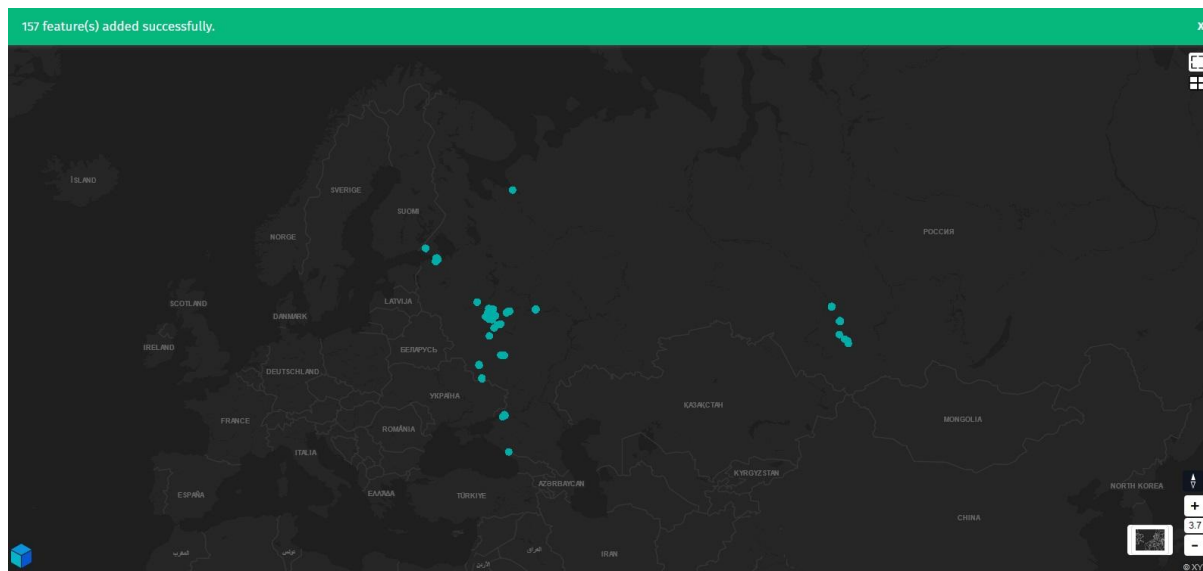


Рис. 11. Визуализация точек-банкоматов с сайта mosoblbank.ru

8. Заключение

По итогам практики были получены следующие результаты:

1. Выбраны сайты для проведения извлечения гео-данных: tanukifamily.ru, mosoblbank.ru.
2. Проведён анализ сайтов и сформулированы алгоритмы извлечения данных.
3. При реализации алгоритмов извлечены гео-данные использовался **scrapy**.
4. Полученные гео-данные оформлены в формате **geojson**.
5. Сделана визуализация полученных гео-данных с помощью ресурса **Geojson Viewer**.
6. Общий объём извлечённых данных составил 247 точек.
7. Изучены учебные материалы по составлению презентаций, и работе с библиотеками извлечения данных с веб-ресурсов.