

Получение данных

Аксентьев Артем Алексеевич

Chillers

17 марта 2020 г.

О чем пойдет речь...

- 1 Что есть данные?
- 2 Хранение данных
- 3 Как получать данные
 - Использование API
 - Web scraping
- 4 Домашнее задание

Зачем собирать данные?

Что такое статистика?

Статистика – плохой фонарный столб. Ничего не освещает, но на неё можно опереться

Зачем собирать данные?

Что такое статистика?

Статистика – плохой фонарный столб. Ничего не освещает, но на неё можно опереться

Пример статистики

Тут должна быть статистика по авариям, но сайт ГИБДД не работал

CSV файлы

Набор данных Ирисы Фишера

```
1  длина чашелистика(см),ширина чашелистика (см),длина  
   ↳ лепестка (см),ширина лепестка (см)  
2  5.1,3.5,1.4,0.2,Iris-setosa  
3  4.9,3.0,1.4,0.2,Iris-setosa  
4  4.7,3.2,1.3,0.2,Iris-setosa  
5  4.6,3.1,1.5,0.2,Iris-setosa  
6  5.0,3.6,1.4,0.2,Iris-setosa  
7  5.4,3.9,1.7,0.4,Iris-setosa  
8  4.6,3.4,1.4,0.3,Iris-setosa  
9  5.2,4.1,1.5,0.1,Iris-setosa  
10 5.5,4.2,1.4,0.2,Iris-setosa  
11 4.9,3.1,1.5,0.1,Iris-setosa  
12 5.0,3.2,1.2,0.2,Iris-setosa  
13 5.5,3.5,1.3,0.2,Iris-setosa  
14 4.9,3.1,1.5,0.1,Iris-setosa
```

XML файлы

Сведения о модернизации региональных систем общего образования

```
1 <part07>
2   <record>
3     <region>Республика Адыгея</region>
4     <avg_general>13043.500000000000</avg_general>
5     <avg_teacher>12855.000000000000</avg_teacher>
6     <procent>1.390000</procent>
7   </record>
8   <record>
9     <region>Республика Башкортостан</region>
10    <avg_general>16459.600000000000</avg_general>
11    <avg_teacher>16459.600000000000</avg_teacher>
12    <procent>1.300000</procent>
13  </record>
```

JSON файлы

Перечень подведомственных учреждений Минтруда РБ

```
1 [{
2   "name": "ГКУ Центр занятости населения города Уфы",
3   "short_name": "ЦЗН г. Уфы",
4   "address": "450006, г. Уфа, бульвар Ибрагимова, 47/1",
5   "phone": "(347) 251-51-55",
6   "email": "ubtczn@bashkortostan.ru",
7   "site": "http://ufa.bashzan.ru",
8   "coords": {
9     "lat": 54.742819,
10    "lng": 55.962913
11  }
12 }, {
```

API

API (application programming interface)

Описание способов (набор классов, процедур, функций, структур или констант), которыми одна компьютерная программа может взаимодействовать с другой программой

Виды API

- 1 Web-API;
- 2 Программное API.

web API

Используется в веб-разработке, как правило, определённый набор HTTP-запросов, а также определение структуры HTTP-ответов, для выражения которых используют XML или JSON форматы.

API

HTTP

Определение

HTTP (HyperText Transfer Protocol — «протокол передачи гипертекста») — протокол прикладного уровня передачи данных изначально — в виде гипертекстовых документов в формате «HTML», в настоящий момент используется для передачи произвольных данных.

GET запросы

Используется для запроса содержимого указанного ресурса.
<https://mai.ru/education/schedule/detail.php?group=M3O-111Б-20&week=14>

API

Пишем код

Вывод погоды в командную строку

- 1 Пользователь открывает нашу программу с передачей в качестве аргумента города;
- 2 Программа выводит температуру, силу и направление ветра.

web scraping

Определение

Web scraping — это технология получения веб-данных путем извлечения их со страниц веб-ресурсов. Веб-скрейпинг может быть сделан вручную пользователем компьютера, однако термин обычно относится к автоматизированным процессам, реализованным с помощью кода, который выполняет GET-запросы на целевой сайт.

Виды

- 1 «Копипаста» вручную;
- 2 Сопоставление текстовых шаблонов;
- 3 Синтаксический анализ HTML;
- 4 Вертикальная агрегация.

Как формируется html - страница

Back-end

- Обработка поступившего запроса;
- Запросы информации из БД;
- Наполнение страницы данными;

Как формируется html - страница

Back-end

- Обработка поступившего запроса;
- Запросы информации из БД;
- Наполнение страницы данными;

Front-end

- Разметка страницы;
- Дизайн страницы;
- Минимальные, невидимые глазу операции (например, проверки на правильность задания паролей при регистрации на сайтах)

web scraping

Как происходит?

1. Определить цель;

web scraping

Как происходит?

- 1 Определить цель;
- 2 Скачать файлы html;

web scraping

Как происходит?

- 1 Определить цель;
- 2 Скачать файлы html;
- 3 Проанализировать архитектуру файла;

web scraping

Как происходит?

- 1 Определить цель;
- 2 Скачать файлы html;
- 3 Проанализировать архитектуру файла;
- 4 Распарсить информацию с html в удобный вид (json, xml, csv, plain text).

web scraping

Python

Модуля для парсинга

- 1 Re - регулярные выражения `a\.?;`

web scraping

Python

Модуля для парсинга

- 1 Re - регулярные выражения `a\.?;`;
- 2 BeautifulSoup;
- 3 lxml;

web scraping

Python

Модуля для парсинга

- 1 Re - регулярные выражения `a\.*;`;
- 2 BeautifulSoup;
- 3 lxml;
- 4 scrapy - фреймворк;

web scraping

Парсер расписания с сайта МАИ

Зачем?

Иногда удобно видеть расписание занятий не на неделю, а на месяц или даже весь семестр.

Что нужно?

Пройти по всем страницам с расписанием и сложить полученную информацию.

Домашнее задание

Домашнее задание

Решить какую-либо задачу с помощью API и технологии web-scraping. Иными словами необходимо реализовать работу с любым API и реализовать получение любой информации с помощью web-scraping.

Как сдать?

Послать решение в сообщения группы. В каждом файле обязательно укажите, что вы сделали и сайт которым пользовались.

Домашнее задание

Пример задания

- API:
 - ❶ Сделать простой конвертор валют (у ЦБ РФ есть API);
 - ❷ Реализовать поиск маршрута (например на чем можно добраться) между двумя городами (API Яндекс.Расписаний);
- Web Scraping:
 - ❶ С доски объявления спарсить все объявления о продаже одной марки автомобиля. Найти среднюю цену автомобиля по годам (авито, авто ру и т.д.);
 - ❷ У какой-либо публичной персоны в инстаграмм узнать какая локация чаще всего используется на фотографиях.

Спасибо за внимание!