

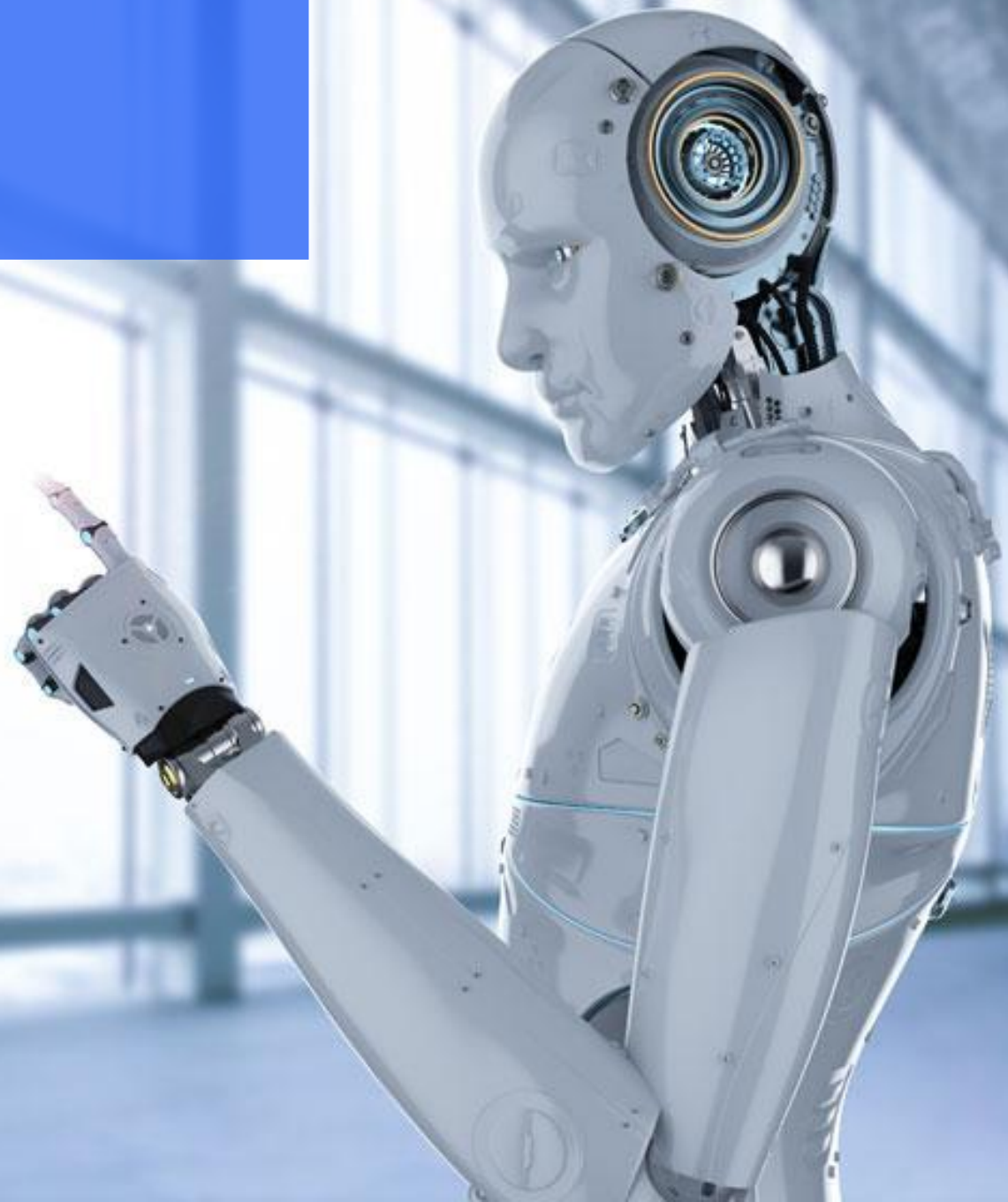


УНИВЕРСИТЕТ
ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

WEB-SCRAPPING

ПЛАН

- Интернет и Данные
- Web Scrapping
- Web APIs
- Web scrapping vs API
- Requests + BeautifulSoup
- Scrappy

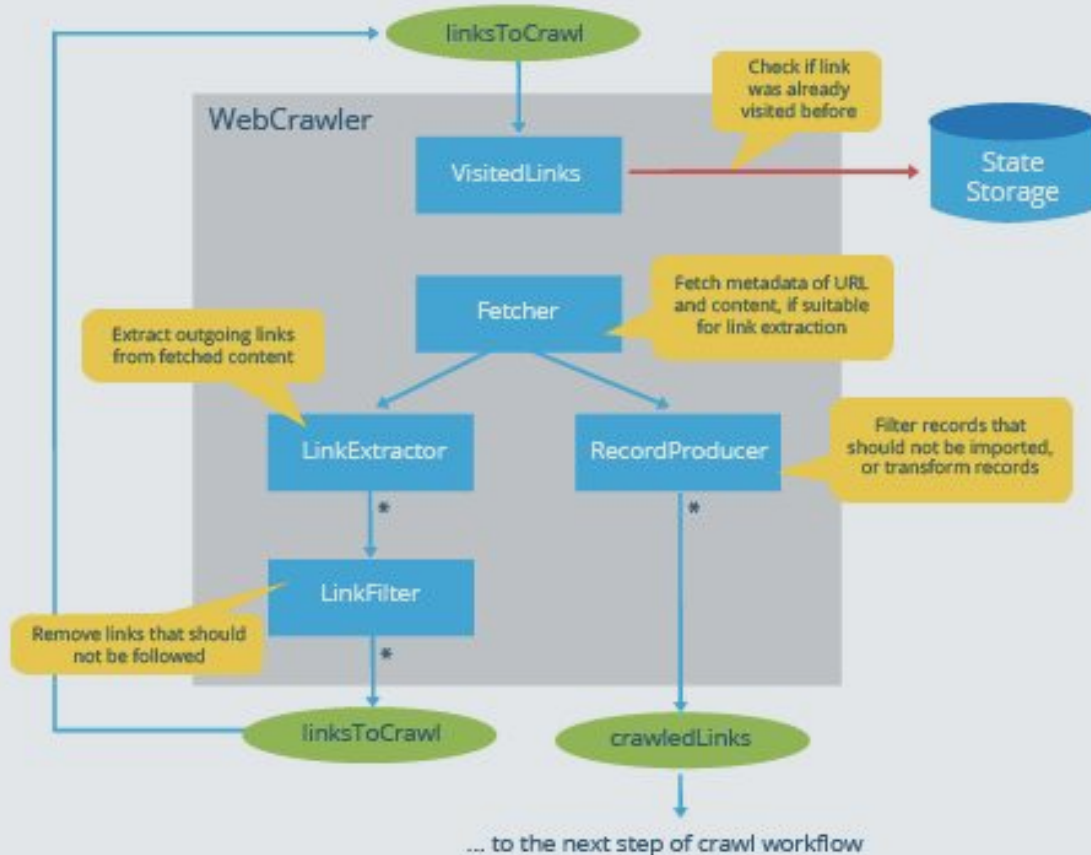


Данные и Интернет



- Более 474,000 Tweets в минуту в 2019!
- Youtube загружается > 400 hours видео в минуту! В 2019, пользователи в среднем просматривали 4,333,560 роликов в минуту.
- Instagram > 100 миллионов фото в день
- 1,209,600 новых данных производится в среднем в день медиа порталами.
- Большая 4-ка (Apple, FB, Google, Amazon) хранит более 1,200 Петабайт данных

Web – Crawling/Scrapping



Web Crawler («Bot» или «Spiden») – программа для автоматизированного поиска документов и информации в интернете. Основная задача – автоматизация множественных переходов по ссылкам. Поисковики используют «Веб-Краулеры» для обхода новых документов и обновления поискового индекса

Самый известный Веб-Краулер на сегодняшний день это Googlebot.

Web APIs

APIs - Набор правил, определяющих, как можно взаимодействовать с конкретным приложением (Веб-сайт)



Плюсы:

- Простота в использовании
- Легкость в интерпретации результата
- Легкость интегрирования
- Не зависят от изменений структуры Веб-сайта
- Хороши для прототипирования и проверки гипотез

Минусы:

- Могут быть отключены в любой момент времени
- Могут быть расхождения с информацией на Веб странице
- Часто существуют лимиты на использование
- Набор данных ограничен (на странице данных часто больше)

Search the Largest API Directory on the Web

Search Over 22,485 APIs

SEARCH APIS

Filter APIs

By Category



☐ Include Deprecated APIs

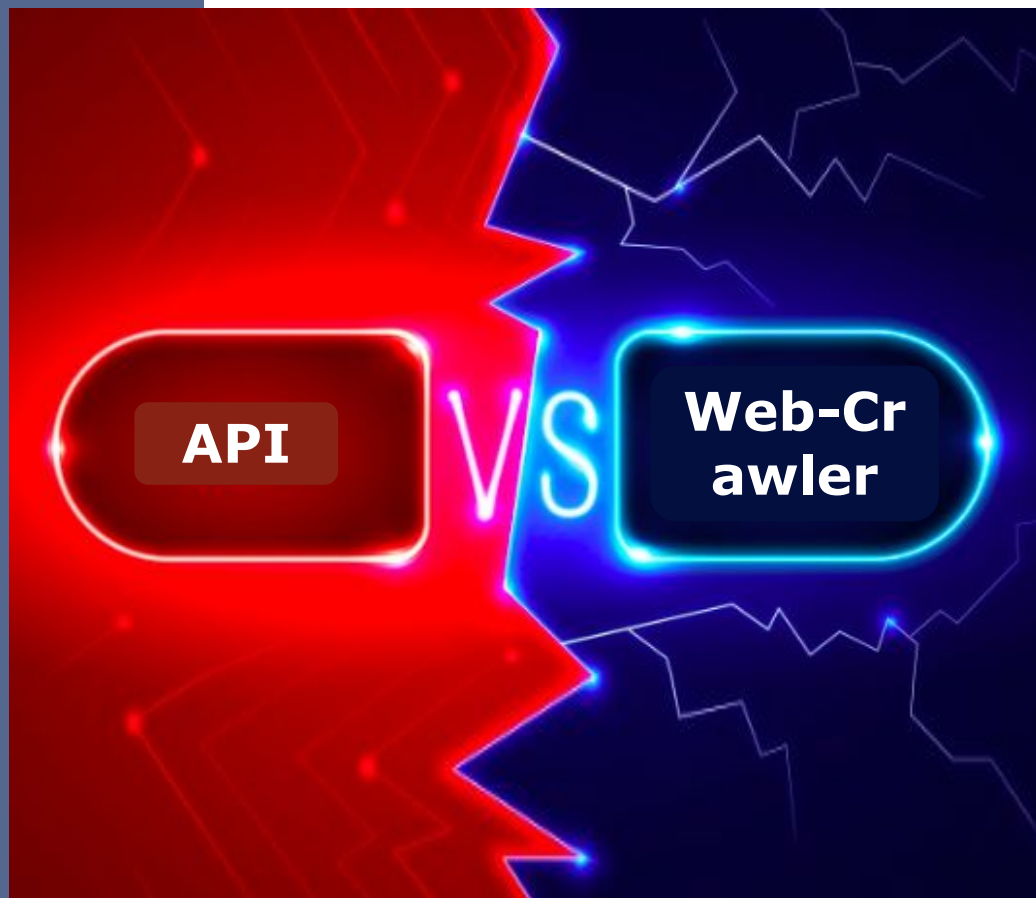
API Name	Description	Category	Versions
Google Maps API	[This API is no longer available. Google Maps' services have been split into multiple APIs, including the Static Maps API, Street View Image API, Directions APIs, Distance Matrix API, Elevation API,...]	Mapping	REST v0.0
Twitter API	[This API is no longer available. It has been split into multiple APIs, including the Twitter Ads API, Twitter Search Tweets API, and Twitter Direct Message API. This profile is maintained for...]	Social	Version ▾
YouTube API	The Data API allows users to integrate their program with YouTube and allow it to perform many of the operations available on the website. It provides the capability to search for videos, retrieve...	Video	REST v0.0

Web APIs

Ресурс содержит
более 16.000 APIs

<https://www.programmableweb.com/category/all/apis>

Web-Crawling vs API



Web scraping более надежен.

Доступен 24/7, и не требует поддержки со стороны кого-то еще кроме своего разработчика.

Web scraping как правило предоставляет более достоверные данные.

Web scraping не имеет ограничений на выдачу информации.

Большинство APIs имеют ограничения на выдачу

Web scraping часто выгружает более структурированную информацию.

Хотя API разработаны для предоставления структурированной информации, часто их схемы плохо поддерживаются

REQUESTS

Beautifulsoup

```
import requests
user_id = 12345
url = 'http://www.kinopoisk.ru/user/%d/votes/list/ord/date/page/2/#list' % (user_id) #
url для второй страницы
r = requests.get(url)
with open('test.html', 'w') as output_file:
    output_file.write(r.text.encode('cp1251'))
```

```
from bs4 import BeautifulSoup
from lxml import html
```

BeautifulSoup

```
soup = BeautifulSoup(text)
film_list = soup.find('div', {'class': 'profileFilmsList'})
```

lxml

```
tree = html.fromstring(text)
film_list_lxml = tree.xpath('//div[@class = "profileFilmsList"]')[0]
```

КиноПоиск
найди своё кино!

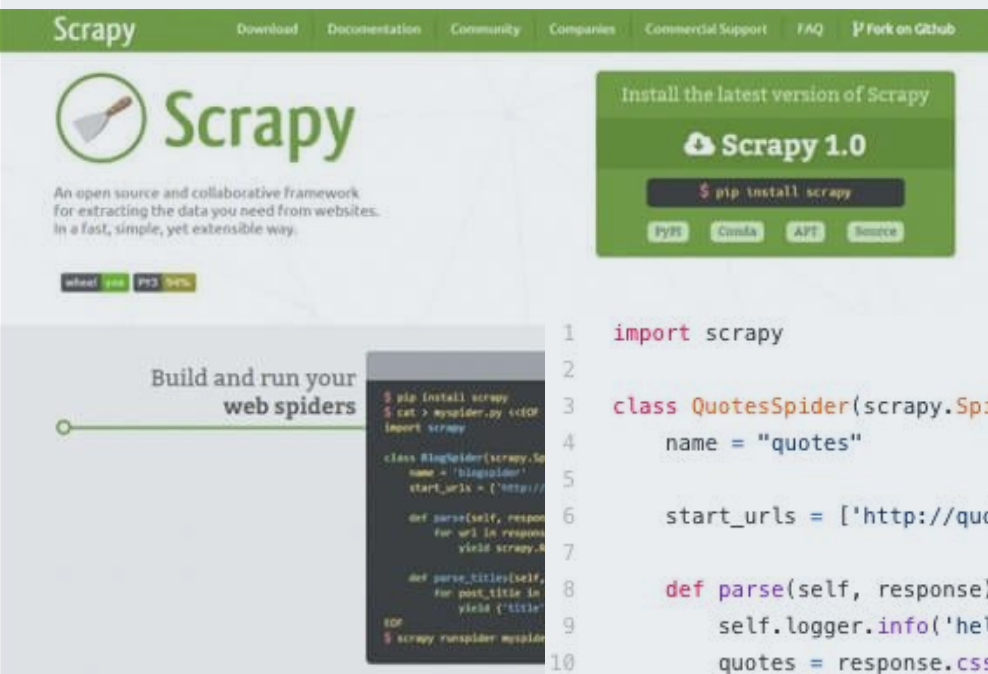
Если вы видите эту страницу, значит с вашего IP-адреса поступило необычно много запросов. Система защиты от роботов (C3oP) решила, что вместо вас действует программа, и ограничила доступ.

Для автоматического получения рейтинга, пожалуйста используйте [xml](#) версию.

6	Звёздные войны: Эпизод 3 – Месть Ситхов (2005)	04.03.2016, 16:27	8.070 (137 139) 140 мин.
7	Криминальное чтиво (1994)	04.03.2016, 16:25	8.621 (242 135) 154 мин.
8	Зверополис (2016)	04.03.2016, 10:53	8.646 (46 488) 108 мин.
9	Стажёр (2015)	28.02.2016, 21:38	7.555 (66 424) 121 мин.
10	Статус: Свободен (2015)	28.02.2016, 21:33	5.862 (12 143) 95 мин.
11	Дэдпул (2016)	23.02.2016, 20:32	6.666 (12 143) 95 мин.

```
IL > CSS Сценарий DOM Сеть Cookies
profileFilmsList < div < tr < tbody < table#lis...oryVotes < td.graphs < tr < tbody < table < td < tr < tbody < table < td < tr
  <div class="item">
  <div class="item even">
  <div class="item">
  <div class="item even">
    <div class="num">8</div>
    <div class="info">
      <div class="nameRus">
        <a href="/film/775276/" data-popup-info="enabled">Зверополис (2016)</a>
      </div>
      <div class="nameEng">Zootopia</div>
      <div class="rating">
        <b>8.646</b>
        <span class="text-grey">(46 488)</span>
        <span class="text-grey">108 мин.</span>
      </div>
    </div>
    <div class="date">04.03.2016, 10:53</div>
    <div class="selects vote_widget">
    <div class="clear"></div>
    <script>
  </div>
  <div class="item">
  <div class="item even">
```


Scrapy



```
1 import scrapy
2
3 class QuotesSpider(scrapy.Spider):
4     name = "quotes"
5
6     start_urls = ['http://quotes.toscrape.com']
7
8     def parse(self, response):
9         self.logger.info('hello this is my first spider')
10        quotes = response.css('div.quote')
11        for quote in quotes:
12            yield {
13                'text': quote.css('.text::text').get(),
14                'author': quote.css('.author::text').get(),
15                'tags': quote.css('.tag::text').getall(),
16            }
17
18        next_page = response.css('li.next a::attr(href)').get()
19
20        if next_page is not None:
21            next_page = response.urljoin(next_page)
22            yield scrapy.Request(next_page, callback=self.parse)
```

Scrapy – Open Source платформа для совместной работы для извлечения данных с веб-сайтов с очень высокой производительностью.

Особенности:

- Имеет встроенную поддержку для извлечения данных из источников HTML с использованием XPath и CSS.
- Кроссплатформенность.
- Легко расширяемая платформа
- Быстрая.
- Потребляет намного меньше памяти и ресурсов процессора.
- Имеет хорошую поддержку за счет большого сообщества разработчиков.

APIs

API (Application programming interface) — контракт, который предоставляет программа. «Ко мне можно обращаться так и эдак, я обязуюсь делать то и это».

REST (Representational State Transfer) - согласованный набор архитектурных принципов для создания масштабируемой и гибкой сети, описывающий компоненты системы и их взаимодействия)

RESTful —сеть, которая отвечает ограничениям Филдинга:

- Определение ресурсов
- Управление ресурсами через представления
- Гипермедиа
- Код по требованию
- Система слоёв
- Кэширование