

Web Scraping with Scrapy

Akshay Krishna

S5 CSE A

October 18, 2019

Objective

- **Scrap information from webpages of a given website using Scrapy**

What is Web scraping ? What is it's relevance ?

Web scraping refers to the process of collecting and organizing data from a web page. Today various methods are used to scrape data from web pages. The process of web scraping mainly involves two steps.Fetching and Extracting. First the agent requests the web page. When the web page is fetched as acknowledgement for the request,in the next step we extracts the required information from the fetched web page.

Web scraping nowadays used for purposes like, contact scraping,and as a component of webapplications used for web indexing,web mining and data mining,price comparison,weather data monitoring,researching etc.

Existing methods

1. Human Copy-Paste

A human being manually examines the webpage and copy paste required data. Failure rate is almost zero in this traditional technique

2. Text pattern matching

An approach to extract information from webpage using UNIX **grep** command or regular expression matching facilities of programming languages

3.Using web-scraping frameworws/softwares

These frameworks/softwares gives the facilities to create web scraping bots(spiders/crawlers) to scrape data from webpages. These spiders try to recognize structure of the page automaticaly and extracts the data and gives the facility to process and store them in specific format

4.Computer vision webpage analysis

Employing the features of machine learning and computer vision to idenetify and extract data from webpage. This can be described as computerised version of human copy-paste with a failure rate $\neq 0$

Proposed Methodology

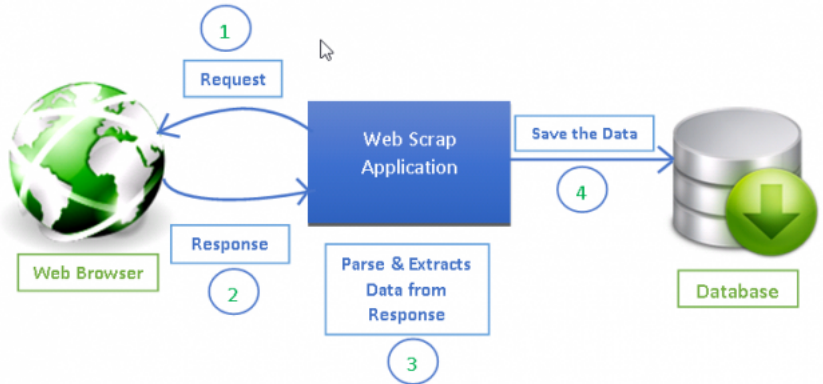
Using web-scraping framework

The proposed methodology is to scrape webpages using a software/framework. Here we are using "Scrapy", a framework written in python to generate and deploy spiders to extract data from a webpage

Why Scrapy ?

- 1 Faster and automated way to scrape webpages
- 2 Easy to use & implement
- 3 Have selectors to specify data field needs to be extracted
- 4 Have pipelines to refine extracted data

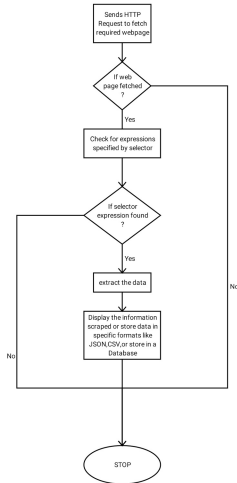
How Scrapy works



Barriers to Web Scraping

- Not all web pages allows data scraping
- Captchas block the scraper from proceeding further
- Frequent structural changes made to website
- IP blocking

Implementation Algorithm



Conclusion

Webpages of given Website is scraped and extracted data displayed in a specific format