Web Scraping with Scrapy

Akshay Krishna

S5 CSE A

October 23, 2019

Objective

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

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

Web scraping refers to the process of collecting and organizing data from a web page.The steps involved in web scraping are Fetching web pages and Extracting data from it

Applications of web scraping

- For contact scraping
- For web indexing
- For data mining
- For comparing and doing researches on data

Existing methods

1.Human Copy-Paste

A human being manuals 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 framewors/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 ¿0

Proposed Methodology

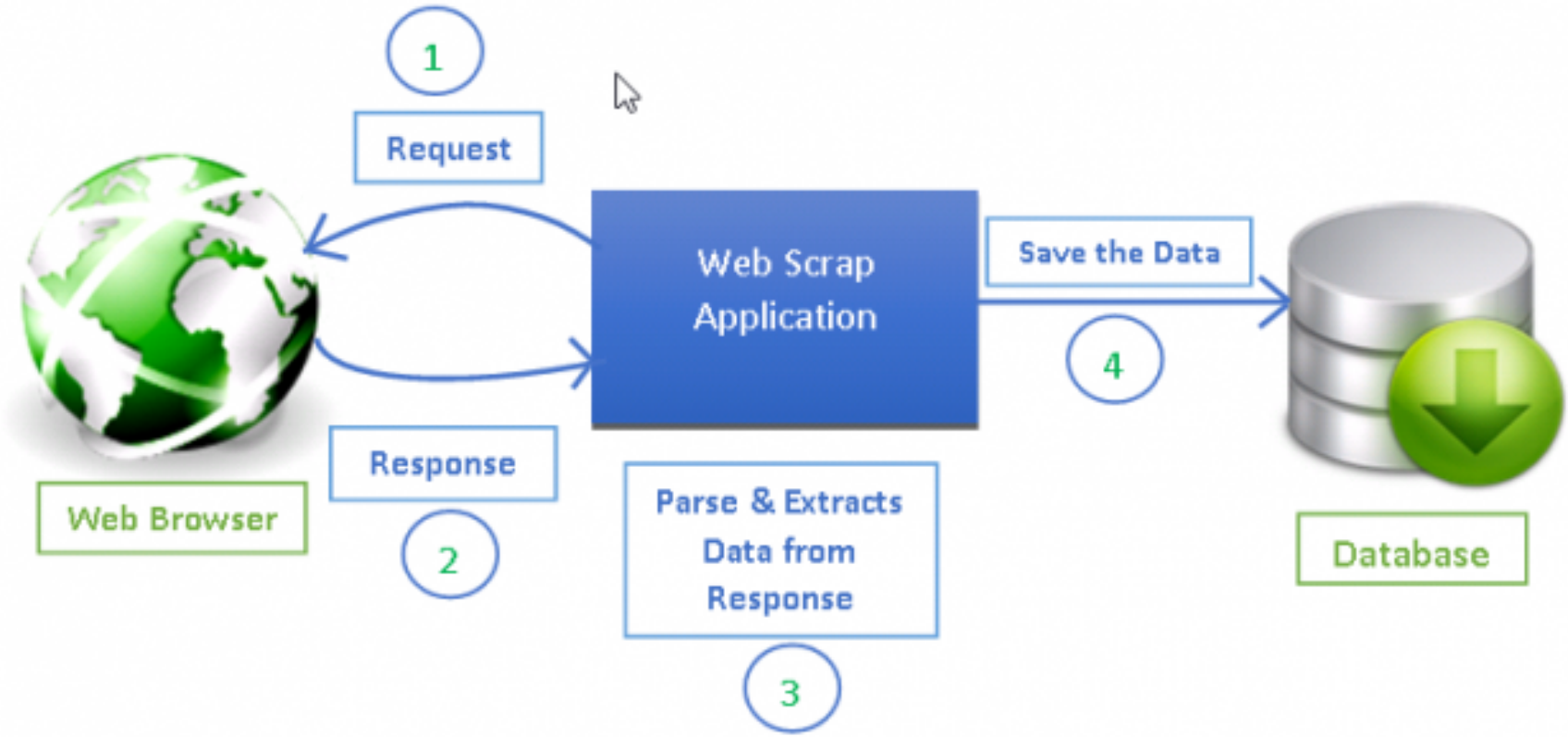
Using web-scraping framework

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

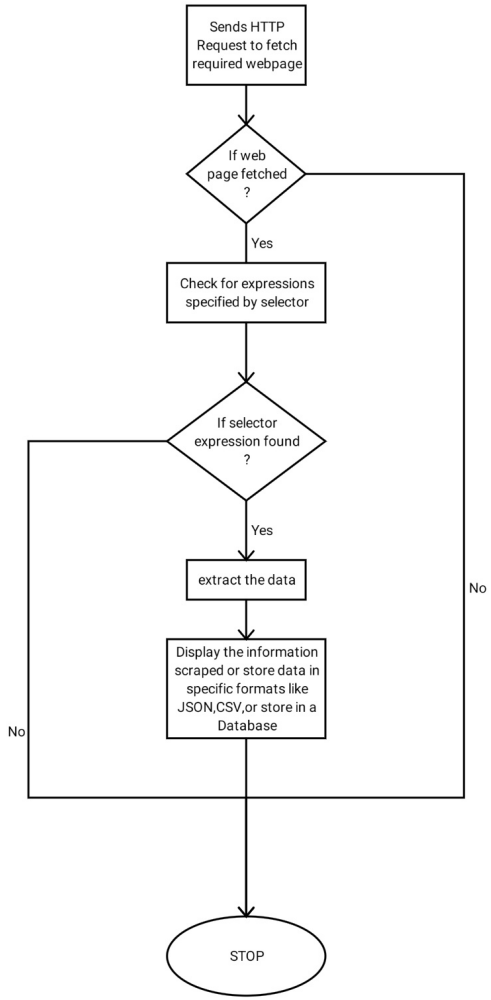
How Scrapy works



Barriers to Web Scraping

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

Implementation Algorithm



Conclusion

- Scrapy is a package of python used to scrape web pages
- It contains modules for processing and extracting data from webpages
- Using selectors we can select the fields of web pages we want
- using **extract()** we can extract the required data from web page
- We can display the data scraped of we can store it in files or in a database