PRIME MINISTER'S DEPARTMENT
DEPARTMENT OF STATISTICS MALAYSIA

Web Scraping of E-Commerce Website Using BeautifulSoup in Python

20 OCT

2016 - 2030

PSSN

StatsMalaysia
www.DOSM.gov.my

Outlines

- Introduction
- Web crawling vs Web Scraping
- Challenging in Web Scraping
- Components of web scraper
- Comparison packages in web scraping
- Methodology
- Step by step of web scraping of e-commerce website

Introduction

- Web scraping is an automatic process of extracting information from website.
- Traditionally : data provided by agencies, organization, online data platform or questionnaire
- Why web scraping (the data from web can be used for)
  - E-commerce :Competitor price monitoring
  - Data Science : Academic purpose, learn data cleaning, forecasting etc.
  - Monitoring of trends of current issues
- Web scrapping also known as "web data mining" or web Scraping in extracting data

https://www.youtube.com/watch?v=Pm1P5hvsc-k

WEB CRAWLING VS WEB SCRAPING

Web crawling, also known as Indexing, is used to index the information on the page using bots also known as crawlers. Crawling is essentially what search engines do. It's all about viewing a page as a whole and indexing it. When a bot crawls a website, it goes through every page and every link, until the last line of the website, looking for ANY information.

The web crawling process usually captures generic information. whereas web scraping hones in on specific data set snippets.

Web Crawling	Web Scraping
Refer to downloading and string the contents of large number of websites	Refers to extracting individual data element from the website by using a site specific structure.
Mostly done in large scales	Can be implemented at any scale
Yields most of generic information	Yields specific information
Used by major search search engines line Google, Yahoo,	The information extracted using web scraping can be used to replication in some other website or can be used to perform data analysis. For examples, the data elements can be names, address , price etc.

Challenging in Web Scraping

- Most of website developers are trying to avoid from their website being scrapped.
- Some the scrapper will damage the website if it does not do correctly.

Variety : every website is different. Each website is unique and will need personal treatment if we want to extract the relevant information.

Durability : Websites constantly change.

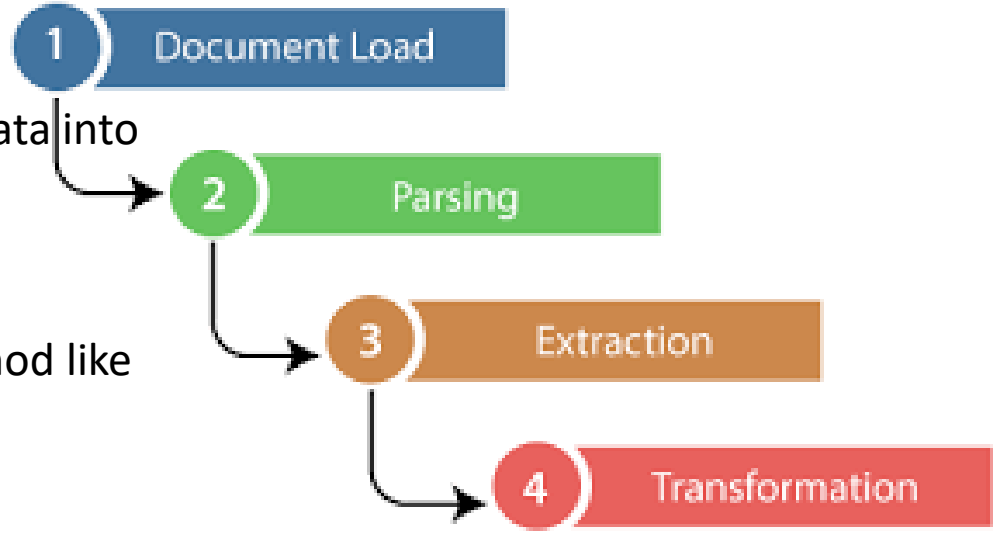
Components of a Web Scraper

Web Crawler Module
This module is used to navigate the target website by making HTTP or HTTPs request to the website.

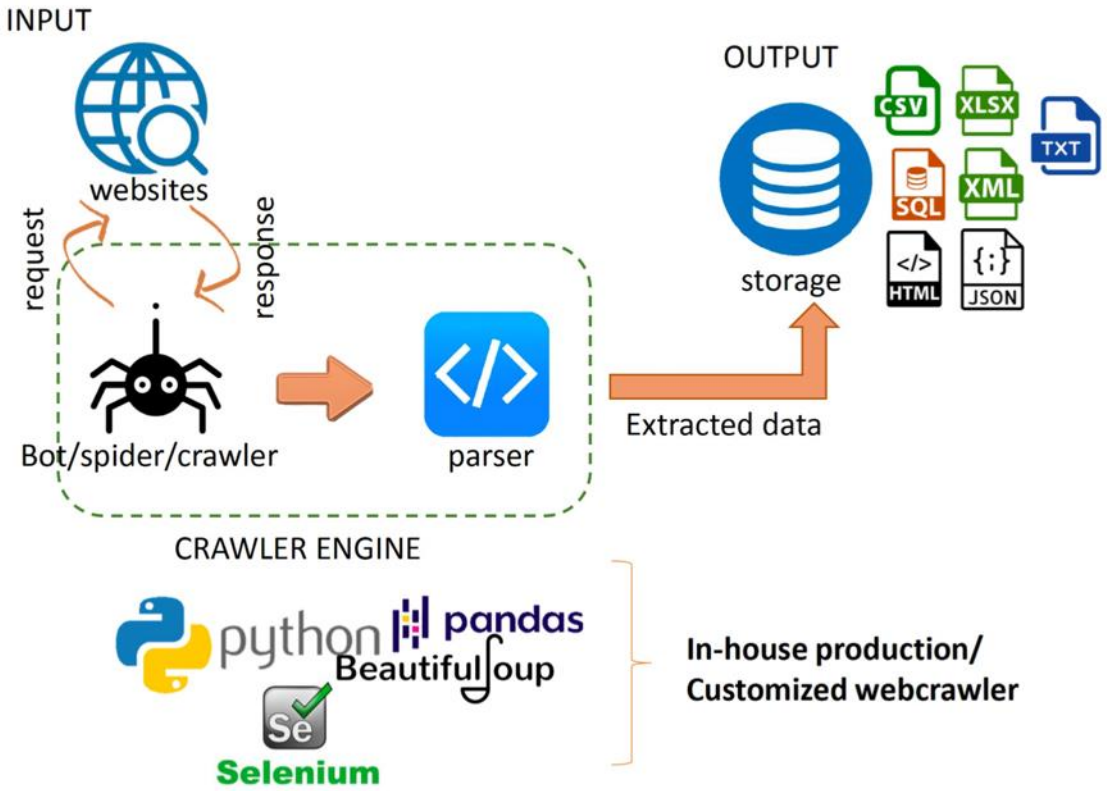
Extractor
The extractor processes the fetched HTML content and extracts the data into semi structured format.

Data Transformation and cleaning module
The data will do some cleaning module so that can be used. The method like String manipulation or regular expression can be used

Storage Module
We need to store it as per our requirement



Web Scraping



Webcrawling Tools

botsol
Web Scraper
Octoparse
parsehub

Pros

- Simple and easy to use (coding free)
- Point and click system

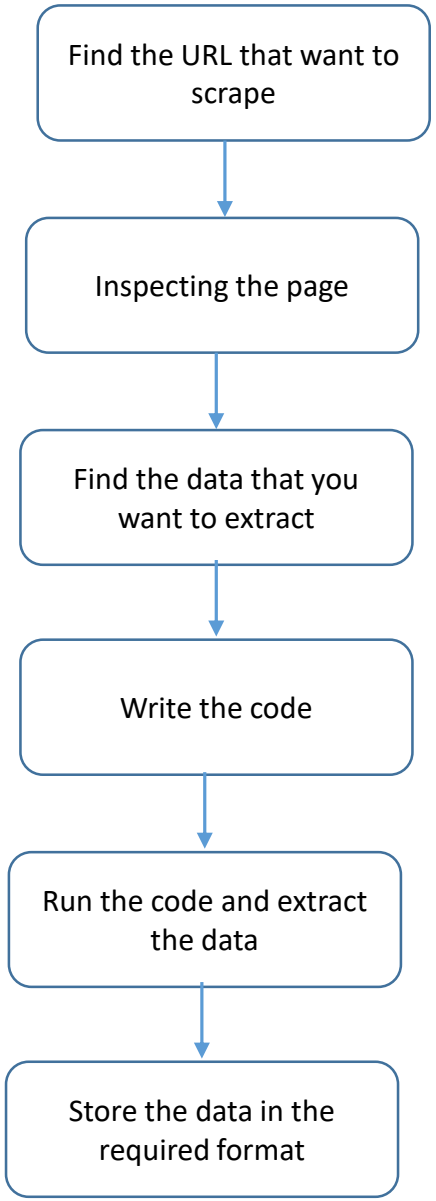
Cons

- Paid subscription for full version (free version – function limited)
- Dependencies on bulit in scraper

Comparison Packages in Web Scraping

Packages	Requests	Beautiful Soup	lxml	Selenium
Purpose	Simplify making HTTP requests	Parsing	Parsing	Simplify making HTTP requests
Ease-of-use	High	High	Medium	Medium
Speed	Fast	Fast	Very Fast	Slow
Learning Curve	Very easy(beginner friendly)	Very easy(beginner friendly)	Easy	Easy
Documentation	Excellent	Excellent	Good	Good
Javascript Support	None	None	None	Yes
CPU and Memory usage	Low	Low	Low	High
Size of web scraping project supported	Large and small	Large and small	Large and small	Small

Methodology



Hands-on

E-commerce website : www.bookdepository.com
: category of book : Big Data

Data to scrape :

- Title
- Author
- Date Published
- Cover type
- Current price
- Previous/initial price

Platform : Jupyter Notebook/Google Colab
Installation of Jupyter : https://coreteambda.wixsite.com/blog/python

Data format : CSV

Lets Do it

"STATISTICS BLOOM IN HARMONY"

Doesn't matter far or near
Strength in numbers
we don't live in fear

Birds of feather flock together
Statistics our form of adour
We, will always live it up

So let us live in solidarity
And in the world arena we'll succeed
It is statistics that will come to be
The reason we will bloom in harmony

Everybody undivided
Data's where our hearts reside in
There will always be a bind

Just like fire that ignites
That's how brightly lit our dreams are
We'll reach higher than the stars

Sending love to one another
Leaving no one in a slumber
We will stand with unity

Mustering our courage while
Embracing our disparities
We'll achieve our victory

One dream with unity
One love with harmony

" STATISTICS BLOOM IN HARMONY" VIDEO

https://bit.ly/StatisticsBloomInHarmony

THANK YOU

#KELUARGAMALAYSIA

StatsMalaysia

www.DOSM.gov.my

2012-2022 | 2022 | 20 OCT | 2016 - 2030 | PSSN