Fastcampus Data Science SCHOOL

Python Basic

Python External Library

PIP("Pip Installs Packages" or "Pip Installs Python")

- a package management system used to install and manage software packages written in Python

command

To install or uninstall package,

```
$ pip install package-name
```

```
$ pip uninstall package-name
```

install packages listed in requirements.txt

```
$ pip install -r requirements.txt
```

print the list of installed packages

```
$ pip freeze
```

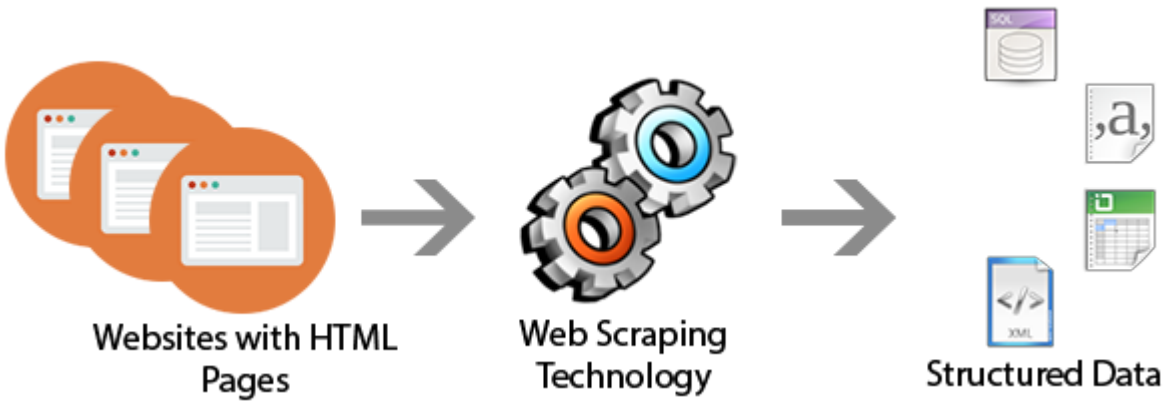
Generate a requirements file

```
$ pip freeze > requirements.txt
```

Web Crawling with python

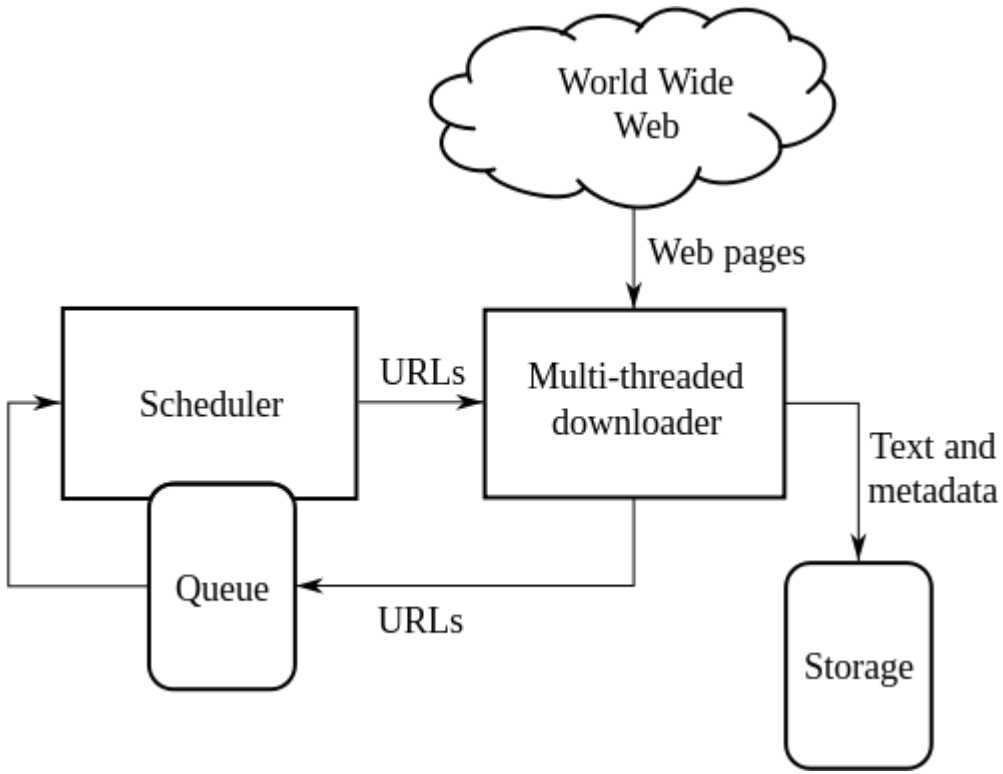
Scraping vs Crawling vs Parsing

Scraping: 데이터를 수집하는 행위



Scraping vs Crawling vs Parsing

Crawling: 조직적 자동화된 방법으로 월드 와이드 웹을 탐색하는 것



Scraping vs Crawling vs Parsing

Parsing: 문장 혹은 문서를 구성 성분으로 분해하고 위계관계를 분석하여 문장의 구조를 결정하는 것

Caution!!

저작권 침해 위반 소지

- 웹사이트 운영자의 크롤링 금지 룰을 어길경우
- 월권하여 데이터베이스에 접근
- 타인의 경제적 이익을 침해할 경우
- 개인정보를 수집할 경우(전화번호, 주소, ..)

Beautiful Soup

Web Scraping with Beautiful Soup

```
$ pip install beautifulsoup4
```

```
$ pip install requests
```

```
$ pip list
```

Web Scraping with Beautiful Soup

```
from bs4 import BeautifulSoup
html = """

uglified html code

"""
soup = BeautifulSoup(html, "html.parser")
print(soup.prettify())
```

Web Scraping with Beautiful Soup

```
curl https://www.rottentomatoes.com
```

```
import requests
from bs4 import BeautifulSoup


url = "https://www.rottentomatoes.com"
html = requests.get(url)
source = html.text

soup = BeautifulSoup(source, "html.parser")
print(soup)

table = soup.find(id="Top-Box-Office")
print(table)
```

Web Scraping with Beautiful Soup

```
all_tr = table.find_all("tr")

for tr in all_tr:
    all_td = tr.find_all("td")
    score = all_td[0].find("span", attrs={"class":"tMeterScore"
    movie_name = all_td[1].a.text
    amount = all_td[2].a.text
    print(score, movie_name, amount)
```

Let's Crawl naver search results