Crawling Day 2 230228

Homework Review

```
from selenium import webdriver        # 웹 브라우저 컨트롤 (크롬)
from bs4 import BeautifulSoup as bs   # 데이터 분석을 용이하게 정제
import pandas as pd                   # 데이터 분석 관련 모듈

driver = webdriver.Chrome('chromedriver.exe')  # 버전 주의
url = 'https://www.genie.co.kr/chart/top200'
driver.get(url)

txt = driver.page_source      # 이때 읽어온 데이터는 그냥 글자
html = bs(txt)

songs = html.select('tbody > tr')

song_data = []
rank = 1

for song in songs:
    title = song.select('a.title')[0].text.strip()
    singer = song.select('a.artist')[0].text.strip()

    song_data.append(['Genie',rank,title,singer])
    rank += 1


url = 'https://www.genie.co.kr/chart/top200?ditc=D&ymd=20230228&hh=14&rtm=Y&pg=2'
driver.get(url)

txt = driver.page_source      # 이때 읽어온 데이터는 그냥 글자
html = bs(txt)

songs = html.select('tbody > tr')

for song in songs:
    title = song.select('a.title')[0].text.strip()
    singer = song.select('a.artist')[0].text.strip()

    song_data.append(['Genie',rank,title,singer])
    rank += 1

df = pd.DataFrame(song_data, columns = ['서비스','순위','타이틀','가수'])
df


df.to_excel('Genie.xlsx', index=False)
```

from pprint import pprint

출력을 pretty print

Naver Webtoon Title

```
from selenium import webdriver        # 웹 브라우저 컨트롤 (크롬)
from bs4 import BeautifulSoup as bs   # 데이터 분석을 용이하게 정제
import pandas as pd                   # 데이터 분석 관련 모듈

driver = webdriver.Chrome('chromedriver.exe')  # 버전 주의
url = 'https://comic.naver.com/webtoon/weekday'
driver.get(url)

txt = driver.page_source      # 이때 읽어온 데이터는 그냥 글자
html = bs(txt)

# titles = html.select('a.title')  select 문법

titles = html.findAll('a', {'class':'title'})
titles = [title.text for title in titles]
pprint(titles)
```

네이버 날씨

```
import requests

txt = requests.get('https://weather.naver.com/')
html = bs(txt.text)

temp = html.select('strong.current')[0].text.strip()
temp[5:]
```

Bitcoin Price

```
from selenium import webdriver        # 웹 브라우저 컨트롤 (크롬)
from bs4 import BeautifulSoup as bs   # 데이터 분석을 용이하게 정제
import pandas as pd                   # 데이터 분석 관련 모듈
from pprint import pprint

driver = webdriver.Chrome('chromedriver.exe')  # 버전 주의
url = 'https://www.bithumb.com/react/'
driver.get(url)

txt = driver.page_source      # 이때 읽어온 데이터는 그냥 글자
html = bs(txt)


coins = html.findAll('strong', {'class':'MarketRow_sort-real__5zeND'})
coin = coins[0]

coins = [coin.text for coin in coins]

coins[1]
```

Naver Webtoon Thumbnail

```
###################################
# 썸 네일 모두 자동으로 다운받기
###################################

from urllib.request import urlretrieve  # 파일 다운 모듈
import re
import requests
from bs4 import BeautifulSoup as bs   # 데이터 분석을 용이하게 정제
import pandas as pd                   # 데이터 분석 관련 모듈


url = 'https://comic.naver.com/webtoon/weekday'
txt = requests.get(url)
html = bs(txt.text)

tlist= html.findAll('div',{'class':'col_inner'})
li_list = []

for data in tlist:
    #제목과 썸네일 영역 추출
    li_list.extend(data.findAll('li'))  # 해당 부분을 찾아 병합
```

```
# pprint(li_list)

# 썸네일과 제목 추출
for li in li_list:
    img = li.find('img')
    title = img['title']
    img_src = img['src']
    # print(title, img_src)
    # 특수 문자 제거 정규 표현식
    title = re.sub('[^0-9a-zA-Zㄱ-힣]', '',title)
    urlretrieve(img_src, title+'.jpg')
```