Lab. Using Beautiful Soup

1. BeautifulSoup module 이용하기
   1)pip install BeautifulSoup4
   2)Scraping을 위한 module
   3)루이스 캐럴의 [이상한 나라의 앨리스]에 나오는 동명의 시에서 이름을 따왔다.
   4)일반적으로 HTML tag들이 start tag와 end tag가 서로 pair 되지 않을 경우가 많다.
   5)pair 되지 않아도 아름답게 처리해 주는 module
   6)https://www.crummy.com/software/BeautifulSoup/bs4/doc/

```
from bs4 import BeautifulSoup
import requests

html_data = requests.get('https://www.naver.com')
soup = BeautifulSoup(html_data.text, 'html.parser')
soup.title
---------------------------------
<title>NAVER</title>
```

   7)첫번째 인자는 Python 객체로 바꿀 string을 넣어주고 두 번째 인자로 parser를 넣는다.
   8)Parser란, 원시 코드인 순수 문자열 객체를 해석할 수 있도록 분석하는 것을 의미한다.
   9)Python에서 사용되는 parser는 다음과 같다.
      a. lxml
      b. html5lib
      c. html.parser

   10)각 parser의 장단점을 알아보자.
      a. lxml
         -XML 해석이 가능한 parser이다.
         -Python 2.x와 3.x 모두 지원 가능하다.
         -다른 parser에 비해 매우 빠른 속도로 처리한다.
         -그 이유는 C언어로 구현되어 있기 때문이다.
      b. html5lib
         -웹 브라우저 방식으로 HTML을 해석한다.
         -하지만 처리 속도가 매우 느리다는 단점이 있다.
         -그리고 2.x 전용이다.
      c. html.parser
         -최신 버전의 Python에서는 사용이 불가

2. 각 parser 비교하기
   1)lxml

```
from bs4 import BeautifulSoup
html = """<p>test</p>"""
soup = BeautifulSoup(html, 'lxml')
print(soup)
-----------------------------------------------
<html><body><p>test</p></body></html>


from bs4 import BeautifulSoup
html = """<html><p>test</p></html>"""
soup = BeautifulSoup(html, 'lxml')
print(soup)

html = """<body><p>test</p></body>"""
soup = BeautifulSoup(html, 'lxml')
print(soup)
-------------------------------------------
<html><body><p>test</p></body></html>
<html><body><p>test</p></body></html>
```

      -html과 body tag가 포함된 형태로 만들어 준다.

   2)html5lib

```
$ pip install html5lib

from bs4 import BeautifulSoup
html = """<p>test</p>"""
soup = BeautifulSoup(html, 'html5lib')
print(soup)
-------------------------------------------------
<html><head></head><body><p>test</p></body></html>
```

      -html5lib도 html처럼 해석하기 때문에 html, head, body tag가 포함된 형태로 만들어 준다.

3. 한빛미디어 책 제목 읽어오기

```
hanbit = requests.get('http://www.hanbit.co.kr/media/')
soup = BeautifulSoup(hanbit.text, 'html.parser')
soup
```

```
    ------------------------------------------------------------
    <!DOCTYPE html>
    <html lang="ko">
    <head>
    <!--[if lte IE 8]>
    <script>
      location.replace('/support/explorer_upgrade.html');
    </script>
    <![endif]-->
    <!-- Google Tag Manager -->
    <script>(function(w,d,s,l,i){w[l]=w[l]||[];w[l].push({'gtm.start':
    new Date().getTime(),event:'gtm.js'});var f=d.getElementsByTagName(s)[0],
    j=d.createElement(s),dl=l!='dataLayer'?'&l='+l:'';j.async=true;j.src=
    'https://www.googletagmanager.com/gtm.js?id='+i+dl;f.parentNode.insertBefore(j,f);
    })(window,document,'script','dataLayer','GTM-W9D5PM3');</script>
    <!-- End Google Tag Manager -->
    ...
    ...

    for book in soup.find_all('p', class_='book_tit'):
        print(book.find('a').text)
    ------------------------------------------------------------
    마이크로서비스 아키텍처 구축 가이드
    맛있는 디자인 프리미어 프로&애프터 이펙트 CC 2023
    맛있는 디자인 포토샵&일러스트레이터 CC 2023
    실전에서 바로 쓰는 Next.js
    (NO.1 영상 편집 유튜버) 비됴클래스의 유튜브 영상 편집 with 프리미어 프로
    이것이 Windows Server다(개정판)
    업무에 바로 쓰는 AWS 입문
    성공하는 유튜브 채널은 따로 있다
    프로덕트 매니저는 무슨 일을 하고 있을까
    우리가 사랑한 한국 PC 게임
    혼자 공부하는파이썬(개정판)
    혼자 공부하는 머신러닝+딥러닝
    회사에서 바로 통하는 실무 엑셀+파워포인트+워드&한글(모든 버전 사용 가능, 개정판)
    혼자 공부하는 컴퓨터 구조+운영체제
    혼자 공부하는 C 언어
    맛있는 디자인 포토샵&일러스트레이터 CC 2022
    혼자 해도 프로 작가처럼 잘 그리는 아이패드 드로잉 with 프로크리에이트
    구글 엔지니어는 이렇게 일한다
    국내 최다 이모티콘 승인 작가 씨엠제이가 알려주는 승인율 99.9% 이모티콘 만들기
    이것이 취업을 위한 코딩 테스트다 with 파이썬


4. Naver 영화 평점 Scraping 하기

    from bs4 import BeautifulSoup

    html_data = requests.get('https://movie.naver.com/movie/point/af/list.nhn?page=1')
    soup = BeautifulSoup(html_data.text, 'html.parser')
    titles = soup.find_all(class_='movie')

    title_list = []
    for title in titles:
        print(title.text)
    ---------------------------------------------
    메간
    애프터썬
    터미네이터
    열한번째 엄마
    접속
    영웅
    바빌론
    가재가 노래하는 곳
    더 퍼스트 슬램덩크
    더 퍼스트 슬램덩크

    title_list = []
    for title in titles:
        title_list.append(title.text)

    point_list = []
    points = soup.find_all(class_='list_netizen_score')
    for point in points:
        point_list.append(point.find('em').text)

    review_list = []
    reviews = soup.find_all(class_='title')
    for review in reviews:
         rev = review.text
         rev = rev.strip()
         rev = rev.replace('\t', '')
         rev = rev.replace('\n', '')
```

```
        rev = rev.replace('신고', '')
        review_list.append(rev)

    df = pd.DataFrame(title_list, columns=['Title'])
    df['Point'] = point_list
    df['Review'] = review_list
    df
    ------------------------------------------------
       Title                                Point      Review
    0  자백    9  자백별점 - 총 10점 중9기대이상입니다. 반전이 일단 최고네요. 간만에 긴장감있게...
    1  타이타닉   10 타이타닉별점 - 총 10점 중10여운의 끝을 보여주는 영화….
    2  자백    7  자백별점 - 총 10점 중7생각보다 재밌게 봤습니다. 기대 안했는데 영상미가 돋보였...
    3  메간    8  메간별점 - 총 10점 중8근래 본 공포영화 중에선 가장 볼만 했습니다.
    4  누구의 딸도 아닌 해원10 누구의 딸도 아닌 해원별점 - 총 10점 중10찌질하다는 것은 그릇이 작은 것이다....
    5  압꾸정 8  압꾸정별점 - 총 10점 중8다들 재미없다고 하시는데.. 저는 그냥 아무생각없이 웃...
    6  영웅    10 영웅별점 - 총 10점 중10이거 극장가서 안보면 완전 후회합니다 초강추요영상미 넘...
    7  더 퍼스트 슬램덩크    10 더 퍼스트 슬램덩크별점 - 총 10점 중10감동 그 자체입니다..
    8  타이타닉   10 타이타닉별점 - 총 10점 중10ost 흘러 나올 때마다 울었다..
    9  타이타닉   10 타이타닉별점 - 총 10점 중10영화관에서는 처음 보네요. 재밌고 감동적이네요
```

5. Naver 평점 1page부터 100page까지 scraping 하기

```
    from bs4 import BeautifulSoup
    import requests
    import pandas as pd

    url = 'https://movie.naver.com/movie/point/af/list.nhn?page='

    title_list = []
    point_list = []
    review_list = []

    for pge in range(1, 101):
        url = url + str(pge)
        print(url)
        html_data = requests.get(url)
        soup = BeautifulSoup(html_data.text, 'html.parser')
        titles = soup.find_all(class_='movie')
        points = soup.find_all(class_='list_netizen_score')
        reviews = soup.find_all(class_='title')
        for title in titles:
            title_list.append(title.text)
        for point in points:
            point_list.append(point.text)
        for review in reviews:
            rev = review.text
            rev = rev.strip()
            rev = rev.replace('\t', '')
            rev = rev.replace('\n', '')
            rev = rev.replace('신고', '')
            review_list.append(rev)
        url = url.split('=')[0] + '='

    df = pd.DataFrame(title_list, columns=['Title'])
    df['Point'] = point_list
    df['Review'] = review_list

    df.info()
    ------------------------------------------------------
    https://movie.naver.com/movie/point/af/list.nhn?page=1
    https://movie.naver.com/movie/point/af/list.nhn?page=2
    https://movie.naver.com/movie/point/af/list.nhn?page=3
    https://movie.naver.com/movie/point/af/list.nhn?page=4
    ...
    ...
    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 1000 entries, 0 to 999
    Data columns (total 3 columns):
     #   Column  Non-Null Count  Dtype
    ---  ------  --------------  -----
     0   Title   1000 non-null   object
     1   Point   1000 non-null   object
     2   Review  1000 non-null   object
    dtypes: object(3)
    memory usage: 23.6+ KB
```

6. Coupang의 상품정보 Scraping

    1)Web 문서들은 서로 다양한 문서 구조로 출력된다.

    2)따라서 Python에서 web 문서로부터 scraping을 하기 위해서는 추출하고자 하는 정보들이 구성되어 있는 영역을 먼저 확인해야 한다.

    3)Social Commerce의 대표적인 online market인 Coupang의 상품 정보 추출을 해보자.

4)'여성패션' 중 '여성 크로스백' 목록 item을 살펴보자.
5)Scraping하려는 web page의 URL 구조와 문서 구조를 파악한다.
6)URL 구조
http://www.coupang.com/np/search?q=여성크로스백
7)문서 구조
-상품명 : class="title"
-가격 : class="discount-price"

7. 한국일보 headline 기사 Scraping하기
1)한국일보 첫 page의 기사를 Scraping 해보자.
2)먼저 scraping 하려는 web page의 URL 구조와 문서 구조를 파악해야 한다.
3)URL 구조
http://www.hankookilbo.com/
4)문서 구조
-기사 제목 : class="title"