웹스크래핑

yes24 홈페이지 구조 변경

실험 데이터세트

- **YES24 도서구매 사이트**
  - YES24 베스트셀러 부문
  - url : http://www.yes24.com/24/Category/BestSeller

- **실험 내용**
  - 베스트셀러 웹 페이지 분석
  - 베스트셀러 도서 정보를 수집
  - 베스트셀러의 도서 제목, 저자, 가격을 표로 작성

실습2.3 find_all() 활용

- **find_all() 로 파싱**
  - 'div' 태그의 속성 class="itemUnit" 모두 수집 : 24개 베스트셀러 도서

```
# total number of cards
cards_data = data.find_all('div', attrs={'class':'itemUnit'})
print('Total Number of Cards Found : ', len(cards_data))
```

```
Total Number of Cards Found :  24
```

```
<li data-goods-no="122120495">
    <div class="itemUnit">
        <div class="item_img">
            <div class="img_canvas">
```

태그로 정보 확인

- **책제목, 저자, 가격 태그로 정보 확인**
  - 책제목, 저자는 <a>태그, 가격은 태그에

```
<div class="info_row info_name">

        <span class="gd_res">[도서]</span>

    <a class="gd_name" href="/Product/Goods/122120495" onclick="wiseLogV2(&#39;BS&#39;, &#39;001_005_001&#39;, &#39;&#39;); ">마흔에 읽는 쇼펜하우어</a>

        <span class="gd_nameE">마음의 위기를 다스리는 철학 수업</span>
                    <a href="/Product/Goods/122120495" target="_blank" class="bgYUI ico_nWin" onclick="wiseLogV2(&#39;BS&#39;, &#39;001_005_001&#39;, &#3
</div>
    <div class="info_row info_pubGrp">

            <span class="authPub info_auth" onclick="wiseLogV2(&#39;BS&#39;, &#39;001_005_002&#39;, &#39;&#39;);">
                <a href="https://www.yes24.com/Product/Search?domain=ALL&query=강용수&authorNo=293665&author=강용수" target="">강용수</a> 저
            </span>
                                <span class="authPub info_pub" onclick="wiseLogV2(&#39;BS&#39;, &#39;001_005_003&#39;, &#39;&#39;);"><a href="https://www
                                <span class="authPub info_date">2023년 09월</span>
    </div>
                <dl class="info_present">
        <dt><em class="bgYUI ico_present">구매혜택</em></dt>
        <dd>
            [11월의 얼리리더] 헬로키티 문진 마그넷 증정 (포인트 차감)
        </dd>
    </dl>
                    <div class="info_row info_price">
                <span class="txt_sale"><em class="num">10</em>%</span>
            <strong class="txt_num"><em class="yes_b">15,300</em>원</strong>
            <span class="txt_num dash"><em class="yes_m">15,300</em>원</span>
                <span class="yPoint"><em class="bgYUI ico_point">포인트적립</em>850원</span>
        </div>
```

실습2.5 태그로 정보 수집

- 베스트셀러 정보 출력 (책제목, 저자, 가격 정보)

```
for card in cards_data:
    print(card.find('a', attrs={'class':'gd_name'}).text)
    tag = card.find('span', attrs={'class':'authPub info_auth'})
    print(tag.find('a').text)
    print(card.find('em', attrs={'class':'yes_b'}).text)
```

```
마흔에 읽는 쇼펜하우어
강용수
15,300
전지적 푸바오 시점
에버랜드 동물원
19,800
더 마인드
하와이 대저택
17,820
트렌드 코리아 2024
김난도
17,100
```

실습2.6 베스트셀러 표 만들기

- 베스트셀러 정보 딕셔너리로 리스트 만들기
  - 책제목, 저자, 가격 Label과 값으로 card_details 딕셔너리를 생성하고, scraped_data 리스트에 저장

```
scraped_data = []
for card in cards_data:
    card_details = {}
    card_details['book_name'] = card.find('a', attrs={'class':'gd_name'}).text
    tag = card.find('span', attrs={'class':'authPub info_auth'})
    card_details['author'] = tag.find('a').text
    card_details['price'] = card.find('em', attrs={'class':'yes_b'}).text
    # append the scraped data to the list
    scraped_data.append(card_details)
```

영상 수집

- **HTML 웹 페이지에서 영상 수집**
  - 사례: YES24 베스트셀러 책 표지 사진 수집
  - HTML 웹 페이지에서 <img > 태그의 src속성은 사진이 비어있어서, data-original 속성에서 가져오기

```
<em class="img_bdr">
    <img class="lazy" data-original="https://image.yes24.com/goods/122120495/L"
src="https://image.yes24.com/momo/Noimg_L.jpg" border="0" alt="마흔에 읽는 쇼펜하우어">
```

실습3.1 <img>태그로 수집

- **실습2.3에 이어서**
  - 첫번째 <img> 태그로 수집

```
# find all with the image tag
images=[]
for card in cards_data:
    images.append( card.find('img') )
print('Number of Images: ', len(images))
```

```
Number of Images:  24
```

실습3.2 src 속성 수집

- **베스트셀러 책표지 data-original 속성으로 수집**
  - 첫번째 <img> 태그의 data-original 속성 수집

```
image_src = [x['data-original'] for x in images]
for image in image_src:
    print(image)
```

```
https://image.yes24.com/goods/122120495/L
https://image.yes24.com/goods/123400249/L
https://image.yes24.com/goods/123155346/L
https://image.yes24.com/goods/122426425/L
https://image.yes24.com/goods/103495056/L
https://image.yes24.com/goods/122944685/L
https://image.yes24.com/goods/117014613/L
```

실습3.3 영상 저장

- **베스트셀러 책표지 수집하여 저장**
  - 영상 url을 요청하여 write()로 영상 저장

```
image_count = 1
for image in image_src:
    with open('image_'+str(image_count)+'.jpeg', 'wb') as f:
        res = requests.get(image)
        f.write(res.content)
    image_count = image_count+1
```

  - 수집하여 저장된 영상


image_1.jpeg


image_2.jpeg

image_3.jpeg


image_4.jpeg

End