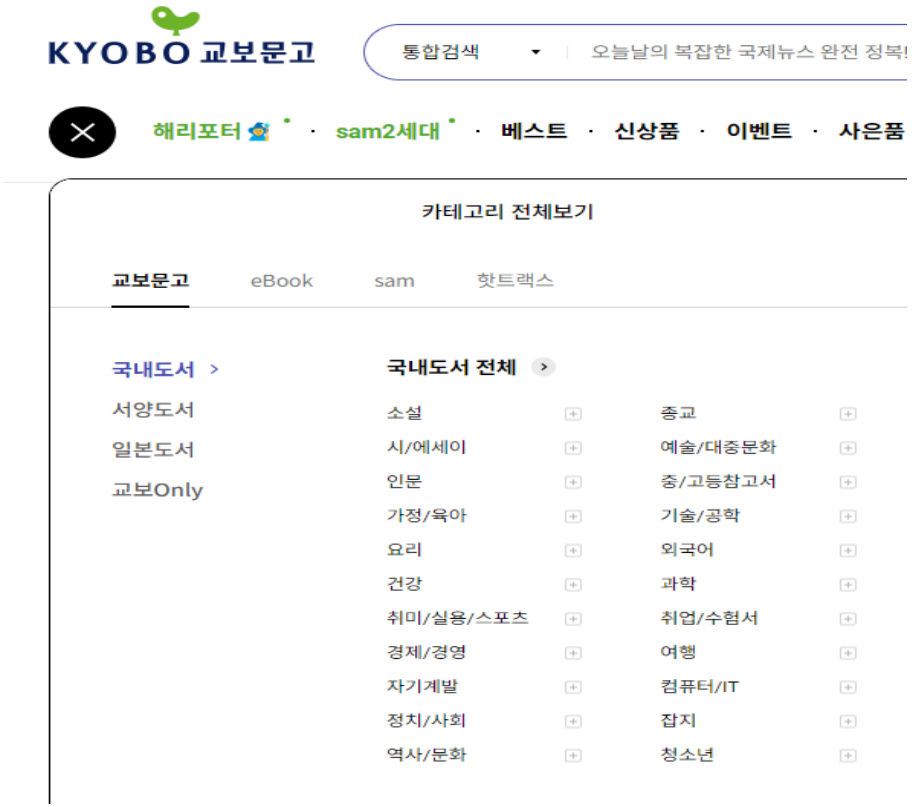
04. 프로젝트 수행 경과 – 데이터 수집

- 크롤링 : 웹페이지를 그대로 가져와서 거기서 데이터를 추출하는 방식




```
[ ]  1 from selenium import webdriver
     2
     3 options = webdriver.ChromeOptions()
     4 options.add_argument("--headless") # GUI 화면을 띄울지 유무. 보통 코딩 완료후에는 필요없기 때문에 끈다
     5 options.add_argument('--disable-dev-shm-usage') # 크롬 메모리 제한 푸는것
     6 options.add_argument("--no-sandbox")
     7 user_agent = "Mozilla/5.0 (Windows NT 6.1; Win64; x64; rv:47.0) Gecko/20100101 Firefox/47.0"
     8 options.add_argument('user-agent=' + user_agent)
     9
    10 driver = webdriver.Chrome(options=options)
```

```
[ ]  1 from tqdm import tqdm
     2
     3 book_info = []
     4
     5 for n in tqdm(kyob['판매상품ID'][746:]) :
     6     url = f'https://product.kyobobook.co.kr/detail/{n}' # 책 페이지
     7     driver.get(url)
     8     time.sleep(3)
     9
    10     try :
    11         genre = driver.find_element(By.CLASS_NAME, 'category_list_item').text # 책 장르
    12         contents = driver.find_element(By.CLASS_NAME, 'intro_bottom').text # 책 소개
    13         index = driver.find_element(By.CLASS_NAME, 'book_contents_item').text # 책 목차
    14         book_info.append([n, genre, contents, index])
    15
    16     except :
    17         book_info.append([n])
    18         continue
```

```
[ ]  1 book = pd.DataFrame(book_info)
```

```
[ ]  1 book.columns = ['판매상품ID', '분야', '책 소개', '책 목차']
```

```
[ ]  1 book = book.reset_index()
     2 del book['index'] # 리셋하면서 새로생긴 index열 삭제
     3 book
```

```
[ ]  1 book_final = pd.concat([kyob, book], axis= 1)
```

```
[ ]  1 del book_final[13] # 중복되는 판매상품ID열 삭제
     2 book_final
```

```
[ ]  1 book_final.columns = ['순위', '판매상품ID', '상품코드', '상품명', '저자', '출판사', '분야', '발행(출시)일자', '정가',
     2                       '판매가', '할인율', '적립율', '적립예정포인트', '분야', '책 소개', '책 목차']
```

```
[ ]  1 book_final.to_pickle('book_data_final.pkl')
```