**미니프로젝트:**

**어떤 분야의 책들이 대중들의 이목을 집중시켰을까?.**

**# 아이디어 구상**

**Process of Brainstorming**

1. 프로젝트 직전 [남 서울도서관] 대출목록 분석함
2. 하…… 뭐로 하지? 내가 좋아하는 거 해보고 싶은데
3. 내가 좋아하게 뭐지? 책 읽기
4. 책을 왜 읽지? 배움/깨달음을 얻으려고
5. 어떤 배움/꺠달음? 내 스스로 중요시 하는 가치(예: 운/노력/ 몰입)에 대한 작가의 접근방식을 책을 통해 간접적으로 체득하고 나의 자아를 확립해 나가는 것.
6. 그럼 나는 위 3개의 가치를 중요시 하는데 다른 사람들은? 표본의 범위를 국가 단위로 확장시켜 그 집단이 어떤 자기개발 도서를 읽는지를 한 번 분석해 보자
7. 그럼 뭘 알 수 있는데? 집단의 가치 및 정체성을 데이터를 바탕으로 추정 가능 할 거 같 같은데 비교대상을 미국과 한국으로 잡고 해보자.

**목표:**

**이번 미니프로젝트에서는 작년 한국/미국을 대표하는 온라인서점(교보문고/아마존)의 배스트샐러 중(1~100위)100권의 책들이 분야별로 차지하는 비율을 분석해보자**

**7월에 계획된 메인 프로젝트에서는 표본의 크기를 지난 10년 단위로 넓혀보자**

**기대효과:**

1. 연도별 어떤 분야가 사랑을 받았으며 관심 트렌드에 대한 통계치 파악
2. 연도 별 어떤 가치의 변화가 있었는지를 데이터를 기반으로 분석

(예를들어 2020년도 코로나 시대를 기준으로 전/후 가치변화 분석)

1. 그 변화 속에서 어떤 Insight를 얻을 수 있을지에 대한 탐구.

**본 프로젝트의 핵심목표는 머신 러닝을 통해 도서 분야 별 가중치를 구하고 국가 단위에서 그룹에 속한 구성원들(모집단(국가) 속 표본(그 국가에 속한 사람들)의 identity 즉, 미국과 한국 사람들의 가치와 정체성 분석을 기조로 세움.**

1. **데이터 수집**

2023년 데이터를web-scrapping을 통해 데이터를 추출 및 분석:

1위부터 100위까지의 데이터에서 분야 별로 도서가 차지 하는 비율을 파악해보고

그 비율이 가지는 의미를 탐구해보는 것으로 정하려고 하였으나

데이터 수집 단계에서 어려움 봉착함.

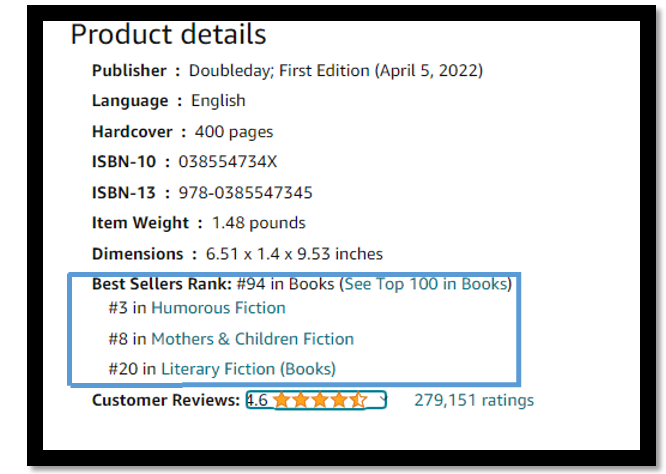
1. **교보문고**

****

<https://product.kyobobook.co.kr/bestseller/total?period=004#?page=1&per=20&period=004&ymw=&bsslBksClstCode=A>

베스트셀러에 대한 정보를 excel파일로 다운받을 수 있게 되어있다.

1. **아마존 (대안: Amazon api활용)**



<https://www.amazon.com/Lessons-Chemistry-Novel-Bonnie-Garmus/dp/038554734X/ref=zg_bsar_g_books_d_sccl_6/146-5497212-4365122?psc=1>

아래 그림은 amazon.com의 웹 페이지의 구성 중 책(product)의 세부정보이다.

주석과 함께 총 3개의 하위 카테고리로 책의 분야가 설명됨

데이터의 특성을 확인해 보면 Fiction이라는 단어가 총 3회 반복 출현된 것을 확인 가능하다.

이를 통해 위 책은 소설 책임을 유추할 수 있음.

# 활용 툴

Python

# web-scrapping : beautifulSoup/selenium

# dataframe : pandas/numpy

# data visualization : matplotlib, power-bi etc.