**2021.11.01.**

프로젝트 진행을 위한 데이터셋 조사를 실시함. Kaggle에서 탐색한 결과

BoardGameGeek Reviews(<https://www.kaggle.com/jvanelteren/boardgamegeek-reviews>)

Board Game Geek Rankings(<https://www.kaggle.com/mseinstein/bgg_top2000>)

등 유용할 수도 있는 데이터셋을 몇가지 찾을 수 있었음. 특히 BoardGameGeek Reviews 데이터셋의 경우 csv 파일의 형태로 특정 유저의 닉네임. 유저가 남긴 보드게임 평점. 보드게임 평론과 같은 데이터를 저장하고 있어서 본 작품을 개발하는데 있어서 큰 도움이 될 것으로 예상됨. 필요한 경우 해당 데이터셋에 더해서 Board Game Geek이라는 웹사이트의 리뷰 데이터를 직접 크롤링하여 데이터를 수집하는 방법을 이용해도 괜찮을 것으로 보임.

**2021.11.11**

BoardGameGeek Reviews 데이터셋에 대한 분석을 진행하였음. 해당 데이터셋을 상세히 살펴본 결과 데이터셋이 협업 필터링에 기반한 추천 시스템 구현에는 적합하지 않을 수도 있을 것으로 생각됨. 한 유저가 다양한 보드게임에 대한 리뷰를 남기기 보다 유저 당 2~3개 정도의 적은 리뷰를 남기는 경우가 많음. 추가적인 데이터셋 탐색이 필요할 것으로 판단됨.

**2021.11.18**

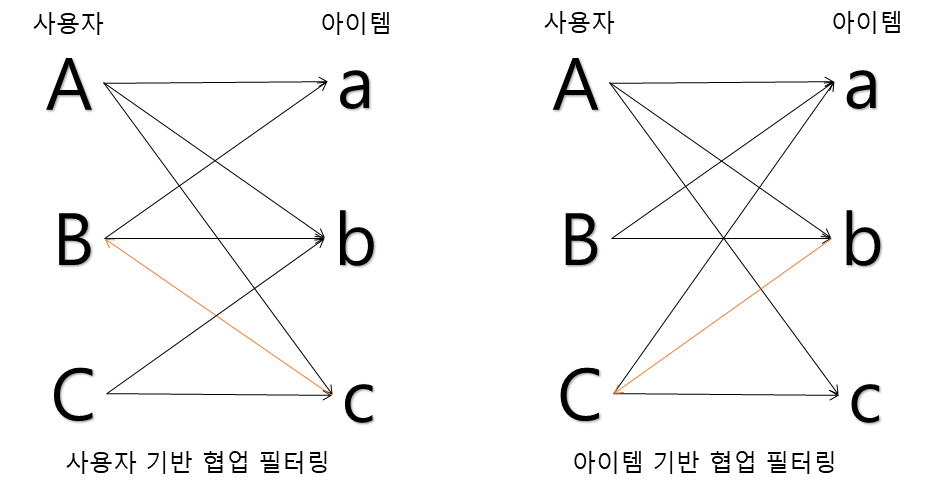
데이터셋의 추가적인 탐색 및 분석을 진행하였음

1. Board Game Data( <https://www.kaggle.com/mrpantherson/board-game-data?select=bgg_db_2018_01.csv> ): 다양한 보드게임의 종류와 각 보드게임 별로 간단한 정보의 요약이 제공됨. 또한 해당 보드게임의 더 상세한 정보를 담고 있는 BoardGameGeek 사이트로의 링크 정보도 제공하고 있기 때문에 크롤링으로 정보를 수집할 경우 유용하게 사용할 여지가 있음.
2. Board Games Dataset( <https://www.kaggle.com/gabrio/board-games-dataset> ): 90000여개의 보드게임에 관한 정보를 담고 있는 데이터셋. 위 Board Game Data에 비해 조금 더 상세한 설명이 포함되어 있는 점을 제외하고는 큰 차이가 없음. 콘텐츠 기반 필터링 구현에 이용할 수 있을 것으로 보임.
3. Board Games( <https://www.kaggle.com/andrewmvd/board-games> ): 위의 데이터셋들과 비교했을 때 Board Game Geeks의 유저 평점 데이터, 전체 랭킹이 추가되었음.

**2021.11.25**

Python을 활용한 협업 필터링 구현 방법에 대한 학습을 진행하였음. 우선 협업 필터링의 개념에 관하여 연구논문작품 제안서에 작성한 내용을 첨부함.

먼저 협업 필터링은 사용자들의 과거 경항이 미래에도 계속될 것이라는 전제를 이용한 알고리즘이다. 협업 필터링은 또다시 사용자 기반 협업 필터링과 아이템 기반 협업 필터링으로 나뉜다. 사용자 기반 협업 필터링은 사용자들의 기호를 분석하여 ‘A 사용자의 기호와 B 사용자의 기호가 서로 비슷하다면 A 사용자가 선호하는 아이템을 B 사용자도 선호할 것이다’라는 추측을 통해 추천하는 방식이다. 그리고 아이템 기반 협업 필터링은 ‘a 아이템을 선호하는 사용자와 b 아이템을 동시에 선호하는 사용자가 많다면 a 아이템과 b 아이템은 서로 유사한 것이다’라는 추측을 통해 추천하는 방식이다.



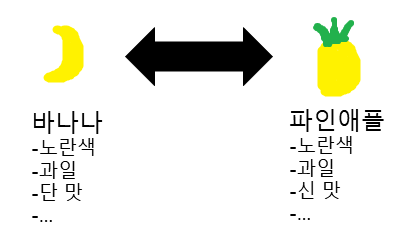
위 그림은 사용자 기반 협업 필터링과 아이템 기반 협업 필터링에 대해 나타내고 있다. 사용자 기반 협업 필터링을 보면 A 사용자와 B 사용자의 기호가 비슷하다고 판단하여 A 사용자만 선호하는 아이템 c를 B 사용자에게도 추천한다. 그리고 아이템 기반 협업 필터링에서는 a 아이템과 b 아이템을 동시에 선호하는 사용자가 많기 때문에 a 아이템을 선호하는 사용자 C에게 b 아이템도 추천하는 모습이다.

Python의 scikit learn을 이용하면 협업 필터링을 쉽게 구현할 수 있음. 사용자가보드게임에 준 평점을 바탕으로 시스템을 구현한다면, 평점은 실수 형태이기 때문에 간단하게 유사도를 구할 수 있기 때문임. Pandas를 활용해 csv 형태로 되어있는 데이터셋을 로드하고, scikit learn의 유사도 계산 함수들을 활용하여 사용자 및 아이템 간의 유사성을 모두 계산하고 이를 이용하여 추천 시스템을 만들면 됨.

**2021.12.02**

Python을 활용한 콘텐츠 기반 필터링 구현 방법에 대한 학습을 진행하였음. 우선 콘텐츠 기반 필터링의 개념에 관하여 연구논문작품 제안서에 작성한 내용을 첨부함.

콘텐츠 기반 필터링은 문자 그대로 각 콘텐츠(아이템)들이 가진 특성들의 유사도를 이용하여 추천하는 방식이다.



위 그림은 콘텐츠 기반 필터링에 대해 나타내고 있다. 바나나라는 아이템과 파인애플이라는 아이템의 특성들이 비슷하므로 바나나를 좋아하는 사용자에게는 파인애플을, 파인애플을 좋아하는 사용자에게는 바나나를 추천하게 될 것이다.

콘텐츠 기반 필터링에서 협업 필터링과는 다르게 주의를 기울여야 할 부분은 콘텐츠 기반 필터링에는 문자열 형태의 데이터가 이용된다는 점임. 따라서 협업 필터링처럼 간단하게 유사도를 계산할 수가 없음. 문자열 간의 유사도를 계산하기 위해서는 먼저 문자를 벡터화하는 작업이 필요함. Python의 scikit learn에서 제공하는 CountVectorizer 및 TfidfVectorizer, gensim의 Word2Vec 등을 고려해 볼 만 함.