선택한 데이터: H&M Personalized Fashion Recommendations

선택한 Evaluation metric selection: MAP @ K, NDCG @ K

선택한 모델: Content-based model & item-based collaborative filtering

해당 데이터는 articles.csv, customers.csv, transactions.csv로 구성되었다.

# 간단한 EDA

이때, article.csv에는 제품의 코드와 이름, 색상, 추가된 색상 정보, 의류 그룹 등이 있다. 그 중에서 몇 가지 열에 대한 그래프를 살펴보자.

article1. product\_group\_name

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

각 그룹에 속한 product가 많은 순으로 나열했을 때의 TOP3: Garment Upper body > Garment Lower body > Garment Full body

article2. Index\_group\_name

텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명

각 군에 속하는 product가 많은 순으로 나열했을 때의 TOP3: Ladieswear>Baby/children >Divided

article3. Number of Product Types per each Product Group (각 그룹당 속해 있는 product type의 수)

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

TOP3: Accessories>Shoes>Garment upper body

Transactions.csv에는 고객의 id, 아이템의 id, 제품의 가격, 제품이 판매된 경로가 기록되어 있다. 따라서 이를 통해 어떤 고객이 어떤 아이템을 구매했는지 알 수 있다. transaction에서는 다른 EDA보다는 시간에 따라 아이템을 몇 개 구매했는지 살펴보았다.

친필, 라인, 폰트, 텍스트이(가) 표시된 사진

자동 생성된 설명

Customer.csv에는 고객의 대한 개인정보가 기록되어 있다.

Customers1. club\_member\_status

텍스트, 영수증, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

각 멤버 상태에 몇 명의 고객이 속해 있는 지 알아보았다.

Customers2. 각 club membership에 어느정도 나이대의 사람들이 있는 가

텍스트, 스크린샷, 번호, 도표이(가) 표시된 사진

자동 생성된 설명

비교적 ACTIVE와 PRE-CREATE에는 다양한 나이대가 있었지만, LEFT CLUB에는 60세 이하의 나이대가 많았다.

# Content-based

One-hot encoding을 할 열과, embedding을 할 열을 나눠서 전처리를 진행하였다.

그 중에서도 처음으로 articles에서는, content의 정보를 뽑아오기 위해, detail\_desc 열에 있는 문장들을 분석하여 따로 전처리를 진행하였고, product\_type\_name, department\_name, colour\_group\_name등의 열들을 이와 합쳐 desc\_full 열을 생성하여 list로 만들어 text\_normalized라는 txt파일을 생성하여 사용하였다.

다음으로, customers에서는 null값이 많아 0이나 평균값으로 채워주었다. 이후에 one-hot encoding을 일부 열에 적용하였다.

전처리 이후에는 transactions 데이터로 train/test 데이터를 만들어 학습을 진행하였다. 이때 test data로는 마지막 하루를 지정하였다. 이후에 진행할 item-based 모델도 이 데이터에 대해 적용하였다.

모델은 tensorflow의 keras를 사용해서 구성하였는데, 활성화함수는 relu를 사용하였다. 자세한 모델의 구성은 코드에서 확인할 수 있다. 또한 model학습 중에 가장 좋은 성능을 보여주는 모델을 선택하여 저장 후, predict에 사용하도록 구성하였다.

# Item-based

각 고객들이 구매한 article의 개수로 transcation.csv를 필터링하여 “bought”열로 추가하여 사용하였다. 또한 positive sample 뿐만 아니라, random으로 negative sample을 만들어 모델에 적용하였다.

모델은 수업시간에 배운 모델의 구성 순서와, 캐글 자료를 참고하여 구성하였다. 해당 모델을 앞선 content-based와 비교하는 것이 목적이기 때문에 epoch등은 일치시켜주었다.

# evaluation method를 통한 비교

선택한 두 method를 통해 각 모델의 결과를 비교하였다. 이때 정확한 비교를 위하여 test\_data는 동일하게 설정하였으며, k 또한 동일하게 설정하였다. 해당 모델들의 결과는 다음과 같았다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

두 method 모두 content\_based 모델의 결과가 훨씬 좋게 나왔으며, 이는 10개의 예측 중 성공한 예측이 content\_based 모델이 더 많았다는 것을 의미한다. 이 점에서 이 method들을 선택한 이유를 알 수 있다. H&M의 경우 쇼핑몰이기 때문에, 첫페이지에 어떤 추천을 보여주는지, 또는 광고에서 어떤 추천을 보여주는 지가 소비자가 클릭할 확률을 높이는 데에 큰 기여를 한다. 만약 전체 데이터에 대한 추천 예측이 훨씬 높더라도, 정작 목표 고객과 맞지 않는 추천을 한다면 광고등을 통해 고객들을 끌어들이지 못할 것이다. 따라서 TOP N개 중 성공한 예측의 개수를 확실히 알 수 있는 두 지표인 MAP와 NDCG를 선택하였다.