AIB CP2 프로젝트 최종 보고서

[개인 프로젝트]

프로젝트 참여자 : 이예지
프로젝트 시작일 : 2022. 09. 15.
프로젝트 명칭 : 잇템(IT:EM)
프로젝트 분류 : DS

1. 프로젝트 제목

본 프로젝트의 명칭은 “잇템**(IT:EM)**”이라 한다.

2. 프로젝트 개요

눈 깜박할 사이에 새로운 시즌이 돌아오고, 그때마다 신상이 쏟아지는 패션업계. 이들 신상은 오프라인 매장과 온라인 쇼핑몰을 통해 만난다. 대부분의 고객은 본인의 스타일을 벗어나지 않는 선에서 주기적으로 패션 제품을 구매하고자 한다. 하지만 본인의 스타일을 몰라 쇼핑하는 데 불편함을 겪는 고객도 다수 있다.

그래서 본 프로젝트는 가장 최근 한 주의 인기 상품과 고객의 구매 내역을 기반으로 이전에 자주 구매한 상품과 이전에 구매한 상품과 유사한 상품을 추천하여 그들만의 스타일을 추천하여 상품을 고르는 시간도 절약하고, 유사한 상품으로 스타일에 대한 만족도도 높이고, 패션 제품을 구매하는 모든 고객의 불편을 줄이고자 하는 것을 목표로 한다.

3. 팀 구성 및 역할

- 프로젝트 기획
- 데이터 수집
- 데이터 시각화
- 추천시스템 개발
    - 마지막 주 인기 상품 10개 추천
    - 고객별 이전에 자주 구매한 상품 10개 추천
    - 고객별 이전에 구매한 상품과 유사한 상품 추천 10개 추천
        - 자연어 처리
        - 텍스트 벡터화
            - countvectorizer
            - tf-idf
        - 코사인 유사도
        - knn
- 보고서 작성

4. 프로젝트 배경 및 목적

코로나 19로 언택트 시대가 되면서 온라인 쇼핑몰은 점점 그 규모가 커지고 있는데 패션 시장도 온라인에서 크게, 또 빠르게 성장하고 있다. 온라인 선택의 확대로 인해 소비자들은 더욱 개인화된 추천 서비스에 대한 요구가 커지고 있다.

이제 패션 업계에서는 모든 소비자를 타깃으로 하는 사업은 실패하기 쉽다고 본다. 따라서 AI 추천을 통해 단 한 사람만을 위해 이뤄지는 추천 시스템이 기본이 되어야 한다.

목적

- 트렌드에 민감하게 반응하는 고객을 위해 빠르게 변화하는 트렌드를 반영하여 최근 인기 상품을 추천하는 시스템을 개발한다.
- 주기적으로 상품을 구매하는 고객에 이전에 자주 구매한 상품을 추천하여 고객의 이탈률을 줄이고자 한다.
- 쇼핑 시간을 줄이고 본인만의 스타일로 상품을 구매하려는 고객을 위해 고객별로 구매했던 상품과 유사한 상품을 추천하는 시스템을 개발한다.
- 고객의 만족도를 더욱 높이기 위해 고객별 구매기록을 기반으로 추천시스템을 개발한다.

5. 방법 설명 및 결과

<데이터 설명>
articles.csv

- 제품 데이터
- (105542, 25)
- 총 105542 개 제품에 대한 정보가 담겨 있음
- 한 행은 제품에 대한 특징 데이터를 의미함

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 105542 entries, 0 to 105541
Data columns (total 25 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   article_id                    105542 non-null  int64
 1   product_code                  105542 non-null  int64
 2   prod_name                     105542 non-null  object
 3   product_type_no               105542 non-null  int64
 4   product_type_name             105542 non-null  object
 5   product_group_name            105542 non-null  object
 6   graphical_appearance_no       105542 non-null  int64
 7   graphical_appearance_name     105542 non-null  object
 8   colour_group_code             105542 non-null  int64
 9   colour_group_name             105542 non-null  object
 10  perceived_colour_value_id     105542 non-null  int64
 11  perceived_colour_value_name   105542 non-null  object
 12  perceived_colour_master_id    105542 non-null  int64
 13  perceived_colour_master_name  105542 non-null  object
 14  department_no                 105542 non-null  int64
 15  department_name               105542 non-null  object
 16  index_code                    105542 non-null  object
 17  index_name                    105542 non-null  object
 18  index_group_no                105542 non-null  int64
 19  index_group_name              105542 non-null  object
 20  section_no                    105542 non-null  int64
 21  section_name                  105542 non-null  object
 22  garment_group_no              105542 non-null  int64
 23  garment_group_name            105542 non-null  object
 24  detail_desc                   105126 non-null  object
dtypes: int64(11), object(14)
memory usage: 20.1+ MB
None
```

	article_id	product_code	prod_name	product_type_no	...	section_name	garment_group_no	garment_group_name	detail_desc
0	108775015	108775	Strap top	253	...	Womens Everyday Basics	1002	Jersey Basic	Jersey top with narrow shoulder straps.

[1 rows x 25 columns]

customers.csv

- 고객 데이터
- (1371980, 7)
- 총 1371980명에 대한 개인정보가 존재함
- 고객 id, 우편번호는 개인정보로 암호화되어 있음
- 하나의 행은 고객의 고유 데이터를 의미함
- FN(패션 뉴스 구독 여부), Active(커뮤니케이션 가능 여부), Club member status(신규, 활성화, 탈퇴), Fashion news frequency(패션 뉴스 알람 주기), Age(나이는 16-99)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1371980 entries, 0 to 1371979
Data columns (total 7 columns):
 #   Column                  Non-Null Count    Dtype
---  ------                  --------------    -----
 0   customer_id             1371980 non-null  object
 1   FN                      476930 non-null   float64
 2   Active                  464404 non-null   float64
 3   club_member_status      1365918 non-null  object
 4   fashion_news_frequency  1355971 non-null  object
 5   age                     1356119 non-null  float64
 6   postal_code             1371980 non-null  object
dtypes: float64(3), object(4)
memory usage: 73.3+ MB
None
```

	customer_id	FN	Active	club_member_status	fashion_news_frequency	age	postal_code
0	00000dbacae5abe5e23885899a1fa44253a17956c6d1c3...	NaN	NaN	ACTIVE	NONE	49.0	52043ee2162cf5aa7ee79974281641c6f11a68d276429a...
1	0000423b00ade91418cceaf3b26c6af3dd342b51fd051e...	NaN	NaN	ACTIVE	NONE	25.0	2973abc54daa8a5f8ccfe9362140c63247c5eee03f1d93...
2	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	NaN	NaN	ACTIVE	NONE	24.0	64f17e6a330a85798e4998f62d0930d14db8db1c054af6...
3	00005ca1c9ed5f5146b52ac8639a40ca9d57aeff4d1bd2...	NaN	NaN	ACTIVE	NONE	54.0	5d36574f52495e81f019b680c843c443bd343d5ca5b1c2...
4	00006413d8573cd20ed7128e53b7b13819fe5cfc2d801f...	1.0	1.0	ACTIVE	Regularly	52.0	25fa5ddee9aac01b35208d01736e57942317d756b32ddd...

transactions.csv

- 거래기록 데이터
- (31788324, 5)
- 2018.09.20 - 2020.09.22 발생한 약 3천만 건의 거래기록이 담겨 있음
- 하나의 행은 제품 1회 구매기록을 의미함
- t_dat(구매 날짜), price(실제 가격의 /590), sales_channel_id(구매방식 1: 오프라인, 2: 온라인)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31788324 entries, 0 to 31788323
Data columns (total 5 columns):
 #   Column            Dtype
---  ------            -----
 0   t_dat             object
 1   customer_id       object
 2   article_id        int64
 3   price             float64
 4   sales_channel_id  int64
dtypes: float64(1), int64(2), object(2)
memory usage: 1.2+ GB
None
```

	t_dat	customer_id	article_id	price	sales_channel_id
0	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	663713001	0.050831	2
1	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	541518023	0.030492	2
2	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	505221004	0.015237	2
3	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687003	0.016932	2
4	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687004	0.016932	2

빠르게 변화하는 패션 업계에서 본인의 스타일을 모르는 소비자가 있다고 생각해 가장 최근에 전체적으로 인기 있는 상품을 추천한다.

<recommendations_popular.py>

- 가장 마지막 주**(20.09.16. - 20.09.22.)**의 인기 상품 **10**개 추천하기

```
924243001 924243002 918522001 923758001 866731001 909370001 751471001 915529003 915529005 448509014
```

또한 고객이 자주 찾는 상품을 추천함으로써 고객의 재구매율을 높인다.

<recommendations_previously.py>

- 고객별 이전에 자주 구매한 상품 **10**개 추천하기
  - 고객별 상품 구매 기록 구하기
    - ex) '8536c0c8b77f15197e75eb25aaf11663732b632f6e2abcadd1907e9f372f108f': {562245001: 1, 562245059: 1}
  - 구매기록으로 인기있는 상품 10개 구하기

```
[706016001, 706016002, 372860001, 610776002, 759871002, 464297007, 372860002, 610776001, 399223001, 706016003]
```

  - 고객별로 상품 구매 기록을 내림차순하기
    - 구매 기록이 10개 초과가 될 경우 : 10개 까지만
    - 구매 기록이 10개가 되지 않지 않을 경우 : 인기있는 상품까지 합쳐 10개 까지(ex) 구매기록 7개 + 인기 상품 top 3)

	customer_id	recommendations
0	00000dbacae5abe5e23885899a1fa44253a17956c6d1c3...	568601006 797065001 625548001 176209023 627759...
1	0000423b00ade91418cceaf3b26c6af3dd342b51fd051e...	811835004 723529001 351484002 689898002 583558...
2	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	663713001 351484002 750424014 870304002 541518...
3	00005ca1c9ed5f5146b52ac8639a40ca9d57aeff4d1bd2...	742079001 732413001 706016001 706016002 372860...
4	00006413d8573cd20ed7128e53b7b13819fe5cfc2d801f...	399061015 634249005 677049001 698286003 707704...

더 나아가 고객의 만족도를 높이기 위해 고객이 구매했던 상품들과 유사한 상품들을 추천한다.

<recommendations_count.py>

- 상품에 대한 설명에 자연어 처리 과정을 거침
  - article에서 object형이 모두 상품에 대한 설명이라고 판단하여 하나로 합쳐 자연어 처리(소문자화, 불용어 제거, 공백 제거 등)를 함

	article_id	text
0	108775015	Strap top Vest top Garment Upper body Solid Bl...
1	108775044	Strap top Vest top Garment Upper body Solid Wh...
2	108775051	Strap top (1) Vest top Garment Upper body Stri...
3	110065001	OP T-shirt (Idro) Bra Underwear Solid Black Da...
4	110065002	OP T-shirt (Idro) Bra Underwear Solid White Li...

	article_id	text
0	108775015	strap top vest top garment upper body solid bl...
1	108775044	strap top vest top garment upper body solid wh...
2	108775051	strap top vest top garment upper body stripe w...
3	110065001	shirt idro bra underwear solid black dark blac...
4	110065002	shirt idro bra underwear solid white light whi...

- 상품 설명에 대해 벡터화를 함
  - 상품에 대한 설명이 맥락을 고려할 필요가 없고, 벡터 공간 내 단어 할당을 임의적으로 줘도 상관없다고 생각해 임베딩이 아닌 텍스트 벡터화를 함
  - 단어 피쳐에 값을 부여할 때, 각 문서에서 해당 단어가 나타나는 횟수, 즉 Count를 부여하는 경우 **=> CountVectorizer**

    - 카운트 벡터화에서는 값이 높을수록 중요한 단어로 인식
- 사용자별 구매기록에 기반하여 유사상품 **10**개를 추천함
    - 임의로 사용자를 지정하고, 그 사용자의 거래 기록에 기반하여 구매 상품의 설명에 대해 벡터화를 함
    - 이전에 구한 전체 상품 설명의 벡터값과 사용자의 거래 기록에 기반한 상품의 벡터값의 코사인 유사도로 비교함
        - 코사인 유사도 : 두 벡터의 코사인 각도를 이용하여 구할 수 있는 두 벡터의 유사도를 의미
    - 코사인 유사도 값이 큰 순서대로 10개를 뽑아 추천함

```
                                        customer_id article_id     score  \
0  8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...  561758003  0.958723
1  8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...  561793001  0.958723
2  8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...  561796001  0.958723
3  8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...  561797002  0.958723
4  8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...  561868003  0.958723
5  8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...  561758001  0.955298
6  8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...  561764003  0.955298
7  8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...  561776005  0.955298
8  8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...  561790001  0.955298
9  8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...  561758017  0.946959

                                                text
0  sunglasses josef sunglasses accessories solid ...
1  fork jogger trousers garment lower body solid ...
2  twiggy shirt garment upper body solid white li...
3  mike shirt shirt garment upper body solid blac...
4  vivi velvet top top garment upper body solid d...
5  sunglasses josef sunglasses accessories patter...
6  attract sweater garment upper body melange gre...
7  class iggy ringpk ring accessories solid light...
8  fargo pkt trousers garment lower body solid da...
9  sunglasses josef sunglasses accessories solid ...
```

단순히 단어의 빈도만 고려한다면 모든 문서에서 자주 쓰일 수 밖에 있는 단어들이 중요하다고 인식될 수 있다는 문제(ex) a 문서에서 'the'가 가장 많이 등장하고 b 문서에도 'the'가 가장 많이 등장한다고 해서 두 문서가 유사한 문서라고 판단할 수 없다)가 있다고 판단하여 이와 같은 문제를 보완하고자 했다.

<recommendations_tfidf.py>

- 상품에 대한 설명에 자연어 처리 과정을 거침 (recommendations_count.py와 동일)
- 상품 설명에 대해 벡터화를 함
    - 상품에 대한 설명이 맥락을 고려할 필요가 없고, 벡터 공간 내 단어 할당을 임의적으로 줘도 상관없다고 생각해 임베딩이 아닌 텍스트 벡터화를 함
    - 개별 문서에서 자주 등장하는 단어에 높은 가중치를 주되, 모든 문서에서 전반적으로 자주 등장하는 단어에 대해서는 페널티를 주는 방식 **=> TF-IDF(Term Frequency - Inverse Document Frequency)**
    - 해당 단어가 실질적으로 중요한 단어인지 검사
    - 문서의 양이 많을 경우에 일반적으로 CountVectorizer보다 TF-IDF 벡터화를 사용
- 사용자별 구매기록에 기반하여 유사상품 **10**개를 추천함
    - 임의로 사용자를 지정하고, 그 사용자의 거래 기록에 기반하여 구매 상품의 설명에 대해 벡터화를 함
    - 이전에 구한 전체 상품 설명의 벡터값과 사용자의 거래 기록에 기반한 상품의 벡터값의 코사인 유사도로 비교함
    - 코사인 유사도 값이 큰 순서대로 10개를 뽑아 추천함

	customer_id	article_id	score	\
0	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561758003	0.977567	
1	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561793001	0.977567	
2	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561796001	0.977567	
3	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561797002	0.977567	
4	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561868003	0.977567	
5	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561758001	0.974681	
6	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561764003	0.974681	
7	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561776005	0.974681	
8	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561790001	0.974681	
9	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561758013	0.973049	

	text
0	sunglasses josef sunglasses accessories solid ...
1	fork jogger trousers garment lower body solid ...
2	twiggy shirt garment upper body solid white li...
3	mike shirt shirt garment upper body solid blac...
4	vivi velvet top top garment upper body solid d...
5	sunglasses josef sunglasses accessories patter...
6	attract sweater garment upper body melange gre...
7	class iggy ringpk ring accessories solid light...
8	fargo pkt trousers garment lower body solid da...
9	sunglasses josef sunglasses accessories solid ...

CountVectorizer는 단순히 빈도를 기반으로 표현을 해주고, TF-IDF 벡터화는 다른 문장에서의 단어 빈도도 고려하여 해당 단어의 중요도를 표현해준다. 상품 설명에 대한 단어가 반복적으로 이루어지는 이 프로젝트에서는 TF-IDF가 적절한 벡터화 방식이라 생각한다.

KNN은 거리기반 분류분석 모델로, 데이터로부터 거리가 가까운 'K개'의 다른 데이터의 레이블을 참조하여 분류하는 알고리즘이다. 간단한 알고리즘이지만 상품 추천 알고리즘으로 많이 사용된다.

<recommendations_knn.py>

- 상품에 대한 설명에 자연어 처리 과정을 거침 (recommendations_count.py와 동일)
- 상품 설명에 대해 벡터화를 함(recommendations_tfidf.py와 동일)
- 사용자별 구매기록에 기반하여 유사상품 열개를 추천함
    - 임의로 사용자를 지정하고, 그 사용자의 거래 기록에 기반하여 구매 상품의 설명에 대해 벡터화를 함 => TF-IDF
    - 가장 가까운 거리에 있는 11개의 데이터를 찾음
    - 이전에 구한 전체 상품 설명의 벡터에서 11개의 이웃을 찾음

```
(array([[0.21181614, 0.21181614, 0.21181614, 0.21181614, 0.21181614,
        0.22503064, 0.22503064, 0.22503064, 0.22503064, 0.23216836,
        0.23216836]]), array([[14193, 14173, 14174, 14178, 14155, 14172, 14170, 14167, 14154,
        14160, 14171]], dtype=int64))
```

    - 유클라디안 거리로 계산한 스코어가 높은 순서대로 10개를 뽑아 추천함
        - 유클라디안 거리 : 피타고라스 정리

	customer_id	article_id	score	\
0	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561793001	0.211816	
1	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561796001	0.211816	
2	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561797002	0.211816	
3	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561758003	0.211816	
4	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561790001	0.225031	
5	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561776005	0.225031	
6	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561764003	0.225031	
7	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561758001	0.225031	
8	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561758013	0.232168	
9	8536c0c8b77f15197e75eb25aaf11663732b632f6e2abc...	561784001	0.232168	

	text
0	fork jogger trousers garment lower body solid ...
1	twiggy shirt garment upper body solid white li...
2	mike shirt shirt garment upper body solid blac...
3	sunglasses josef sunglasses accessories solid ...
4	fargo pkt trousers garment lower body solid da...
5	class iggy ringpk ring accessories solid light...
6	attract sweater garment upper body melange gre...
7	sunglasses josef sunglasses accessories patter...
8	sunglasses josef sunglasses accessories solid ...
9	dallas badge trs trousers garment lower body s...

knn과 tf-idf이 거의 유사하게 추천하지만, countvectorizer는 조금 다르게 추천하는 것을 확인할 수 있다.

6. 기대 효과

빠르게 변화하는 패션 업계에서 모든 고객에 가장 최근에 전체적으로 인기있던 상품을 추천함으로써 모든 고객이 트렌드를 따라갈 수 있도록 한다. 또한 주기적으로 상품을 구매하는 고객에 이전에 자주 구매한 상품을 추천하기 때문에 고객이 재구매할 수 있도록 한다. 고객별로 구매했던 상품과 유사한 상품을 추천하여 본인만의 스타일로 상품을 구매하는 고객들에게 쇼핑 시간을 단축하고 빠른 구매를 결정하는 데 도움을 줄 수 있다. 더욱이 개인의 만족도를 높이기 위해 고객별 구매기록을 기반으로 추천 시스템을 제공하는데 의의를 두고 있다.

7. 프로젝트 회고

프로젝트를 풀스택으로 해보고 싶어 개인 프로젝트를 하게 되었다. 하지만 프로젝트를 진행하면서 무엇을, 어떻게 구현할 지 다방면으로 고민해야하다보니 그 생각이 오만한 생각이었음을 깨달았다. 그래서 선택과 집중을 통해 프로젝트의 핵심 부분을 완성하여 기한 내 최종 결과물을 만들어내고자 했다.

결국 가장 마지막 주에 전반적으로 인기있던 상품, 고객별로 이전에 자주 구매한 상품, 그리고 이전에 구매한 상품과 유사한 상품 번호를 추천하는 잇템(IT:EM)을 구현해내며 나름 만족스러운 결과물을 완성해냈다.

하지만 이번 프로젝트를 진행하면서 아쉬운 점도 많이 있었다. 먼저, 혼자 진행하다보니 나태해지기도 했고, 또 하고 있는 것에 대해 잘 하고 있는 것인지 확신이 없었다. 그래서

앞으로는 올바른 방향성을 제시해주고, 또 프로젝트에 도움을 줄 수 있는 사람과 함께 프로젝트를 진행한다면 완성도가 조금 더 높은 프로젝트를 할 수 있지 않을까라는 생각을 했다. 이번 프로젝트를 통해 기업에서 백엔드, 프론트엔드, DE, DS 등 직무를 나누는 이유를 조금이나마 이해할 수 있었다.