통계 데이터 사이언스 개인 과제

MBTI 와 작성 게시글을 기반으로 한 취미 추천 시스템

손은주

서 울 대 학 교

MBTI와 작성 게시글을 기반으로 한 취미 추천 시스템

손은주

요약

본 분석의 목적은 작성한 게시글로 개인의 성격을 나타내는 MBTI를 파악하고, 어떠한 취미를 추천해주는지 알아보고자 한다. PersonalityCafe forum의 MBTI의 데이터로 텍스트를 분석하여 MBTI의 유형을 분류하는 Logistic모형을 선택했다. FSEV UK의 설문자료를 바탕으로 요인분석을 하여 MBTI의 유형을 파악하고 아이템 기반, 잠재요인 협업 필터링을 통해 어떠한 취미를 추천해주는지 확인해 보았다.

1. 서론

1.1. 주제 선정 이유 및 자료 선택 이유

취미나 흥미는 스트레스를 해소시키고 삶의 질을 높이는 요소 중 하나이다. 그러므로 바쁜 현대 사회에서 취미생활을 알고 갖는 것은 긍정적인 효과를 볼 수 있다. 취미를 탐색하기 위해 자신의 성향을 파악하거나 비슷한 성향을 가진 사람들의 취미생활을 알 수 있다면 취미를 선택하는 것에 어려움이 없을 것이라고 생각한다. MBTI성격유형검사는 성향을 파악할 수 있는 조사 중 하나로 최근 M-Z세대(밀레니얼 세대와 Z세대) 사이에서 크게 유행하고 있다. 이는 코로나19로 인한 상호간의 연결이 줄어듦에 따른 재미를 찾기 위함과 간단한 설문들로 자신의 성향을 쉽게 파악할 수 있다는 장점 때문에 화제가 된 것이라고 볼 수 있다. 자신과 같은 MBTI 성향을 가진 사람들의 흥미와 취미를 고려하여 취미를 선택하게 된다면, 다양한 취미 생활 중 고민의 폭을 좁히면서 보다 개인 맞춤화된 취미나 흥미를 찾을 수 있을 것으로 보인다.

본 연구는 단순히 설문자료를 통해 취미와 흥미를 찾는 것이 아니라 개인의 작성 게시글을 통해 MBTI의 유형을 유추하고, 비슷한 성향을 보이는 사람들과의 취미와 흥미를 알고 추천할 수 있는 서비스를 만들어 보고자 한

다. 즉, 개인의 게시글을 통해 취미를 추천해 줄 수 있다는 것이 이번 연구의 Originality이자 흥미로운 주제라고 생각하였다.

1.2. 데이터

첫번째 데이터는 PersonalityCafe forum을 통해 수집된 데이터로 MBTI의 유형과 마지막으로 게시한 50개의 작성 게시물의 변수로 구성되어 있다. 두번째 자료는 2013년에 슬로바키아에 있는 FSEV UK대학의 통계학 수업에서 진행했던 설문 조사자료이다. 음악 선호도, 영화 선호도, 취미와 흥미, 성격 유형, 건강 습관 등의 질문들로 이루어져 있다. 위의 두 자료는 캐글(https://www.kaggle.com/)에서 다운로드 할 수 있다.

1.3 연구 목적

작성한 게시글을 보고 MBTI를 예측하고, 예측한 MBTI 유형으로 취미를 추천할 수 있는 서비스를 만드는 것이 이번 연구의 목적이다. 이를 위해 게시글의 불필요한 용어 제거, 토큰화와 같은 텍스트 분석과 로지스틱 모형, 랜덤포레스트 등의 모형을 사용하여 MBTI의 유형을 잘 예측하는 모형을 선택한다. MBTI의 유형이 결정되었으면 같은 성향을 가진 사람들과의 유사도 또는 잠재 요인 협업 필터링을 통해 취미를 추천할 수 있는 서비스를 개발해 본다. 작성 게시글로 취미를 파악하여 자신의 취미 선택이나 결정에 도움을 주는 것을 목표로 한다.

2. 본론

2-1. 첫번째 데이터 - 작성 게시글과 MBTI 데이터

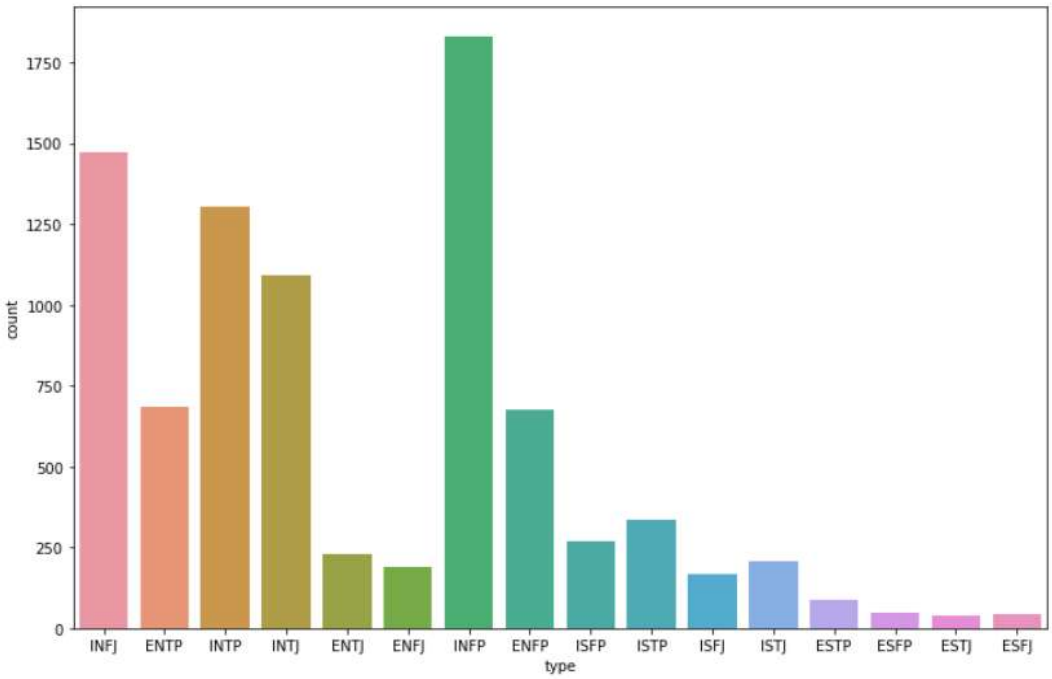
2-1-1. 텍스트 분석 및 데이터 전처리

우선 PersonalityCafe forum 에서 수집된 MBTI 데이터로 텍스트 분석을 해보고자 한다. <표 1>을 보면 MBTI 의 유형과 최근 50 개의 작성 게시글로 구성되어 있는 것을 볼 수 있다. .

	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	ENTP	'I'm finding the lack of me in these posts ver...
2	INTP	'Good one _____ https://www.youtube.com/wat...
3	INTJ	'Dear INTP, I enjoyed our conversation the o...
4	ENTJ	'You're fired.|||That's another silly misconce...
5	INTJ	'18/37 @.@|||Science is not perfect. No scien...
6	INFJ	'No, I can't draw on my own nails (haha). Thos...
7	INTJ	'I tend to build up a collection of things on ...
8	INFJ	I'm not sure, that's a good question. The dist...
9	INTP	'https://www.youtube.com/watch?v=w8-egj0y8Qs||...

<표 1> MBTI 유형과 최근 게시물로 구성된 데이터

위 데이터는 총 8675 개의 행으로 이루어져 있고 MBTI 의 유형의 분포를 알아보기 위해 seaborn 을 활용하여 확인해보았다.



<그림 1> MBTI 유형의 분포

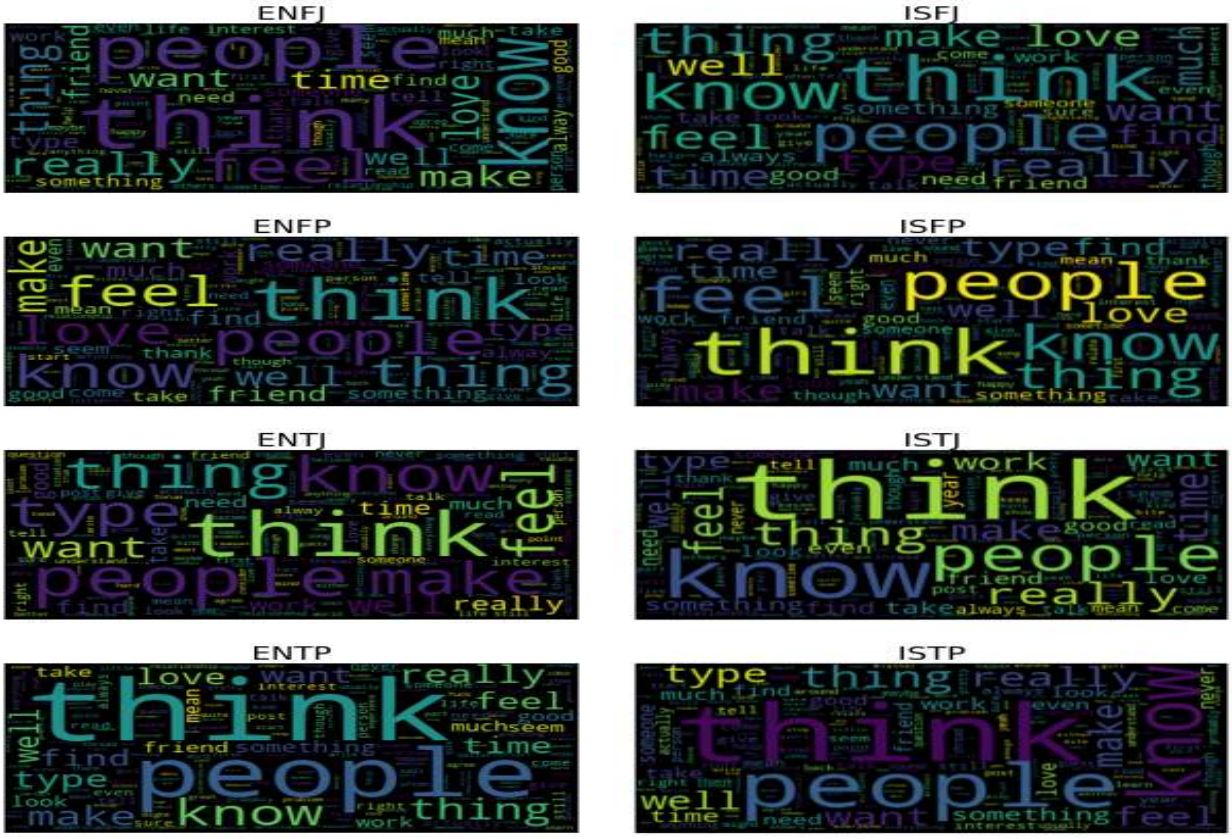
<그림 1>을 보면 INFP, INFJ, INTP, INTJ 순으로 많이 분포 되어있는 것을 확인할 수 있으며 ESTJ 의 유형이 가장 작은 빈도 수를 나타내는 것으로 보인다. 최근 작성 게시글인 <표 1>의 posts 변수를 보면 'http://www...'의 url, '@,|||'의 기호 등 불필요한 글자들을 볼 수 있다. 불필요한 글자들은 텍스트

분석을 진행하는데 영향을 줄 수 있으므로 제거해줘야 한다. 'cleansing_text' 함수를 정의하여 '|||', url, tag, 기호, 구두점, 숫자 등을 제거해 주고 소문자로 변환시켜주었다. 또한 NLTK, Lemmatize 등의 모듈을 사용하여 토큰화와 표제어를 추출하고 불용어, 3 이하인 단어를 제거해주었다.

```
텍스트 분석 전**********
'http://www.youtube.com/watch?v=qsXHcwe3krw|||http://41.media.tumblr.com/tumblr_lfou
y03PMA1qa1rooo1_500.jpg|||enfp and intj moments  https://www.youtube.com/watch?v=iz7
IE1g4XM4  sportscenter not top ten plays  https://www.youtube.com/watch?v=uCdfzelete
c  pranks|||What has been the most life-changing experience in your life?|||http://w
텍스트 분석 후**********
moments sportscenter play prank life change experience life repeat today perc experi
ence immerse last thing friend post facebook commit suicide next rest peace hello so
rry hear distress natural relationship perfection time every moment existence figure
hard time time growth welcome stuff game match prozac wellbrutin least thirty minute
```

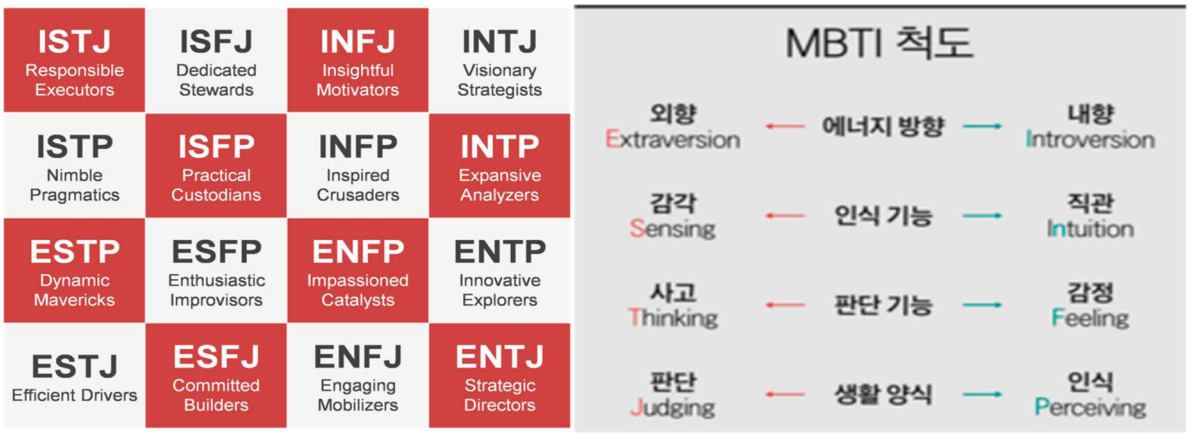
<그림 2> 텍스트 분석 결과

<그림 2>는 첫번째 행의 텍스트 분석 결과를 비교하여 나타냈다. 분석 후의 결과를 보면 url, 불용어, 기호 등이 제거가 된 것을 볼 수 있다.



<그림 3> 워드클라우드 결과

단어의 빈도수가 많은 것을 크게, 적은 것을 작게 하여 MBTI 별로 워드클라우드를 그려보았다. 16 개의 MBTI 유형의 워드클라우드 중 8 개의 결과만 나타낸 것이 <그림 3>이다. 대체로 'think'와 'know', 'people'이 공통적으로 많이 나타나는 것으로 나타났다..



<그림 4> MBTI 척도

MBTI 데이터는 <그림 4>의 왼쪽 그림과 같이 MBTI 의 type 을 16 가지의 유형으로 나누었다. 본 연구에서는 다중 분류를 진행하기 보다는 <그림 4>의 오른쪽 그림과 같이 E(외향형)-I(내향형), S(감각형)-N(직관형), T(사고형)-F(감정형), J(판단형)-P(인식형)으로 각 특성을 고려하여 이진분류로 분리해 분석하고자 한다. 이는 보다 정확한 성향을 파악하는데 도움이 될 것이라고 생각하였다. 그러므로 위의 type 칼럼에 있는 'INFJ, ENTP, ...' 의 MBTI 유형들의 에너지 방향, 인식기능, 판단 기능, 생활 양식의 열을 생성하여 0,1 로 나누었다. <E-I>에서 I 는 0, E 는 1, <S-N>에서 S 는 0, N 은 1, <T-F>에서 T 는 0, F 는 1, <J-P>에서는 P 를 0, J 를 1 로 보았다. <표 2>는 최종 테이블로 MBTI 를 예측하기 위한 준비를 마쳤다.

	type	clean1	E-I	S-N	T-F	J-P
0	INFJ	moments sportscenter play prank life change ex...	0	1	1	1
1	ENTP	find lack post alarm bore position often examp...	1	1	0	0
2	INTP	good course know bless curse absolutely positi...	0	1	0	0
3	INTJ	dear enjoy conversation esoteric gabbing natur...	0	1	0	1
4	ENTJ	fire another silly misconception approach logi...	1	1	0	1
5	INTJ	science perfect scientist claim scientific inf...	0	1	0	1
6	INFJ	draw nail haha professionals nail mean post na...	0	1	1	1
7	INTJ	tend build collection things desktop frequentl...	0	1	0	1
8	INFJ	sure good question distinction dependant perce...	0	1	1	1
9	INTP	position actually person various reason unfort...	0	1	0	0

<표 2> 최종 테이블

2-1-2. 모델링

위의 텍스트 분석을 통해 만들어진 단어들을 Bag Of Words 기법을 사용하여 피처 행렬로 만들어 주었다. 토큰 빈도가 게시글의 95% 이상 나타나는 단어는 무시하고 최대 피처의 개수는 1500 개로 한정하였다. 총 8675 개 행의 작성 게시글에서 1500 개의 단어들의 가중치를 부여하는 TF-IDF 기반의 벡터화를 사용하였다. TF-IDF 는 단어 개수를 그대로 세지 않고 모든 문서에 공통적으로 들어있는 단어의 경우 문서 구별 능력이 떨어진다고 보아 가중치를 부여하거나 축소하는 방법이다.

분류 모델으로는 LogisticRegression, XGBoost, RandomForest, DecisionTree, MultinomialNB 의 5 가지 모형으로 비교해보았다. 평가 지표로는 정확도(Accuracy)로 보았다. LogisticRegression 모형의 분석 결과 <E-I>의 정확도는 0.778, <S-N>의 정확도는 0.864, <T-F>의 정확도는 0.793, <J-P>의 정확도는 0.641 로 나타났다. XGBoost 모형의 분석 결과는 <E-I>의 정확도는 0.763, <S-N>의 정확도는 0.863, <T-F>의 정확도는 0.746, <J-P>의 정확도는 0.616 로 나타났다. RandomForest 모형의 분석 결과 <E-I>의 정확도는 0.769, <S-N>의 정확도는 0.862, <T-F>의 정확도는 0.740, <J-P>의 정확도는 0.607 로 나타났다. DecisionTree 모형의 분석 결과는 <E-I>의 정확도는 0.763, <S-N>의 정확도는 0.854, <T-F>의 정확도는 0.663, <J-P>의 정확도는 0.593 로 나타났다. 마지막으로 MultinomialNB 의 모형 분석 결과는 <E-I>의 정확도는 0.727, <S-N>의 정확도는 0.795, <T-F>의 정확도는 0.778, <J-P>의 정확도는 0.621 로 나타났다.

모델(정확도)	E-I	S-N	T-F	J-P
LogisticRegression	0.778	0.864	0.793	0.641
XGBoost	0.763	0.863	0.746	0.616
RandomForest	0.769	0.862	0.740	0.607
DecisionTree	0.763	0.854	0.663	0.593
MultinomialNB	0.727	0.795	0.778	0.621

<표 3> 모델 정확도 비교

<표 3>은 모델의 정확도를 비교한 것을 나타낸 표이다. 전체적으로 보았을 때 모든 피처에서 LogisticRegression 모델이 높은 정확도를 가지는 것을 볼 수 있다. 그래서 로지스틱 회귀 모형의 최적의 파라미터를 찾기 위해 GridSearchCV 를 진행하였다. C 값을 [0.001, 0.01, 0.1, 1, 10, 100], penalty 를 ['l1','l2']로 주어 비교한 결과 'E-I','S-N','T-F','J-P' 의 모든 열에서 C=1, penalty=l2 가 가장 최적의 파라미터로 도출되었다. LogisticRegression 모형의 기본 default parameter 값과 같으므로 <표 1>의 정확도가 최적의 파라미터로 설정한 결과이다.

로지스틱 회귀 분류 모형 학습 결과, 0.5 이상이면 1, 0.5 이하면 0 으로 대체하여 MBTI 의 유형을 예측할 수 있다.

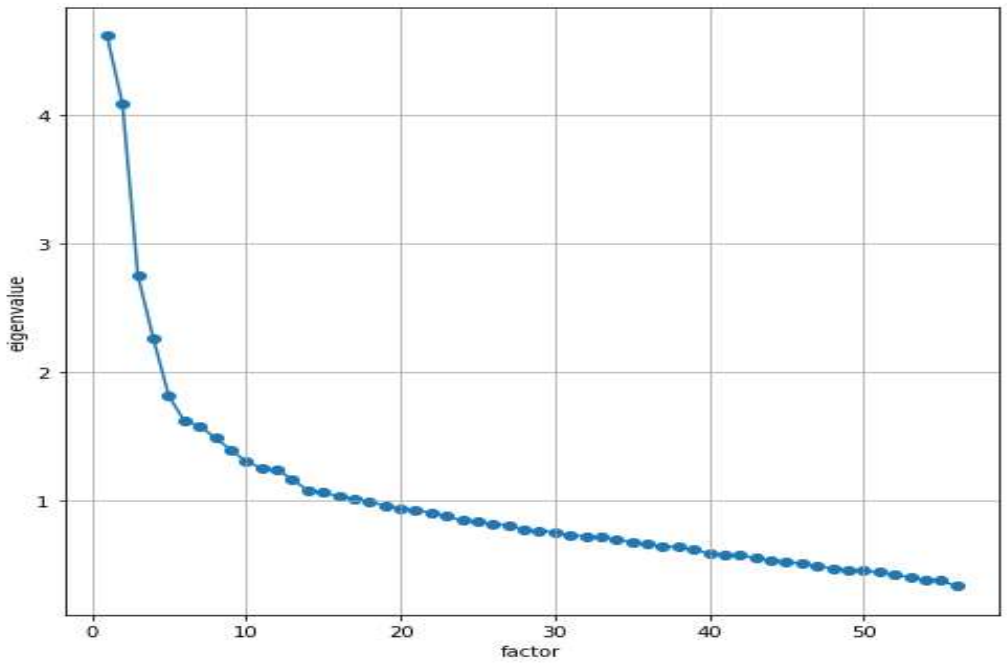
2-2. 두번째 데이터 – 설문조사: 성격유형, 취미와 흥미를 중심으로

설문조사 데이터는 5 점 리쿼트 척도로 음악 선호도, 영화 선호도, 취미와 흥미, 성격 유형, 건강 습관 등의 질문들로 이루어져 있다. <표 4>와 같이 본 주제에 맞게 취미와 흥미(31 번-63 번 항목), 성격 유형(76 번-133 번 항목)의 자료만을 추출하여 분석하였다. 먼저 성격 유형으로 요인 분석을 진행하여 MBTI 유형 형태로 바꾸고자 한다.

Daily events	Prioritising workload	Writing notes	Workaholism	Thinking ahead	Final judgement	Reliability	Keeping promises	Loss of interest	Friends versus money	...	Life struggles	Happiness in life	Energy levels	Small - big dogs	Personal
2.0	2.0	4.0	1.0	4.0	3.0	4.0	4.0	2.0	5.0	...	2.0	3.0	3.0	2.0	3
3.0	2.0	1.0	4.0	2.0	3.0	3.0	3.0	4.0	4.0	...	4.0	4.0	3.0	3.0	3
3.0	1.0	1.0	1.0	4.0	1.0	3.0	5.0	1.0	4.0	...	5.0	3.0	1.0	3.0	2
3.0	1.0	5.0	1.0	3.0	4.0	4.0	4.0	5.0	3.0	...	5.0	3.0	2.0	2.0	4
3.0	5.0	4.0	5.0	4.0	3.0	5.0	5.0	3.0	4.0	...	2.0	4.0	2.0	3.0	4

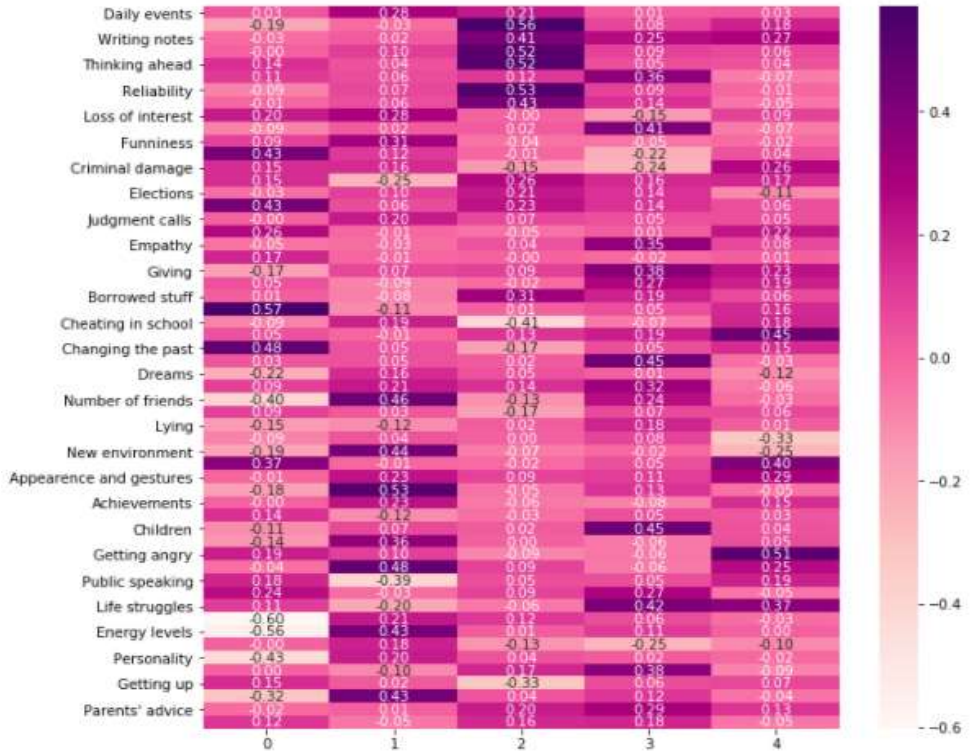
<표 4> 성격 유형의 설문 조사

설문자료 중 57 개의 성격 유형 항목을 통해 MBTI 유형을 파악하고자 한다. 먼저 성격 유형 항목들을 요인 분석을 통해 여러 변수를 하나의 변수로 만들었다. FactorAnalyzer 모듈을 사용하여 요인들의 고유값을 구한 후 요인의 개수를 그래프를 통해 정했다.



<그림 5> 고유값과 요인의 그래프

<그림 5>은 고유값과 요인의 개수를 각 y 축, x 축으로 하여 나타낸 그래프이다. 그래프를 보면 요인의 개수를 5 로 하는 것이 가장 적합해 보임을 알 수 있다. 요인의 개수를 5 로하고 rotation 을 varimax 로 지정하여 FactorAnalyzer 를 통해 요인분석을 다시 진행하였다.



<그림 6> 요인분석 히트맵

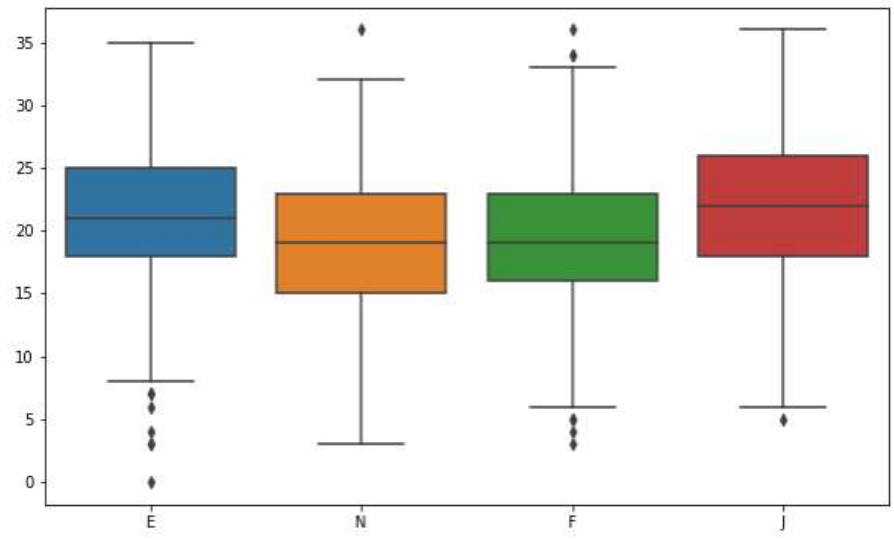
factor	0	1	2	3	4
0	Loneliness	Socializing	Prioritising workload	God	Getting angry
1	Changing the past	Knowing the right people	Reliability	Children	Health
2	Fake	Number of friends	Thinking ahead	Life struggles	Mood swings
3	Self-criticism	New environment	Workaholism	Friends versus money	Life struggles
4	Mood swings	Interests or hobbies	Keeping promises	Giving	Appearence and gestures
5	Hypochondria	Energy levels	Writing notes	Finding lost valuables	Writing notes
6	Unpopularity	Assertiveness	Borrowed stuff	Final judgement	Criminal damage
7	Loss of interest	Funniness	Decision making	Empathy	Knowing the right people
8	Getting angry	Loss of interest	Self-criticism	Charity	Giving

<표 5>는 요인분석 결과

<그림 6>는 요인분석 결과를 히트맵으로 나타낸 것이고 <표 5>는 요인별로 묶어 표로 나타낸 것이다. 사전 지식들과 검색들을 통해 해당 질문들이 MBTI 의 어떤 유형을 나타내는지 알 수 있었다. Factor1 에 해당하는 항목('Socializing', 'Knowing the right people', 'Number of friends', 'New environment','Energy levels', 'Funniness', 'Loss of interest', 'Interests or hobbies','Assertiveness')들은 외향형(Extraversion)을 묻는 질문으로 factor1 을 'E-I'로 볼 수 있다. Factor2 에 해당하는 항목('Writing notes', 'Self-criticism', 'Decision making', 'Borrowed stuff','Keeping promises', 'Thinking ahead', 'Reliability', 'Prioritising workload','Workaholism')들은 성실성(Conscientiousness)의 관련 질문들로 J 와 강한 상관성을 가져 Factor2 를 'J-P'로 볼 수 있다. Factor3 에 해당하는 항목('God', 'Life struggles','Friends versus money', 'Children', 'Finding lost valuables', 'Final judgement', 'Empathy', 'Charity', 'Giving')들은 개방성(Openness) 관련 질문들로 N 과 강한 상관성을 가진다고 하여 Factor3 을 'S-N'으로 볼 수 있다. Factor4 로 묶인 항목('Giving', 'Knowing the right people', 'Criminal damage', 'Writing notes', 'Appearence and gestures', 'Life struggles', 'Mood swings', 'Health','Getting angry')들은 수용성(Agreeableness)관련 질문들로 F 와 강한 상관성을 가져 'T-F'로 볼 수 있다. Factor0 은 불안정성(Neuroticism)관련 항목으로 묶인 것이지만 본 분석에서는 다루지 않기로 하였다.

요인분석으로 묶인 항목들의 점수를 합하여 [E,N,F,J]의 칼럼으로 나타내었다. 점수의 합이 높을수록 해당 열에 더 가까워짐을 뜻하고 점수의 합이 낮을수록 I,S,T,P 에 가까워짐을 뜻한다.

```
********* E
평균:  20.9948586118251941
최댓값:  35
최솟값:  0
최빈값:  [20]
********* N
평균:  18.985861182519281
최댓값:  36
최솟값:  3
최빈값:  [16]
********* F
평균:  19.083547557840618
최댓값:  36
최솟값:  3
최빈값:  [18]
********* J
평균:  21.767352185089976
최댓값:  36
최솟값:  5
최빈값:  [22]
```



<그림 7> 요인으로 묶인 점수의 합산 결과

<그림 7>은 E, N, F, J 요인들의 점수 합산 결과이다. 평균과 빈도 값이 비슷하게 나오는 것을 알 수 있다. 따라서 평균을 기준으로 평균보다 높은 것을 1 로 평균보다 낮은 것을 0 으로 대체하여 MBTI 의 유형을 파악할 수 있다.

id	History	Psychology	Politics	Mathematics	Physics	Internet	PC	Economy Management	Biology	Chemistry	...	Shopping	Science and technology	Theatre	Fun with friends	Adrenal
0	1.0	5.0	1.0	3.0	3.0	5.0	3.0	5.0	3.0	3.0	...	4.0	4.0	2.0	5.0	
1	1.0	3.0	4.0	5.0	2.0	4.0	4.0	5.0	1.0	1.0	...	3.0	3.0	2.0	4.0	
2	1.0	2.0	1.0	5.0	2.0	4.0	2.0	4.0	1.0	1.0	...	4.0	2.0	5.0	5.0	
3	3.0	2.0	3.0	2.0	2.0	2.0	2.0	2.0	3.0	3.0	...	3.0	3.0	2.0	4.0	
4	5.0	3.0	4.0	2.0	3.0	4.0	4.0	1.0	4.0	4.0	...	2.0	3.0	1.0	3.0	

<표 6> 취미와 MBTI 유형 결과

<표 6>은 위에서 요인분석으로 하여 나온 MBTI 유형의 분석결과와 취미와 join 시킨 표이다. 이를 통해 MBTI 의 유형을 통해 어떠한 취미와 흥미를 즐겨하는지를 파악할 수 있다.

2-3. 아이템 기반/ 잠재 요인 협업 필터링 MBTI 기반 취미 추천 : 예시를 바탕으로

게시글을 통해 MBTI 를 예측하고, 예측된 MBTI 유형을 기반으로 취미를 추천하는 시스템을 만들고자 한다. 취미 추천 시스템으로 사용한 추천 알고리즘은 '아이템 기반 협업 필터링'과 '잠재 요인 협업 필터링'이 있다. '아이템 기반 협업 필터링'은 사용자들의 평가 척도가 유사한 아이템을 기준으로 추천이 되는 알고리즘이다. 즉, MBTI 의 유형이 같은 사람들 중에 취미 평점이 비슷한 사람과 유사한 취미를 추천해 주는 방식이다. '잠재 요인 협업 필터링'은 사용자-아이템 평점 행렬 속에 숨어 있는 잠재 요인을 추출해 추천 예측을 할 수 있게 하는 기법이다. 행렬 분해에는 SVD 가 자주 사용되지만 사용자-아이템 행렬에는 사용자가 평점을 매기지 않는 데이터도 존재하기 때문에 주로 SGD 나 ALS 기반의 행렬 분해를 이용한다. 따라서, 게시글로부터 예측된 MBTI 의 사람이 취미 설문에 빈 값이 존재하여도 취미를 추천해 줄 수 있다.

```
ex_post = """Through conversations with customers, Fiveave found that customers were looking for a delicate gripping tool like a human
Piab developers realized that instead of imitating the human hand, they had to develop an inexpensive alternative that would work in
The three-finger gripper incorporates a vacuum cavity that adjusts and controls the gripping force according to the applied vacuum. Th
This has proven to be the perfect size to show off the fruits of the hard work that Piab developers have put into their passion for th
Unwrapped, hollow chocolate eggs, which are a bit weak in color but very brittle, were used to demonstrate the benefits and versatilit
PiSOFTGRIP®, which enables the automation of processes that have not yet been automated, has attracted a lot of interest at the show
In addition to its low cost, its light weight allows it to be lifted lightly when mounted on a robot arm. With a smooth grip for the f
However, our demonstration has led to the possibility that if these technologies can be offered in a variety of sizes, they can also b
Driven by this discussion, our developers have already started designing larger and smaller versions of piSOFTGRIP® with the ultimate
```

<그림 8> 게시글 예시

<그림 8>은 블로그(https://www.piab.com/ko-KR/blog/a-gripping-story/)에서 가져온 게시글이다. 이 게시글로 위의 텍스트 분석 과정을 거쳐 MBTI 를 알아내고, 같은 유형을 가진 사람 중 비슷한 평점을 매긴 사람의 취미를 추천하는 과정을 보이겠다.

게시글 예시를 위의 텍스트 분석 과정을 통해 불용어 처리, 토큰화, 피처 벡터 추출 등을 마치고, 로지스틱 회귀 모형을 통해 MBTI 를 파악하였다.

	mbti_E	mbti_N	mbti_F	mbti_J
0	0	1	0	0

<표 7> 게시글 예시를 기반으로 MBTI 예측 결과

<표 7>은 게시글 예시를 기반으로 나온 MBTI 예측 결과 이며, 해당 게시글의 작성자는 모형 학습 결과 'INTP'의 MBTI 의 유형을 가지는 것으로 확인되었다. 두번째 설문 데이터에서 게시글 작성자와 같은 MBTI 의 유형을 가진 사람들을 추출하였다. 같은 MBTI 의 유형을 가진 사용자 사이에서 아이템 기반 협업 필터링을 진행하여 취미를 추천 받고자 한다.

```
Chemistry                1.000000
Biology                  0.951861
Medicine                 0.932922
Physics                  0.854620
Active sport             0.840531
Fun with friends         0.835040
Internet                 0.829930
Mathematics              0.828035
Foreign languages        0.818322
Science and technology   0.815615
Name: Chemistry, dtype: float64
```

```
Internet                 0.908361
Cars                     0.907159
Science and technology   0.888895
Economy Management       0.879474
Name: PC, dtype: float64
```

<그림 9> 아이템 기반 취미 유사도 결과

먼저 코사인 유사도를 기반으로 아이템 기반 취미 유사도를 파악해보았다. <그림 9>는 아이템 기반 취미 유사도 결과로 왼쪽은 Chemistry 와 유사도가 높은 순서대로 나온 결과이다. Chemistry 를 좋아하는 사람은 Biology, Medicine, Physics 순으로 취미가 추천될 수 있다. 오른쪽은 'PC'와 유사도가 높은 것들로 Internet, Cars, Science and technology 순으로 추천될 수 있다.

임의로 해당 게시물을 작성한 사람이 'Science and technology'를 4 점, 'Bology'를 5 점, 'Foreign languages'를 2 점으로 작성하고 나머지 취미 중 추천을 받으려고 한다고 가정하자. 이를 통해 점수를 부여한 취미에 대해서 예측 성능 평가 MSE(Mean Squared Error)를 구할 수 있다. 아이템 기반 모든 최근접 이웃 MSE 는 1.83 으로 나타났다.

```python
def rating_topsim(ratings_arr, item_sim_arr, n=25):

    pred= np.zeros(ratings_arr.shape)

    for col in range(ratings_arr.shape[1]):
        top_n_items = [np.argsort(item_sim_arr[:,col])[:-n-1:-1]]
        for row in range(ratings_arr.shape[0]):
            pred[row,col] = item_sim_arr[col, :][top_n_items].dot(ratings_arr[row,:][top_n_items].T)
            pred[row,col] /= np.sum(np.abs(item_sim_arr[col, :][top_n_items]))

    return pred
```

<그림 10> rating_topsim 함수

<그림 10>은 rating_topsim 함수로 rating_topsim 을 사용하여 Top-N 유사도를 갖는 취미 유사도 벡터만 예측값을 계산하는데 사용한다. 실제 평점과의 MSE 를 비교하니 1.78 로 기존보다 작은 수로 줄어든 것을 볼 수 있다. 게시글 예시 작성자가 평점을 매긴 것을 제외하고 최종적으로 작성자에게 취미를 추천 결과는 <표 8>이다.

	pred_score
Law	0.440902
Fun with friends	0.440044
Countryside, outdoors	0.438630
Gardening	0.438570
Dancing	0.438314
Economy Management	0.438132
Passive sport	0.437540
Pets	0.433292
Geography	0.431345
Mathematics	0.431115

<표 8> 아이템 기반 협업 필터링 추천 결과

해당 게시글 작성자는 Law, Fun with friends, Gardening 순으로 취미 또는 흥미가 추천되었다.

이번에는 행렬 분해를 이용한 잠재 요인 협업 필터링으로 취미를 추천해보고자 한다.

```
def matrix_factorization(R, K, steps=200, learning_rate=0.01, r_lambda = 0.01):
    num_users, num_items = R.shape

    np.random.seed(1213)
    P = np.random.normal(scale=1./K, size=(num_users,K))
    Q = np.random.normal(scale=1./K, size=(num_items,K))

    prev_rmse = 10000
    break_count = 0

    non_zeros = [(i,j,R[i,j]) for i in range(num_users) for j in range(num_items) if R[i,j]>0]

    # SGD기법 활용
    for step in range(steps):
        for i, j, r in non_zeros:
            eij = r-np.dot(P[i,:],Q[j,:].T)
            # SGD 업데이트
            P[i,:] = P[i,:] + learning_rate*(eij *Q[j,:] - r_lambda*P[i,:])
            Q[j,:] = Q[j,:] + learning_rate*(eij *P[i,:] - r_lambda*Q[j,:])

        rmse = get_rmse(R, P, Q, non_zeros)
        if (step % 10 )== 0:
            print('### iteration step: ', step, 'rmse: ', rmse)

    return P,Q
```

<그림 11> 확률적 경사 하강법을 활용한 행렬 분해 함수

<그림 11>는 확률적 경사 하강법을 활용하여 행렬 분해 함수이다. 이렇게 만들어진 사용자-취미 평점 행렬을 통해 개인 맞춤형 취미를 추천해 볼 수 있다. 잠재 요인 협업 필터링 분석 결과는 <표 9>와 같다.

	pred_score
Biology	4.967848
Medicine	4.068062
Chemistry	3.993577
Science and technology	3.981965
Fun with friends	3.976786
Active sport	3.795098
Theatre	3.257919
History	3.228271
Adrenaline sports	3.215396
Countryside, outdoors	3.070531

<표 9> 잠재 요인 협업 필터링 추천 결과

잠재 요인 협업 필터링 추천 결과 아이템 기반 협업 필터링 추천 결과와는 다른 결과들이 나왔다. Biology, Madicine, Chemisty 순으로 추천이 된 것을 볼 수 있다.

3. 결론

본 분석은 Kaggle 에 있는 PersonalityCafe forum 의 MBTI 데이터와 FSEV UK 대학의 설문 자료 데이터를 바탕으로 작성 게시글을 분석하여 취미를 추천해주는 시스템을 보고자 하였다. MBTI 데이터의 텍스트를 분석하기 위해 텍스트 전처리를 해주는 함수(cleansing_text)를 만들었다. 해당 함수로 기호, url, 구두점, 숫자 등을 제거하고 소문자로 변환시켰다. NLTK 와 Lemmatize 모듈을 사용해 토큰화와 표제어를 추출하고, CountVectorizer 와 TF-IDF 기반의 Bag Of Words 기법을 사용하여 총 8675 개 행의 작성 게시글에서 1500 개의 단어들의 가중치를 부여하는 피처를 추출하였다.

MBTI 데이터에서 MBTI 의 16 가지 유형들은 INFP, INFJ, INTP, INTJ 순으로 많은 비율을 차지하며 ESTJ 의 유형이 가장 적은 비율을 차지함을 볼 수 있었다. 보다 정확한 성향 예측을 위해 16 가지 유형을 .에너지 방향(E-I), 인식기능(S-N), 판단 기능(T-F), 생활 양식(J-P)의 열을 생성하여 이진 분류의 조합들로 보았다.

모델 학습에는 LogisticRegression, XGBoost, RandomForest, DecisionTree, MultinomialNB 의 총 5 개의 모델을 사용하여 정확도를 비교해보았다. 모델 학습 결과 LogisticRegression 이 'E-I'는 0.778, 'S-N'은 0.864, 'T-F'는 0.793, 'J-P'는 0.641 의 정확도를 보이며 다른 모델보다 높은 정확성을 가짐을 확인하였다. 그 후 적절한 파라미터 조정을 진행하여 정확도를 높이기 위해 GridSearchCV 를 이용하여 C 와 penalty 를 조정하였고 C=1, penalty= l2 로 default 값과 같은 결과를 보였다. Logistic 모형으로 예측한 결과를 0.5 로 기준으로 하여 높으면 1, 낮으면 0 으로 값을 놓았다. 즉, MBTI 의 유형을 확인할 수 있었다.

두번째 설문조사 데이터에서는 요인분석을 주로 진행하였다. FactorAnalyzer 모듈을 사용하여 57 개의 성향에 대한 질문들이 5 개의 요인으로 묶이며 그 중 영향도가 높은 것들 45 개만 추출하였다. 5 개의 요인 중 하나의 요인은 해당 분석과 관련 없는 것으로 파악하여 제외한 후 4 개의 요인들(36 개의 항목)로 분석을 진행하였다. 요인분석으로 묶인 항목들의 점수를 합하여 [E,N,F,J]의 칼럼으로 나타내었다. 점수의 합이 높을수록 해당 열에 더 가까워짐을 뜻하고 점수의 합이 낮을수록 I,S,T,P 에 가까워짐을 뜻한다. 점수의 평균과 최빈값을 비교하여 차이가 많이 없음을 확인하였고, 평균보다 높으면 1 값을 낮으면 0 값으로 하여 MBTI 유형을 파악할 수 있었다.

이를 기반으로 아이템 기반 협업 필터링과 잠재 요인 협업 필터링을 활용하여 취미를 추천해주는 과정을 보였다. 한 예시로 어떠한 사람이 게시글을 작성했을 때 Logistic 회귀 모형으로 MBTI 를 'INTP'로 예측하고 , 'INTP' 유형의 사람들과 비교하여 취미를 추천해 주는 시스템이다. 아이템 기반 협업 필터링은 코사인 유사도를 구해 작성한 평점을 바탕으로 유사한 취미를 추천해주는 방식으로 작성자에게 Law, Fun with friends, Gardening 순으로 추천되었고, 잠재 요인 협업 필터링은 행렬을 분해하여 SGD 업데이터를 통해 추천해주는 방식으로 Biology, Madicine, Chemisty 순으로 추천되었다.

4. 참고 문헌

권철민, *파이썬 머신러닝 완벽 가이드*, 위키북스

.https://personalityjunkie.com/09/openness-myers-briggs-mbti-intuition-big-five-iq-correlations/

5. 소감

통계 데이터 사이언스 수업을 들으며 이론부터 실습까지 많은 것을 배울 수 있었다. 금융 데이터 분석, 머신러닝에 필요한 지식들과 분석법들을 익히며 한층 더 성장할 수 있었다. 또한 이번 개인 과제로 MBTI 관련한 흥미있는 주제로 추천 시스템까지 다루어 보면서 평소에 해보고 싶은 분석을 해볼 수 있어 좋았다. 특히, 예시를 통해 실제 게시글로 MBTI 가 예측되고 그것으로 취미를 추천해주는 것이 신기하면서 재미있었다. 나아가 한글로 작성된 게시물을 바탕으로 추천 시스템이 이루어진다면 더욱 흥미롭고 화제가 될 것이라고 생각이 든다. 아쉬운 점이 있다면, MBTI 의 관련 데이터가 부족하여 찾는 것에 많은 시간을 소요하였고 텍스트 분석의 강점으로 작용하는 SVM(Support Vector Machine)을 활용해보지 못했다는 것이다.