머신러닝을 이용한 MBTI 예측 모델 개발

팀 뽀

포트폴리오 목차

참가팀원 ▼

프로젝트 배경 :: 우리는 왜 MBTI에 열광하는가?

MBTI, 대중화의 시작 → **왜 MBTI 인가?** → **우리의 목표**

우리의 목표 :: MBTI, 바르게 알자!

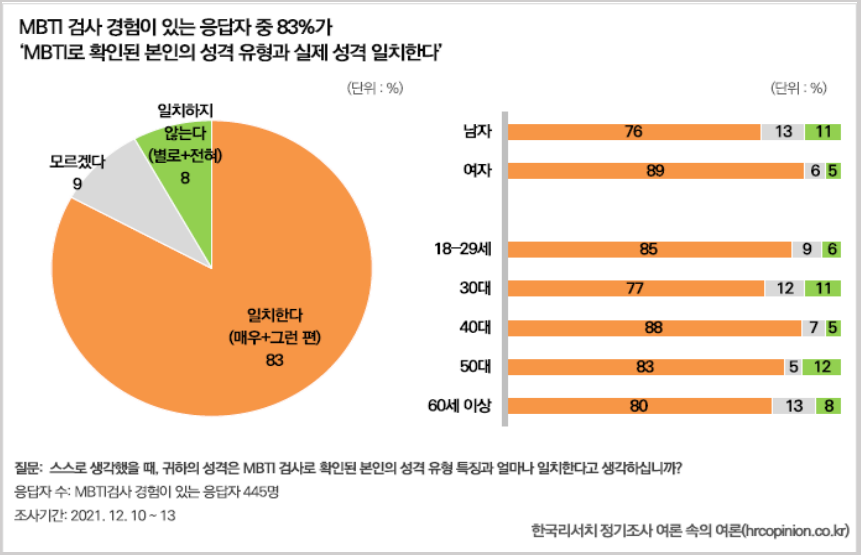
- 전문가들은 대중이 자아탐구에 관심을 갖는 것 자체는 긍정적인 현상이라고 보고 있으나, **특정 유형을 일반화해 배제하거나 잘못된 확증편향으로 이어지는 경우**가 발생하고 있어 이에 대한 **해결책**이 필요한 시점이다.
- 급변하는 사회환경에서 '내'가 누구인지 정의 내릴 수 있는 MBTI 성격유형검사는 우리에게 안정감을 주지만, 어디까지나 하나의 도구이자 성격 지표일 뿐, **올바르게 이해하고 사용할 필요가 있다**.
- 이에 **팀 뽀**는 MBTI가 대중에게 올바르게 받아들여지고 사용될 수 있도록 **머신러닝을 이용한 MBTI 예측 모델 개발 서비스**를 기획하게 되었다.

MBTI, 대중화의 시작

- MBTI 대중화의 시작은 2020년 6월 MBC '놀면 뭐하니' 방송 이후 ,**네이버 데이터랩과 구글 트렌드에서의 'MBTI'에 대한 검색량이 크게 증가했음을 확인할 수 있었음.**

왜 MBTI 인가?

- **MBTI 이전에도 혈액형 성격론과 별자리 성격론으로 대표**되는 타인과 자신을 평가, 규정하는 자아탐구 열풍은 있었지만, 과학적인 근거가 부족한 성격 검사에 불과하다는 인식으로 오래 지속되지 못하였다.
- 하지만 MBTI는 이론적 근거가 앞선 자아탐구 열풍과 함께한 성격검사들에 비해 탄탄했고, **코로나 19로 인한 언택트 문화의 확산**과 맞물려 MBTI는 언택트 시대에서 중요한 자기소개 수단으로 자리잡게 되었다.



프로젝트 배경 :: 우리는 왜 MBTI에 열광하는가?

MBTI, 대중화의 시작

왜 MBTI 인가?

우리의 목표

우리의 목표 :: MBTI, 바르게 알자!

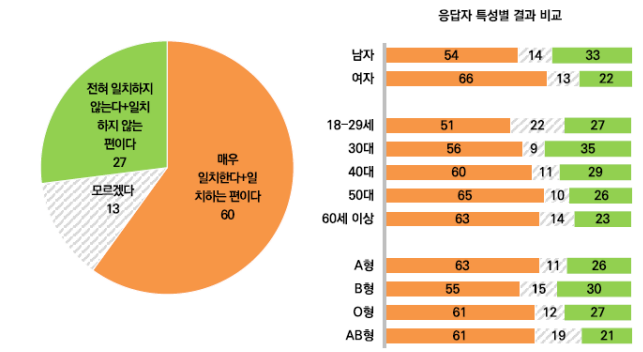
- 전문가들은 대중이 자아탐구에 관심을 갖는 것 자체는 긍정적인 현상이라고 보고 있으나, **특정 유형을 일반화해 배제하거나 잘못된 확증편향으로 이어지는 경우**가 발생하고 있어 이에 대한 **해결책**이 필요한 시점이다.
- 급변하는 사회환경에서 '내'가 누구인지 정의 내릴 수 있는 MBTI 성격유형검사는 우리에게 안정감을 주지만, 어디까지나 하나의 도구이자 성격 지표일 뿐, **올바르게 이해하고 사용할 필요가 있다.**
- 이에 **팀 뽀는** MBTI가 대중에게 올바르게 받아들여지고 사용될 수 있도록 **머신러닝을 이용한 MBTI 예측 모델 개발 서비스**를 기획하게 되었다.

왜 MBTI 인가?

- MBTI 이전에도 **혈액형 성격론**과 **별자리 성격론**으로 **대표**되는 타인과 자신을 평가, 규정하는 자아탐구 열풍은 있었지만, 과학적인 근거가 부족한 성격검사에 불과하다는 인식으로 오래 지속되지 못하였다.
- 하지만 MBTI는 이론적 근거가 앞선 자아탐구 열풍과 함께한 성격검사들에 비해 탄탄했고, **코로나 19로 인한 언택트 문화의 확산**과 맞물려 **MBTI는 언택트 시대에서 중요한 자기소개 수단으로 자리잡게 되었다.**

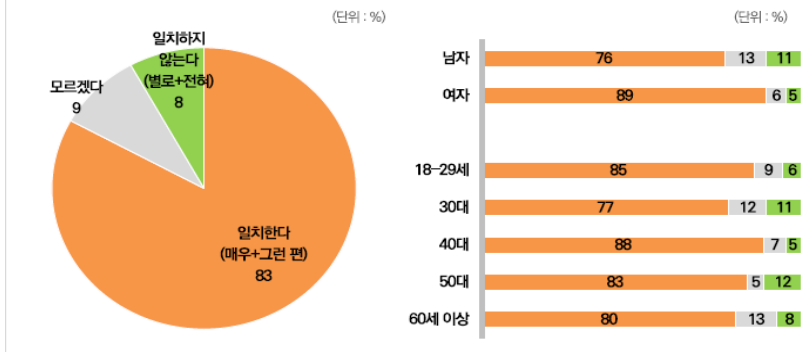
데이터로 팩트체크 :: 혈액형 성격론

나의 성격이 일반적으로 알려져 있는 혈액형 별 성격과 일치한다 60%
20대와 30대에서도 일치한다는 응답 과반 이상



질문: 귀하의 혈액형과 성격을 생각했을 때, 귀하의 성격은 일반적으로 알려져 있는 혈액형 별 성격 특징과 얼마나 일치한다고 생각하십니까?
응답자 수 : 1,000명
조사기간 : 2021. 10. 15. ~ 18
한국리서치 정기조사 여론 속의 여론(hrcopinion.co.kr)

MBTI 검사 경험이 있는 응답자 중 83%가
'MBTI로 확인된 본인의 성격 유형과 실제 성격 일치한다'



질문: 스스로 생각했을 때, 귀하의 성격은 MBTI 검사로 확인된 본인의 성격 유형 특징과 얼마나 일치한다고 생각하십니까?
응답자 수: MBTI검사 경험이 있는 응답자 445명
조사기간: 2021. 12. 10 ~ 13
한국리서치 정기조사 여론 속의 여론(hrcopinion.co.kr)

팀 구성 및 역할

팀장 오병진

- 데이터 리서치 및 수집
- 학습용 데이터 수집 및 전처리
- 예측 모델 정확도 테스트 및 개선
- 예측 결과 분석

팀원 권묘선

- 예측용 데이터 수집 및 전처리
- R을 이용해 수집 자료 빈도표 시각화
- 예측 결과 분석 및 시각화

팀원 김아형

- 예측용 데이터 크롤링 및 수집
- 데이터 전처리 및 ML학습
- 분석 내용 시각화

분석과정 및 방법 :: 프로세스 개요

참가팀원 ▼



데이터 수집

- 최적 학습 모델 탐구
- 최적 모델에 대한 하이퍼 파라미터 설정

데이터셋 구축

- 데이터 이상치, 결측치 제거
- 데이터셋 특성 확인(시각화)
- 데이터셋 최적화 진행

텍스트 클렌징

- Html 태그 제거
- 특수문자 제거
- 알파벳 제외한 모든 기호 제거
- 1~2자의 짧은 글자 제거

텍스트 전처리

- 텍스트 자연어 처리 (Tokenize, Lemmatization)
- Text Vectorization : Tf-IDF

예측모델 학습

- 최적 학습 모델 탐구
- 최적 모델에 대한 하이퍼-파라미터 설정

분석과정 및 방법 :: 데이터 수집

참가팀원 ▼

데이터 선정

학습용 데이터 수집

- Reddit의 MBTI 관련 subreddit 유저 게시글 데이터(post, comment)

ESTJ	ESTP	ISTJ	ISTP
ESFJ	ESFP	ISFJ	ISFP
ENFP	ENFJ	INFP	INFJ
ENTP	ENTJ	INTP	INTJ

(출처 : https://www.reddit.com/r/****/)

출처 및 수집방법

Reddit API 이용
(출처: reddit.com/dev/api/)

예측 및 분석용 데이터 수집

- Reddit의 별자리 관련 subreddit 유저 게시글 데이터(post, comment)

양자리	황소자리	쌍둥이자리	게자리
사자자리	처녀자리	천칭자리	전갈자리
사수자리	염소자리	물병자리	물고기자리

(출처 : https://www.reddit.com/r/****/)

- www.webnovel.com의 장르별(12개) 소설 총 1882작품에 대한 프롤로그 텍스트

출처 및 수집방법

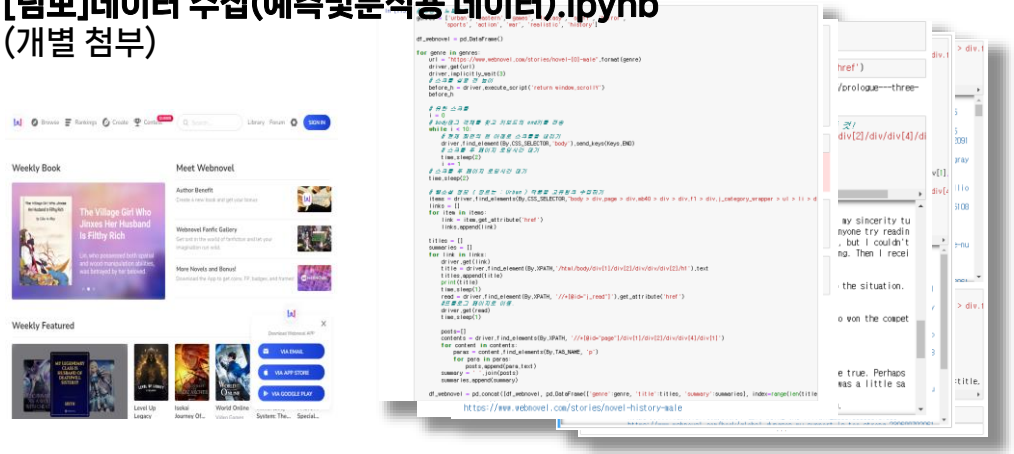
Reddit API 이용
(출처: reddit.com/dev/api/)

Selenium
동적 크롤링 사용
(출처:webnovel.com/)

데이터 수집

[팀뽀]데이터 수집(학습용데이터).ipynb
(개별 첨부)

[팀뽀]데이터 수집(예측및분석용 데이터).ipynb
(개별 첨부)



Raw 수집 데이터

- Reddit에서 mbti별 총 4377명에 대한 post를 수집했음.

```
array(['enfj', 'enfp', 'entj', 'entp', 'esfj', 'esfp', 'estj', 'estp',
       'infj', 'infp', 'intj', 'intp', 'isfj', 'isfp', 'istj', 'istp'],
      dtype=object)
```

	type	authors	posts
0	enfp	astridchan__	Hi ENFPs, have you ever experienced ending up ...
1	enfp	Fancy-Assignment-467	Hey guys, quick question: what do most guys in...
2	enfp	FutureCookies	A GIRL SPERM 😭😭😭 honestly right, this is a les...
3	enfp	CosmicJoke5067	Self imposed standards + Fi is a good thought....
4	enfp	foxinator2_0	it'd be a fun death :D|||why are people just s...

(4377, 3)

- 각 행의 posts 에는 각 유저별로 작성된 게시글 100개가 '|||'를 구분자로 하나의 문자열로 합쳐져 있는 형태이다.

```
In [29]: data_set.posts[0]
Out[29]: 'Hi ENFPs, have you ever experienced ending up being someone' s MPDG? I would like to believe that sometimes it wasn' t unintentional w
hen it happens like I am so deadass wanting to make emotionally unexpressive guys into someone who is able to feel but when it finally
happens - I start being resentful as if the only reason why he became a bit expressive was because I 'coaxed' him into doing that and
it no longer feel genuine. It' s a bit messed up because that was originally the main goal but once you' re in that situation and had a
time to reflect on every action you did to get to the point - you' ll realize that you should never have to ask someone to change.|||No
t a man but I brought this up with my fiance. We met online and I' ve been crushing on him a few months prior to me messaging him. Lolo
lol. I just laid the groundwork and let things unfold. I was the one to confess first as well. Well, basically, he told me that he' s t
hankful I did and taking such initiative is super attractive to him xD|||It' s hard because I have the tendency to blame people whenever I get triggered but lately I' m owning up and trying to acknowledge the fact that I' m responsible for my own actions. \n\nAnd that m
y BPD is not a free pass for hurting others. I cannot count in a single hand anymore how many times I' ve lashed out on people and push
ed them to their limits but lately whenever I have calmed down, I am trying to see the whole situation in a more logical perspective an
d own up my share of mistakes and responsibility.\n\nI think being self-aware and working towards making yourself better is actually a
good thing because you knew these traits are things you shouldn' t just "accept because they are who we are" but we should actually l
earn to fully accept them but at the same time work on them.|||I can hold down a job or friendship for a long time. Although, I would u
sually get this looming fear that friends would leave me.\n\nWhen it comes to romantic relationships, that' s where I' m really bad. I
would lash out at my partner and would usually split.\n\nI am also the worst person when it comes to holding down my temper when it com
es to family, so I would usually get into a fight with my dad.\n\nI have learned to equate my self-worth and sense of identity with my
job because it' s the only thing I find that is consistent in my life.|||Hey, the mere fact that you thought about this means you' re s
elf-aware. \n\nI believe you' re a good person or else this thought won' t cross your mind. There are days when it feels like you' re l
```

시각화 전 데이터 전처리 진행

- 유저의 posts가 null인 행 삭제

```
# posts 가 null 인행 삭제.
data_set[data_set['posts'].isnull()]
for i in data_set[data_set['posts'].isnull()].index:
    data_set.drop(index=[i], inplace=True)
```

- 한 유저가 복수의 mbti 갖는 경우 삭제

```
# 1명의 유저가 여러개의 MBTI를 가지고 있는 경우 확인
stand = authors['authors'].value_counts() >= 2
```

```
# 1명의 유저당 1개의 mbti만 갖고있도록 처리
stand[stand==True].index
for user in stand[stand==True].index:
    delete = authors[authors['authors']==user]
    authors.drop(delete.index, inplace=True)
```

- 기본 전처리 된 데이터

	type	authors	posts
0	enfp	astridchan__	Hi ENFPs, have you ever experienced ending up ...
1	enfp	Fancy-Assignment-467	Hey guys, quick question: what do most guys in...
2	enfp	FutureCookies	A GIRL SPERM 😭😭😭 honestly right, this is a les...
3	enfp	CosmicJoke5067	Self imposed standards + Fi is a good thought....
4	enfp	foxinator2_0	it'd be a fun death :D|||why are people just s...

- 위 데이터에서 ML에 사용할 두 열만 추출

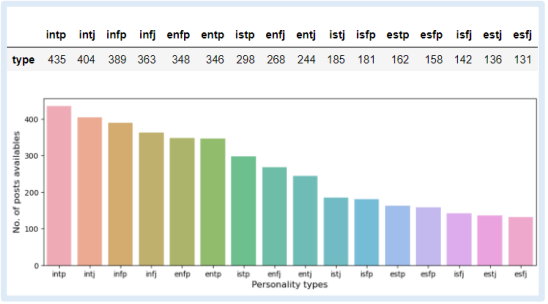
	type	posts
0	enfp	Hi ENFPs, have you ever experienced ending up ...
1	enfp	Hey guys, quick question: what do most guys in...
2	enfp	A GIRL SPERM 😭😭😭 honestly right, this is a les...

- 1차 전처리 결과 : (4377, 3) → (4190, 2)

시각화 진행

- EDA : 탐색적 실측 데이터 분석 진행
- 1차 전처리 된 결과에서 MBTI 별로 분포 시각화

	intp	intj	infp	infj	enfp	entp	istp	enfj	entj	istj	isfp	estp	esfp	isfj	estj	esfj
type	435	404	389	363	348	346	298	268	244	185	181	162	158	142	136	131

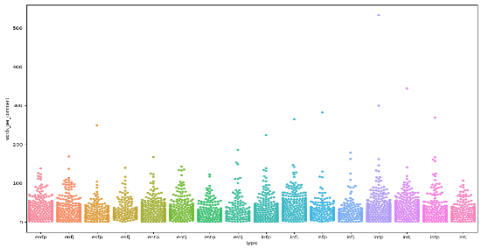


- 개별 글(포스트, 댓글)을 ||| 을 기준으로 분류

```
# 일 함수를 이용해 한 글 당 몇개의 단어가 있는지 확인!
def var_row(row):
    l = []
    for i in row.split('|||'):
        l.append(len(i.split()))
    return np.var(l)

# 1개의 글당 몇개의 단어가 있는지 확인!
df['words_per_comment'] = df['posts'].apply(lambda x: len(x.split()))/100
plt.figure(figsize=(15,10))
sns.swarmplot(x="type", y="words_per_comment", data=df)
```

- 파생변수 생성 : 개별 글(포스트, 댓글) 당 단어 수 'words_per_comment"



참가팀원 ▼



1차 전처리 데이터셋

- Reddit에서 mbti별 총 4190명에 대한 post를 수집하여 진행

- 텍스트 클렌징 전 데이터셋

	type	posts
0	enfp	Hi ENFPs, have you ever experienced ending up ...
1	enfp	Hey guys, quick question: what do most guys in...
2	enfp	A GIRL SPERM 😱😱😱 honestly right, this is a les...
3	enfp	Self imposed standards + Fi is a good thought....
4	enfp	it'd be a fun death :D|||why are people just s...

- 텍스트 클렌징 후 데이터셋

	type	posts
0	enfp	hi enfps have you ever experienced ending up ...
1	enfp	hey guys quick question what do most guys in...
2	enfp	a girl sperm honestly right this is a les...
3	enfp	self imposed standards fi is a good thought ...
4	enfp	it d be a fun death d why are people just s...

- 텍스트 클렌징 된 데이터셋 스플릿 진행(테스트 사이즈를 0.25로 설정)

```
# 데이터 분할!
train_data,test_data=train_test_split(data,test_size=0.25,random_state=42,stratify=data.type)
```

- 텍스트 전처리(자연어 처리) 진행 (초기 진행 시 max_features=5000으로 설정하였음)

```
vectorizer=TfidfVectorizer( max_features=5000,stop_words='english',tokenizer=Lemmatizer())
vectorizer.fit(train_data.posts)
```

```
TfidfVectorizer(max_features=5000, stop_words='english',
                tokenizer=<__main__.Lemmatizer object at 0x000001BE2BBD29A0>)
```

```
train_post=vectorizer.transform(train_data.posts).toarray()
test_post=vectorizer.transform(test_data.posts).toarray()
```

- 트레인 데이터셋(글 정보)

```
train_post
array([[0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.04251327, ..., 0.        , 0.        ,
        0.        ],
       ...,
       [0.        , 0.0108102 , 0.        , ..., 0.02513937, 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ]])
```

(3142, 5000)

- 테스트 데이터셋(글 정보)

```
test_post
array([[0.        , 0.01741397, 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.01402627, 0.        , ..., 0.        , 0.        ,
        0.        ],
       ...,
       [0.06025536, 0.        , 0.        , ..., 0.        , 0.        ,
        0.07723346],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.0069016 , 0.        , ..., 0.        , 0.        ,
        0.        ]])
```

(1048, 5000)

텍스트 클렌징 진행

- 링크에 해당하는 문자열 삭제

```
# 링크에 해당하는 내용 삭제
sentence=re.sub('https?://[^₩s<>"]+|www₩.[^₩s<>"]+',' ',sentence)
```

- 글자를 제외한 특수기호 문자열 삭제

```
# 글자를 제외한 특수기호 삭제
sentence=re.sub('[^0-9a-z]',' ',sentence)
```

텍스트 전처리 진행

- TF-IDF 벡터화 진행

```
vectorizer=TfidfVectorizer(max_features=5000,stop_words='english',tokenizer=Lemmatizer())
vectorizer.fit(train_data.posts)
```

```
TfidfVectorizer(max_features=5000, stop_words='english',
                tokenizer=<__main__.Lemmatizer object at 0x000002892E296940>)
```

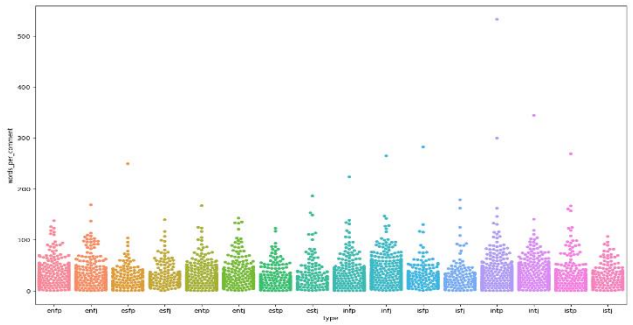
- TfidfVectorizer 사용하며, tokenizer 파라미터에 Leammatizer() 클래스를 넘겨줌으로서 단어들에 대한 Lemmatizing(표제어추출)도 동시에 진행
- Lemmatizer 클래스 설정 시, def __call__ 함수에 3개 이상의 알파벳으로 이루어졌을 때에만 토큰(피처)로 활용하도록 클렌징 조건을 걸어두었음.

```
class Lemmatizer(object):
    def __init__(self):
        self.lemmatizer = WordNetLemmatizer()
    def __call__(self, sentence): # 단어가 3개의 알파벳 이상으로 이루어졌을 때에만 토큰(피처)로 활용하도록!
        return [self.lemmatizer.lemmatize(word) for word in sentence.split() if len(word)>2]
```

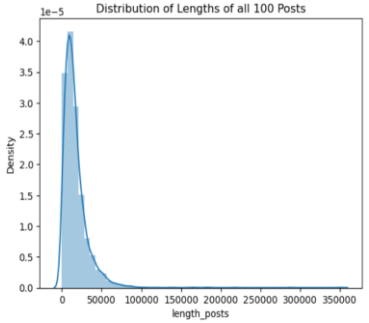
참가팀원 ▼

데이터셋 전처리

- 파생변수 생성 : 개별 글(포스트, 댓글) 당 단어 수
  'words_per_comment" 변수를 통해 데이터 분포 시각화



- 개별 글 별 길이 분포 확인
  - 오른쪽으로 꼬리가 긴 분포 확인



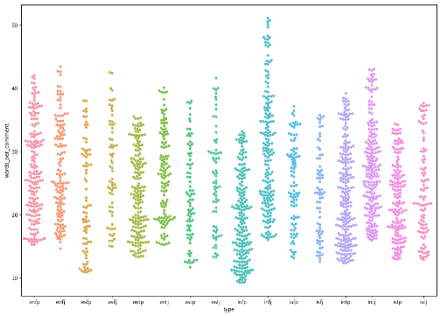
- 모든 MBTI의 'words_per_comment'에서 이상치를 제거한 데이터셋을 추출
  - IQR(Inter Quantile Range)에 해당하는 데이터만 사용

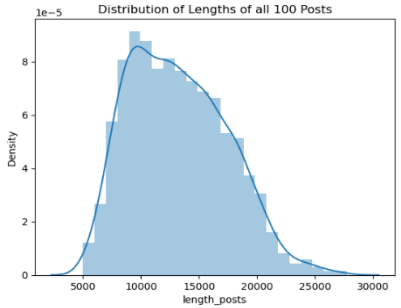
```
# 시각화로 각 데이터셋에 이상치가 있음을 확인.
# mbti 별로 IQR 범위 구해 df 정리하자!
mbtis = ['enfp','enfj','esfp','esfj','entp','entj','estp','estj',
         'infp','infj', 'isfp', 'isfj', 'intp','intj','istp','istj']
df_iqr = pd.DataFrame()
for mbti in mbtis:
    q3 = df2[df2['type']==mbti]['words_per_comment'].quantile(0.75)
    q1 = df2[df2['type']==mbti]['words_per_comment'].quantile(0.25)
    print(mbti,'₩n',q3, q1)
    iqr = df2[(df2['type']==mbti) & (df2['words_per_comment']<=q3) & (df2['words_per_comment']>=q1)]
    df_iqr = pd.concat([df_iqr, iqr])
```

	enfp	enfj	esfp	esfj	entp	entj	estp	estj	infp	infj	isfp	isfj	intp	intj	istp	istj
Q3	42.0675	43.540	38.2675	43.385	35.6300	40.255	38.1975	41.70	33.16	51.440	37.14	35.845	39.555	43.0500	35.1925	37.64
Q1	15.2500	14.605	10.8350	14.960	13.1575	15.255	11.5425	13.25	9.26	15.935	13.14	12.485	12.330	15.9175	12.8625	12.86

개선된 데이터셋

- 각 MBTI 별 'words_per_comment' 값이 IQR 범위에 있는 데이터만 시각화
  - 데이터 분포가 훨씬 고르고, 개별 글 길이 분포도 정규분포에 가까워졌음을 확인 가능





- **2차 전처리 데이터셋**(4377에서 2090으로 데이터 수 감소)

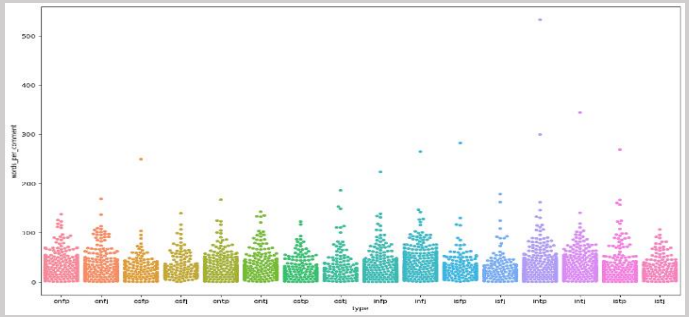
	type	posts	words_per_comment
0	enfp	hi enfps have you ever experienced ending up ...	42.03
1	enfp	i m not in yet but i m going in for 1n0 and s...	27.54
2	enfp	scored in the bottom 8 for order and having a...	38.54
3	enfp	good points he definitely struggles with work...	33.16
4	enfp	been seeing a lot of comments referring to go...	21.59
...	...	...	...
2085	istj	pretty average i do the last one isn t that...	18.37
2086	istj	sounds like me i really feel you i m in a...	21.77
2087	istj	we are considering a loan from a commercial le...	17.13
2088	istj	there s some easy trails that i like right by ...	19.93
2089	istj	removed yes i used to read three bo...	25.43

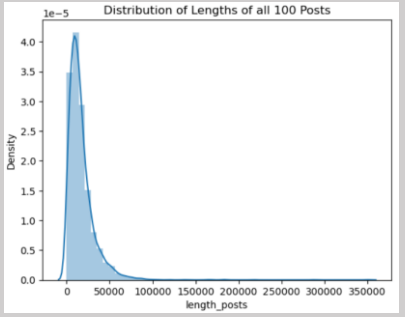
2090 rows × 3 columns

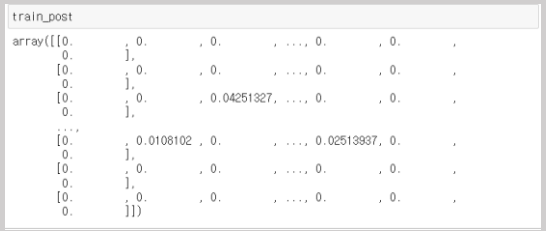
분석과정 및 방법 :: 예측모델 학습 (1)

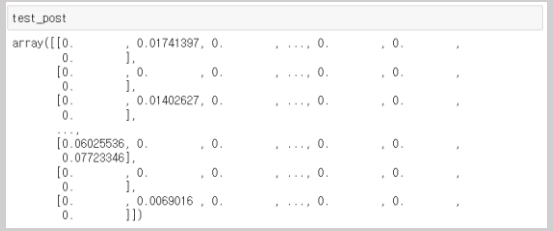
참가팀원 ▼

1차 전처리 데이터셋





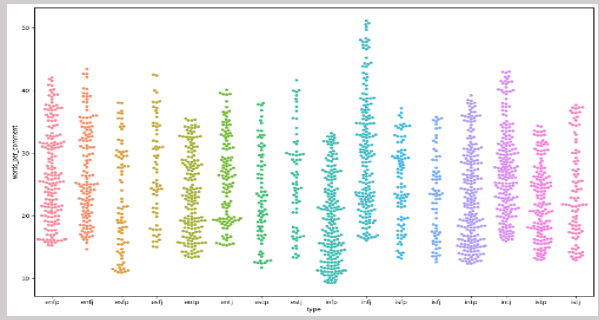


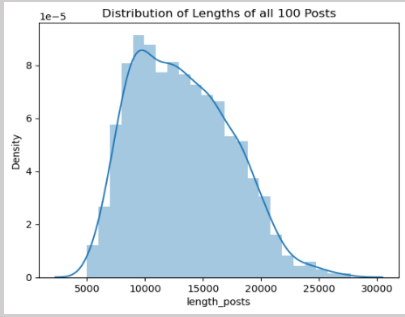


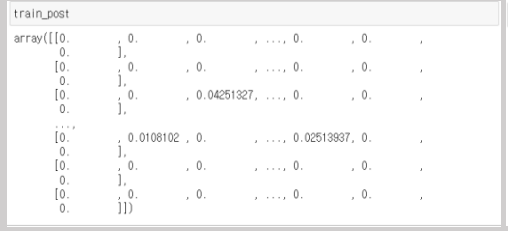
트레인 세트(3142, 5000)

테스트 세트(1048, 5000)

2차 전처리 데이터셋







트레인 세트(1567, 5000)

테스트 세트(523, 5000)

데이터세트 간 훈련 모델 정확도 비교 : XGB, Linear_SVC, SVC, 랜덤포레스트, 선형회귀 모델 5개를 기본 파라미터로 학습시켜 비교

예측 평가 지표 : 정확성(accuracy) 점수 사용해 우수 모델 선정

1차 전처리 데이터셋	Models	Test accuracy
0	XGBoost Classifier	0.479008
1	Linear Support Vector classifier	0.426527
2	Random Forest Classifier	0.360687
3	logistic regression	0.358779
4	Support Vector classifier	0.351145

2차 전처리 데이터셋	Models	Test accuracy
0	XGBoost Classifier	0.447419
1	Linear Support Vector classifier	0.399618
2	Random Forest Classifier	0.370937
3	logistic regression	0.328872
4	Support Vector classifier	0.321224

머신러닝 1차 결론

- 1차, 2차 데이터셋 모두 XGB모델이 가장 우수한 모습을 보여, 최종 모델로 선정하였다.
- 1차 데이터셋에 비해 2차 데이터셋이 전반적으로 정확도가 낮은 모습을 보였으나, 총 데이터 수가 2000개 가까이 차이남에도 0.03의 차이만을 보였기에, 2차 데이터셋이 우수하다 판단하였음.
- 최종 머신 러닝은 **2차 데이터셋**과 **XGB모델**을 이용해 진행하는 것으로 결정하였음.

분석과정 및 방법 :: 예측모델 학습 (2)

최종 선정 모델 XGBoost 최적화

- GridSearchCV를 이용한 최적 파라미터 도출하기.
  - 본 팀 모델인 다중분류에 사용되는 파라미터는 GridSearchCV 이전에 미리 부여했음.
  - n_estimators 파라미터의 경우, 최적 파라미터 검증 시간을 줄이기 위해 값 100으로 임의설정하였음.

다중분류 파라미터

Column Sampling By Tree

```
#XGBoost 정의
# 피처 수가 너무 크기 때문에, 트리 100개로만 시험하자.
xgb= XGBClassifier(objective='multi:softmax', tree_method='exact', n_estimators=100)

xgb_param_grid={
    'learning_rate' : [0.01,0.05,0.1,0.15],
    'max_depth' : [3,5,7,10],
    'colsample_bytree' : [0.8,0.9],
}
# scoring 은 위와 같은 accuray로 설정
xgb_grid=GridSearchCV(xgb, param_grid = xgb_param_grid, scoring="accuracy", n_jobs=-1, verbose = 2
xgb_grid.fit(train_post,train_target)

#best v 수치와 best parameter확인
print("best accuracy : {0: .4f}".format(xgb_grid.best_score_))
print("best param : ",xgb_grid.best_params_)

#dataframe으로 랭킹순보기
result_df = pd.DataFrame(xgb_grid.cv_results_)
result_df.sort_values(by=['rank_test_score'],inplace=True)

#plot
result_df[['params','mean_test_score','rank_test_score']].head(10)
```

: 그라디언트 부스팅에서 민감한 하이퍼파라미터 중 하나로, 텍스트 분류와 같이 칼럼 차원이 많은 경우 모형 성능 향상에 도움이 되는 것으로 알려져있다.

```
Fitting 5 folds for each of 32 candidates, totalling 160 fits
best accuracy :  0.4378
best param :  {'colsample_bytree': 0.8, 'learning_rate': 0.15, 'max_depth': 7}
```

본 팀의 텍스트 분류 모델을 고려해 찾아낸 최적의 파라미터 값

	params	mean_test_score	rank_test_score
14	{'colsample_bytree': 0.8, 'learning_rate': 0.1...	0.437775	1
13	{'colsample_bytree': 0.8, 'learning_rate': 0.1...	0.437142	2
12	{'colsample_bytree': 0.8, 'learning_rate': 0.1...	0.434586	3
29	{'colsample_bytree': 0.9, 'learning_rate': 0.1...	0.434576	4
27	{'colsample_bytree': 0.9, 'learning_rate': 0.1...	0.433304	5
31	{'colsample_bytree': 0.9, 'learning_rate': 0.1...	0.432667	6
30	{'colsample_bytree': 0.9, 'learning_rate': 0.1...	0.432042	7
8	{'colsample_bytree': 0.8, 'learning_rate': 0.1...	0.432038	8
15	{'colsample_bytree': 0.8, 'learning_rate': 0.1...	0.432032	9
11	{'colsample_bytree': 0.8, 'learning_rate': 0.1...	0.432032	10

- 이후 최적의 성능을 보이는 n_estimators값을 찾았고, 이를 200으로 설정 그리고 다항 분류 평가 지표로 쓰이는 eval_metirc='merror' 파라미터를 추가로 설정해주었다.

```
xgb_test1= XGBClassifier(objective='multi:softmax', tree_method='exact',
                         n_estimators=200, eval_metric='merror',
                         colsample_bytree= 0.8,learning_rate= 0.15,
                         max_depth= 7)
xgb_test1.fit(train_post,train_target)
accuracy_score(test_target,xgb_test1.predict(test_post))
```

```
0.47992351816443596
```

최적환경 조성

```
xgb_test1= XGBClassifier(objective='multi:softmax', tree_method='exact',
                         n_estimators=200, eval_metric='merror',
                         colsample_bytree= 0.8,learning_rate= 0.15,
                         max_depth= 7)
xgb_test1.fit(train_post,train_target)
accuracy_score(test_target,xgb_test1.predict(test_post))
```

```
0.47992351816443596
```

0.48에 가까운 정확도로 향상됨

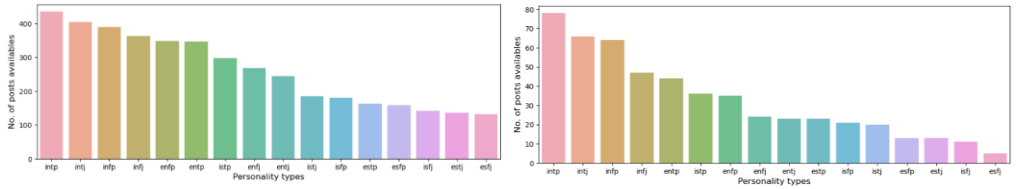
예측모델 완성

최적화 전	Models	Test accuracy
	XGBoost Classifier	0.447419

최적화 후	Models	Test accuracy
	XGBoost Classifier	0.479923

데이터 (원데이터, 예측데이터)간 예측 분포 확인

- 원 데이터와 예측 데이터가 유사한 MBTI 분포를 보임을 확인할 수 있음.



- 총 4개 그룹으로 나누어 비율이 유지되는 모습을 확인할 수 있었음.
  - 이 4개 그룹은 MBTI별 수집된 데이터 수에 따라 나누어짐을 확인할 수 있었음.

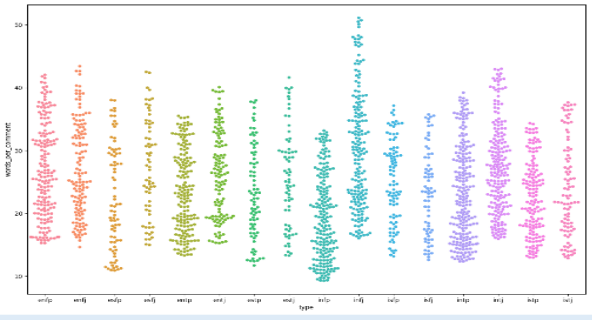
- 상위그룹 intp, intj, infp, infj 4개 그룹
- 2위그룹 entp, istp, enfp, enfj 4개 그룹
- 3위그룹 entj, estp, isfp, istj 4개 그룹
- 4위그룹 esfp, estj, isfj, esfj 4개 그룹

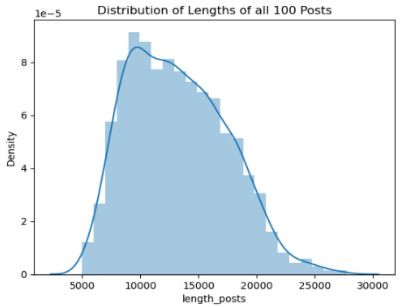
:: 즉 최종모델은 인터넷에서 수집한 게시글을 예측할 때에 최적화된 예측모델임을 확인 ::

가설 1 : IQR -> 0.2 ~ 0.8까지

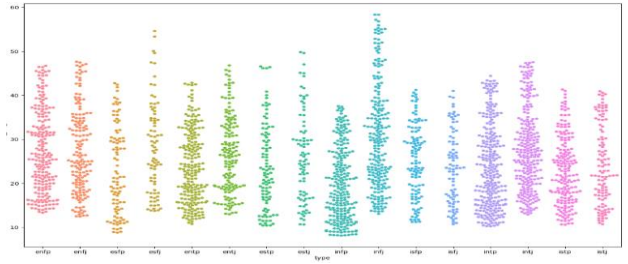
- 이론을 모두 적용한 뒤, 우리 모델의 특성에 보다 맞는 방법으로 모델의 성능을 높일 수 있지 않을까란 가설을 몇 개 추가로 확인해보았음.

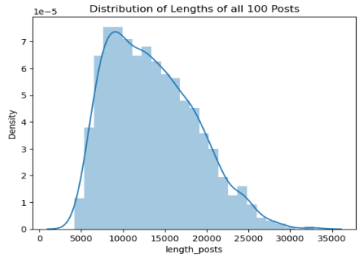
- 기존 IQR 까지 전처리한 데이터셋 분류와 분포.





- 위 IQR까지의 데이터와 0.2~0.8까지의 데이터의 분류와 분포가 거의 유사한 모습을 보임.





- 이에 따라 IQR이 아닌 0.2~0.8까지의 데이터셋이 이후 Reddit 수집 데이터가 아닌 타 커뮤니티 수집 데이터까지도 예측하는 데 있어 보다 좋은 성능을 보일 것이라 가정하고 데이터셋 변경하였음

가설 2 : max_features=5000 => 3000

- 가설 1에서 수정된 데이터셋으로 최적 파라미터를 이용해 정확도를 도출해보니 이전보다 높은 0.5에 가까운 수치가 도출되었음.

```
xgb_test1= XGBClassifier(objective='multi:softmax', tree_method='exact',
                         n_estimators=200, eval_metric='merror',
                         colsample_bytree= 0.8,
                         learning_rate= 0.15, max_depth= 7)
xgb_test1.fit(train_post,train_target)
accuracy_score(test_target,xgb_test1.predict(test_post))
```

0.4968152866242038

- 현재 총 피처값은 5000개로, mbti 예측에 있어 중요도가 낮은 단어까지 지표로 사용하고 있었기에 3000개로 낮추고 정확도를 확인해 보았다.

- 총 피처값 3000에서도 정확도가 급격히 하락하지 않고, 최종모델 1의 정확도인 0.48에 가까운 수치를 보였다.

```
xgb_test2= XGBClassifier(objective='multi:softmax', tree_method='exact',
                         n_estimators=100, eval_metric='merror',
                         colsample_bytree= 0.8,
                         learning_rate= 0.15, max_depth= 7)
xgb_test2.fit(train_post2,train_target2)
accuracy_score(test_target2,xgb_test2.predict(test_post2))
```

0.47770700636942676

최종 예측모델 완성

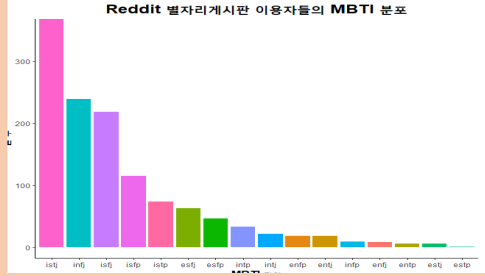
- 위 추가적인 가설 1, 2를 거친 xgb_test2를 팀 뽀의 최종 예측모델로 사용하기로 결정내림

데이터로 팩트체크 :: 별자리 성격론

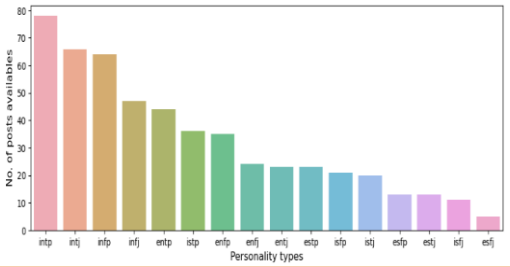
- 앞서 MBTI 배경 설명에서, 혈액형 성격론에 대해 휴먼리서치 자료를 통해 대중의 인식과 신뢰도가 MBTI에 비해 떨어진다는 점을 시사했습니다.
- 혈액형 성격론과 같이 제시한 별자리 성격론에 대해서는 앞서 완성한 최종 MBTI 예측 모델을 이용해 밑 가설을 증명해 보겠습니다.
  - 가설 : 별자리 성격론은 신뢰하기에 그 근거가 부족하다.

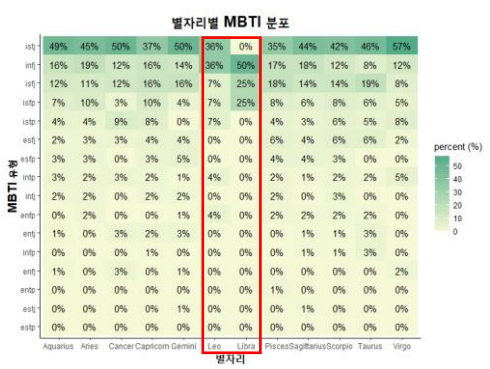
- 아래 게시글에 따른 예측분포를 통해, 예측모델이 의도대로 작동하였음을 확인함

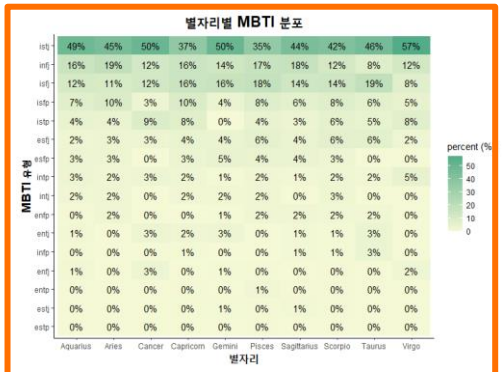
Reddit-별자리 게시글 예측분포



Reddit-MBTI 게시글 예측분포







- 별자리에 상관없이 모든 MBTI 비중이 거의 동일한 분포를 보임을 확인할 수 있음
- 즉 각 별자리 성격이 크게 구분되지 못함을 뜻함.

데이터로 살펴보자 :: 웹소설 장르별 MBTI 분석

- MBTI 예측 모델은 다양한 영역에서 활용 여지가 있음
- 예시사례) 장르별 웹소설 MBTI 예측을 마케팅에 활용해보자.

- 예측 모델에서 가장 큰 비중을 차지하는 ISTP, ISFJ, INFJ, ESFJ 4개의 MBTI를 중요 지표로 사용하였음
- 위 결과 다른 장르들과 다른 분포를 보이는 특이 장르들을 도출해낼 수 있었음.

타겟 가능한 4개 장르
History / Realistic / Sports / War

장르별 MBTI 분포 히트맵

	action	eastern	fantasy	games	history	horror	realistic	scifi	sports	urban	war
istp	29%	38%	32%	31%	32%	34%	29%	33%	17%	36%	25%
istj	5%	3%	5%	5%	5%	2%	2%	2%	2%	2%	5%
isfp	7%	8%	5%	3%	7%	8%	6%	5%	7%	8%	6%
isfj	37%	26%	33%	32%	18%	27%	31%	32%	40%	30%	28%
intp	5%	3%	2%	2%	2%	5%		8%	2%	2%	4%
intj			2%	1%	1%	1%	2%		2%	2%	2%
infp		1%		5%				4%		1%	
infj	12%	11%	5%	8%	24%	10%	18%	8%	21%	7%	19%
estp	1%	1%	1%			1%	2%	1%	1%		1%
estj			1%	1%	1%	1%	1%				1%
esfp	1%		1%	1%	2%	7%	1%		2%	1%	4%
esfj	2%	5%	8%	8%	3%	5%	4%	5%	2%	7%	3%
entp		2%	3%	1%	2%		1%		1%	1%	1%
entj	1%		1%	1%	2%	1%	2%			1%	1%
enfp	2%	1%	1%	1%	1%	1%	1%	2%	1%	2%	1%
enfj		1%	1%	1%	1%	1%	1%	1%	1%	1%	1%

장르

- History, Realistic 장르의 경우 타 장르에 비해 istp과 infj의 비율이 상대적으로 높게 나타났음을 확인할 수 있었음
- Sports 장르의 경우 타 장르와 다르게 istp의 비율이 굉장히 낮게 나타났고, istj와 infj의 비율이 높게 나타났음을 확인할 수 있었음
- War 의 경우 istp, isfj, infj 세 MBTI가 모두 근사하게 나타났다.

MBTI 예측모델을 이용한 마케팅 활용

- 웹소설 마케팅에 있어 ISTP, ISFJ, INFJ 3개의 유형을 타겟팅해 마켓팅을 진행하는 것이 바람직하다는 결과를 도출할 수 있었음.
- Sport 장르의 경우, 모든 장르에서 최고 비중을 차지한 ISTP이 적게 나타난 것으로 보아 비주류 장르임을 알 수 있어, 해당 장르에 대한 마케팅 비용을 줄여도 좋다는 결론 도출 가능
- War 장르의 경우, 상위 3개 MBTI가 가장 고른 분포를 보이는 장르임으로 마케팅을 진행함에 있어 호불호가 갈리지 않는 장르임을 확인할 수 있음.

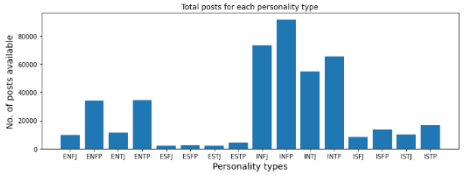
참가팀원 ▼



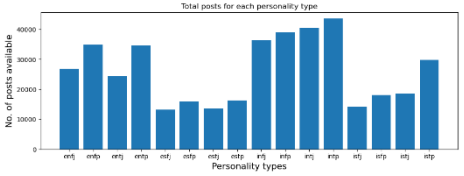
데이터 수집의 한계

- 불균형한 데이터셋
  - 인터넷을 사용하는 MBTI 유형별 비율이 정해져 있는 듯 보임.
    (수집한 커뮤니티 이용 인원 차이가 MBTI별 로 크게 나타남)
  - 일반적인 방법을 사용해서는 균형적인 데이터를 만들 수 없었음.

- Kaggle 에 있는 데이터셋



- 팀 뽀가 수집한 데이터셋



머신 러닝의 한계

- TF-IDF 기반 학습의 한계

- 단어의 빈도를 중심으로 텍스트의 중요도를 부여하기 때문에, 문맥상에서의 단어 간 관계와 유사어 등을 학습하는 데에 한계가 있었다.

- 부족했던 데이터 수

- 전반적인 데이터의 숫자가 너무나 부족해서 보다 의미있는 머신 러닝이 이루어지지 못했다.

데이터 수집 극복안

- 오버샘플링

  - Smote 등을 이용한 오버샘플링 기법을 사용해 조금 더 균형적인 데이터셋을 구성하는 방안을 확인해야 했음.

- 데이터 수집 개선

  - MBTI별로 균형적인 데이터를 다수 수집하는 방안을 마련해야 함.
  - 이벤트 진행 및 유료 API 이용한 대량의 데이터 수집 등

머신 러닝 극복안

- 심화 자연어 처리 진행

- 품사 태깅, 유사어 처리와 같은 심화 텍스트 전처리 기법을 시도한다.

- 딥러닝 활용

- 보다 많은 데이터를 확보하여 텐서플로를 이용한 텍스트 딥러닝 예측기법을 사용한다.

참가팀원 ▼



일상이나 자신이 느낀 점들을 일기 형식으로 기록
-> 예측 데이터 이용자가 직접 생성

매달 작성한 글을 토대로 MBTI 성격 지표 유형을 예측할 수 있음

내가 경험하는 하루 하루와 느끼는 감정들을 기록해
나를 조금 더 이해하고 사랑해보세요

2023 MAY

이번 달 나의
성격 유형 지표는

ENFJ

May 25

오늘은 날씨가 너무 좋아서 기분이 날아갈 것 같아! 이따가 퇴근하고 수진이랑 떡볶이도 먹고 빙수도 먹어야지 ~ 너무 신난다 !!

May 27

오늘 과장님께 크게 혼났다. 내가 잘못한 일도 아닌데 왜 내가 혼나야 하지? 정말 너무 너무 화가 난다 그렇지만 서울에 내 집 마련 하려면 조금 더 버텨야지 ! 파이팅!

_____님의 2023년 3월
MBTI 성격유형 지표는

ISTP

_____님의 2023년 4월
MBTI 성격유형 지표는

ESFP

_____님의 2023년 5월
MBTI 성격유형 지표는

ESFJ