텍스트 데이터 기반
MBTI 예측

프로젝트 개요

01 주제 선정 배경

02 데이터 수집

03 데이터 전처리

04 모델 훈련 및 평가

05 웹 플라스크 구현

주제 선정 배경

최근 한 채용사이트 구인공고에서 특정 MBTI 성향을 거론하며 이에 해당하는 사람은 입사 지원을 자제해 달라는 문구를 넣어 논란이 일었다.

나의 성향을 확인하고 상대방과의 공감대 형성을 위한 도구였던 MBTI는 어느 순간 그보다 더 큰 의미로 작용하는 필터링 수단이 되고 있다.

데이터 수집

캐글 데이터셋 활용 (MBTI) Myers-Briggs Personality Type Dataset

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8675 entries, 0 to 8674
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   type    8675 non-null   object
 1   posts   8675 non-null   object
dtypes: object(2)
memory usage: 135.7+ KB
```

data.head()

	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...

데이터 수집

mbti 'type' 열의 고유 값 확인

```
np.unique(np.array(data['type']))
```

```
array(['ENFJ', 'ENFP', 'ENTJ', 'ENTP', 'ESFJ', 'ESFP', 'ESTJ', 'ESTP',
       'INFJ', 'INFP', 'INTJ', 'INTP', 'ISFJ', 'ISFP', 'ISTJ', 'ISTP'],
      dtype=object)
```

mbti 16가지의 성격 유형이 모두 존재

결측치 확인

```
data.isnull().any()
```

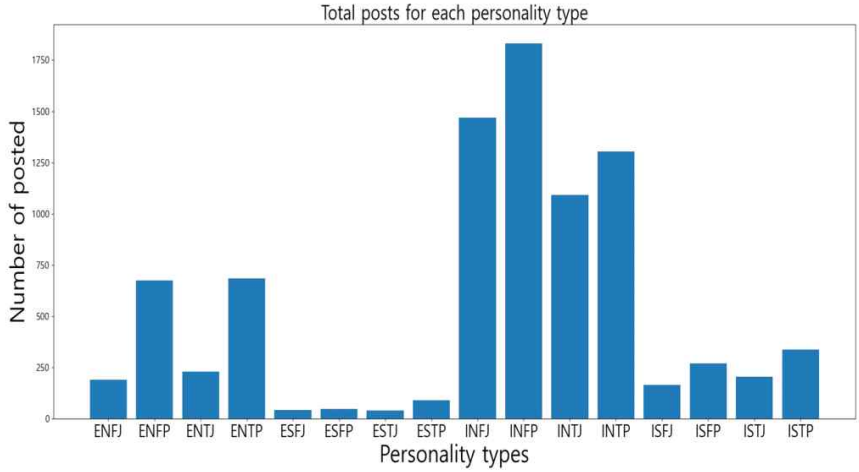
```
type     False
posts    False
dtype: bool
```

결측치 없음

데이터 수집

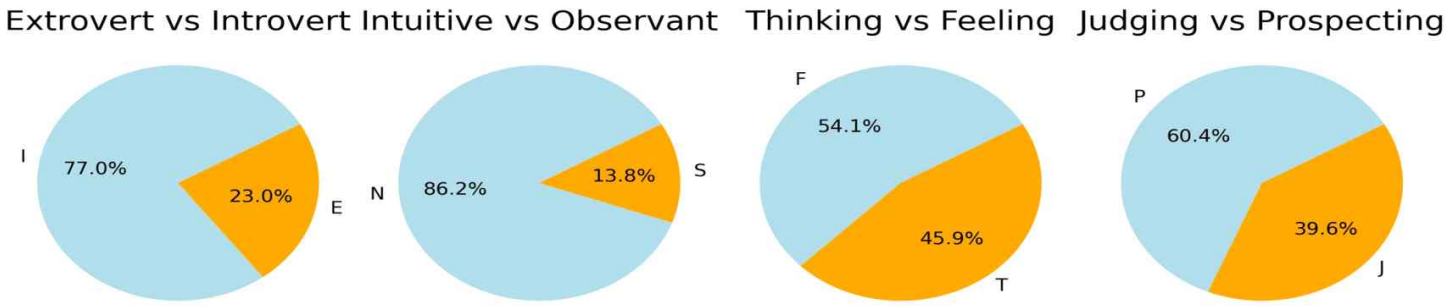
mbti 타입별 게시글 수

type	posts
ENFJ	190
ENFP	675
ENTJ	231
ENTP	685
ESFJ	42
ESFP	48
ESTJ	39
ESTP	89
INFJ	1470
INFP	1832
INTJ	1091
INTP	1304
ISFJ	166
ISFP	271
ISTJ	205
ISTP	337



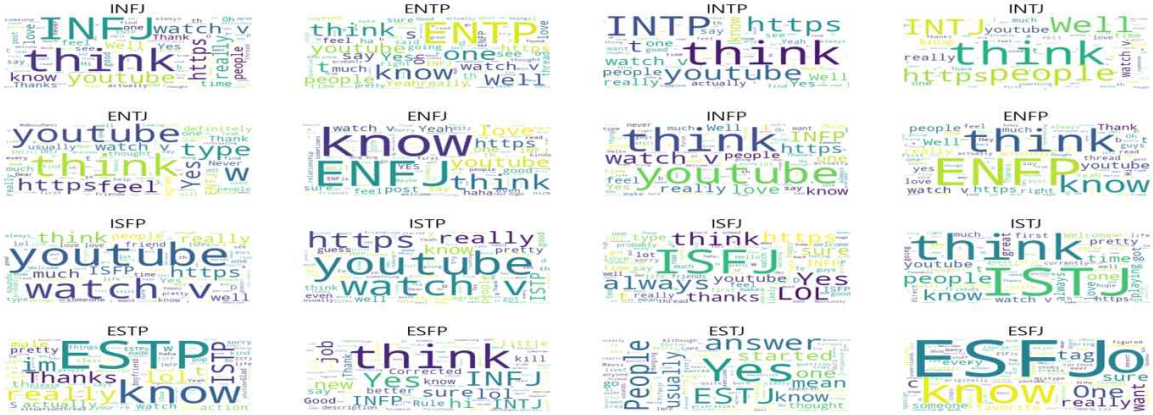
데이터 수집

mbti 타입별 게시글 수



데이터 수집

mbti 타입별 워드클라우드



데이터 전처리

```python
def preprocess_text(text):
    # ||| 로 나뉘어 있는 글 나누기
    text = text.replace('|||', ' ')

    # url 주소 삭제
    text = re.sub(r'https?:\/\/.*?[\s+]', ' ', text)
    # s?는 's' 문자가 선택사항이므로 'http' 및 'https'와 모두 일치함

    # 길이가 1~2인 단어들을 정규 표현식을 이용하여 삭제
    text = re.sub(r'\W*\b\w{1,2}\b', '', text)

    # 영어가 아닌 문자 공백으로 대체
    text = re.sub('[^a-zA-Z]', ' ', text)

    # 영문소문자 변경
    text = text.lower()

    # Remove punctuation : 특수문자 제거
    text = ''.join(ch for ch in text if ch not in string.punctuation)

    # mbti 이름 제거
    mbti_types = ["enfj", "enfp", "entj", "entp", "esfj", "esfp", "estj", "estp",
                  "infj", "infp", "intj", "intp", "isfj", "isfp", "istj", "istp"]
    for mbti_type in mbti_types:
        text = text.replace(mbti_type, ' ')

    # 공백 제거
    text = ' '.join(text.split())

    # 불용어 제거
    stop_words = set(stopwords.words('english'))
    text = ' '.join(word for word in text.split() if word not in stop_words)

    return text
```

||| 으로 나뉘어 있는 글 나누기

url 주소 삭제

길이가 1~2인 단어들 삭제

영어가 아닌 문자 공백으로 대체

영어 소문자 변경

특수문자 제거

각 mbti 이름 제거

공백 제거

불용어 제거

모델 훈련 및 평가

나이브베이즈 모델

```
# 텍스트 데이터 수치로 변경
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.preprocessing import LabelEncoder

cvect = CountVectorizer()
dtm = cvect.fit_transform(df["posts"])

DTM_array = dtm.toarray()
DTM_array.shape # (8675, 97342)

# 타겟 labelencoder
label_encoder = LabelEncoder()
target = df['type']
target = label_encoder.fit_transform(target)

# train/test split : 훈련셋(70) vs 테스트셋(30)
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(
    DTM_array, target, test_size=0.3)

print((x_train.shape), (y_train.shape), (x_test.shape), (y_test.shape))

# Naive Bayes 분류기
from sklearn.naive_bayes import MultinomialNB # nb model
from sklearn.metrics import accuracy_score

# 학습 모델 만들기 : 훈련셋 이용
nb = MultinomialNB()
model = nb.fit(X= x_train, y = y_train)

# 학습 model 평가 : 테스트셋 이용
y_pred = model.predict(X = x_test)
```

훈련셋(70) vs 테스트셋(30)

모델 훈련 및 평가

분류 정확도

```
# 분류정확도
acc = accuracy_score(y_true = y_test, y_pred = y_pred)
print('분류정확도 :', acc)
```

```
분류정확도 : 0.311179408374952
```

분류정확도가 현저히 낮음

훈련셋, 테스트셋 비율을 바꾸고, 다른 모델들도
테스트 해본 결과 정확도가 비슷함

16개의 카테고리를 분류하는데 있어 데이터셋의
양이 부족하다 생각됨

데이터의 양이 훨씬 많은
데이터셋(posts가 이미 전처리 되어있음)을 활용

데이터 수집

캐글 데이터셋 활용 MBTI Personality Types 500 Dataset

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 106067 entries, 0 to 106066
Data columns (total 2 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   posts   106067 non-null  object
 1   type    106067 non-null  object
dtypes: object(2)
memory usage: 1.6+ MB
```

```
data.head()
```

	posts	type
0	know intj tool use interaction people excuse a...	INTJ
1	rap music ehh opp yeah know valid well know fa...	INTJ
2	preferably p hd low except wew lad video p min...	INTJ
3	drink like wish could drink red wine give head...	INTJ
4	space program ah bad deal meing freelance max ...	INTJ

모델 훈련 및 평가

분류 정확도

```
# 텍스트 데이터 수치로 변경
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.preprocessing import LabelEncoder

cvect = CountVectorizer()
dtm = cvect.fit_transform(df["posts"])

DTM_array = dtm.toarray()
DTM_array.shape # (8675, 97342)

# 타겟 labelencoder
label_encoder = LabelEncoder()
target = df['type']
target = label_encoder.fit_transform(target)

# train/test split : 훈련셋(70) vs 테스트셋(30)
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(
    DTM_array, target, test_size=0.3)

print((x_train.shape), (y_train.shape), (x_test.shape), (y_test.shape))

# Naive Bayes 분류기
from sklearn.naive_bayes import MultinomialNB # nb model
from sklearn.metrics import accuracy_score

# 학습 모델 만들기 : 훈련셋 이용
nb = MultinomialNB()
model = nb.fit(X= x_train, y = y_train)

# 학습 model 평가 : 테스트셋 이용
y_pred = model.predict(X = x_test)
```

```
# 분류정확도
acc = accuracy_score(y_true = y_test, y_pred = y_pred)
print('분류정확도 :', acc)
```

```
분류정확도 : 0.7364633418182961
```

나이브베이즈 모델로 같은방식으로 훈련한 결과
분류정확도가 두배이상 높아짐

모델 훈련 및 평가

mbti 예측 모델

```
# 문서분류기 함수
def classifier(texts):
    global model, cvect, label_mapping

    DTM_test = cvect.transform([texts])
    X_test = DTM_test.toarray()

    y_pred = model.predict(X=X_test)
    y_pred_result = label_mapping[y_pred[0]]

    return y_pred_result


my_posts = """Hi I am 21 years, currently, I am pursuing my graduate degree
in computer science and management (Mba Tech CS ),
It is a 5-year dual degree.... My CGPA to date is 3.8/4.0 .
I have a passion for teaching since childhood.
Math has always been the subject of my interest in school.
Also, my mother has been one of my biggest inspirations for me.
She started her career as a teacher and now has her own education trust
with preschools schools in Rural and Urban areas. During
the period of lockdown, I dwelled in the field of blogging and content creation on Instagram.
to spread Love positivity kindness . r
I hope I am able deliver my best to the platform and my optimistic attitude helps in the growth that is expected.
Thank you for the opportunity."""

y_pred_result = classifier(my_posts)
print("MBTI 결과:", y_pred_result)
```

```
MBTI 결과: INTJ
```

웹플라스크 구현


127.0.0.1/service2
Predict Your Personality Type
arts, literature, and music, often
leading to a strong sense in these
areas. Their interest frequently
translates into their actions,
resulting in many people with
outstanding talents in these fields.
They tend to overthink. This is an
inevitable combination, as their few
words, rich im
submit

웹플라스크 구현


127.0.0.1/service2_result
Your Personality Type
ENFJ
INTJ
INTP
ENTJ
ENTP
INFJ
INFP
ENFJ
ENFP
ISTJ
ISFJ
ESTP
ESFJ
ISTP
ISFP
INTP
ESFP

THANK YOU