**나를 분석해봐 : 개인 텍스트 기반 MBTI 테스트**

조 이름: 런닝머신

201723825 임윤진

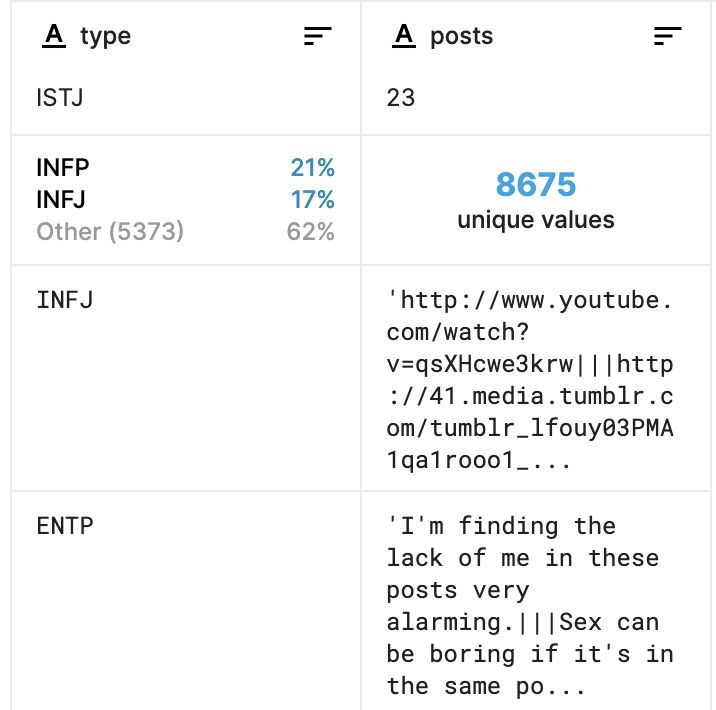
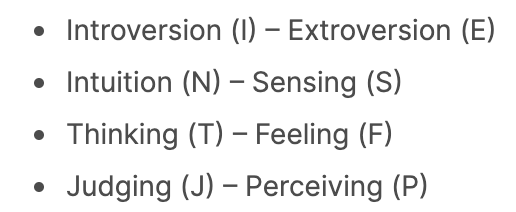
201820742 고예준

201820793 박민석

201820730 배해진

201421092 박효성

**1.Using Data** : MBTI Myers-Briggs Personality Type Dataset. <https://www.kaggle.com/datasnaek/mbti-type>.

-(데이터 부분예시)

총 8675개의 데이터. Feature는 type, posts로 이루어져 있다.

Type: MBTI 16가지 유형 중 하나.

Posts: personalityCafe forum(<https://www.personalitycafe.com/forum/>)에서 특정 유형에 해당하는 사람이 쓴 글 50개를 ‘|||’로 분리.

**2.기존의 문제**

기존의 mbti 성격유형 검사는 굉장히 많은 양의 질의응답이 필요했고, 많은 시간이 소비되었다. 

예를 들어 16personality 사이트에서는 12분의 검사시간이 소요되고 질문의 갯수도 60개로 많다.

[기존의 mbti 유형 검사 사이트](https://www.16personalities.com/ko/%EB%AC%B4%EB%A3%8C-%EC%84%B1%EA%B2%A9-%EC%9C%A0%ED%98%95-%EA%B2%80%EC%82%AC)

-(기존테스트문항)

**3.문제 해결 방안**

사람들이 영어로 쓴 글을 기반으로 하여 mbti 유형을 알려주는 심리테스트 방식을 제안한다. 별도의 검사 없이 자신이 쓴 글의 성격에 따라 mbti 유형을 분류할 수 있다는 가정을 전제로 한다.

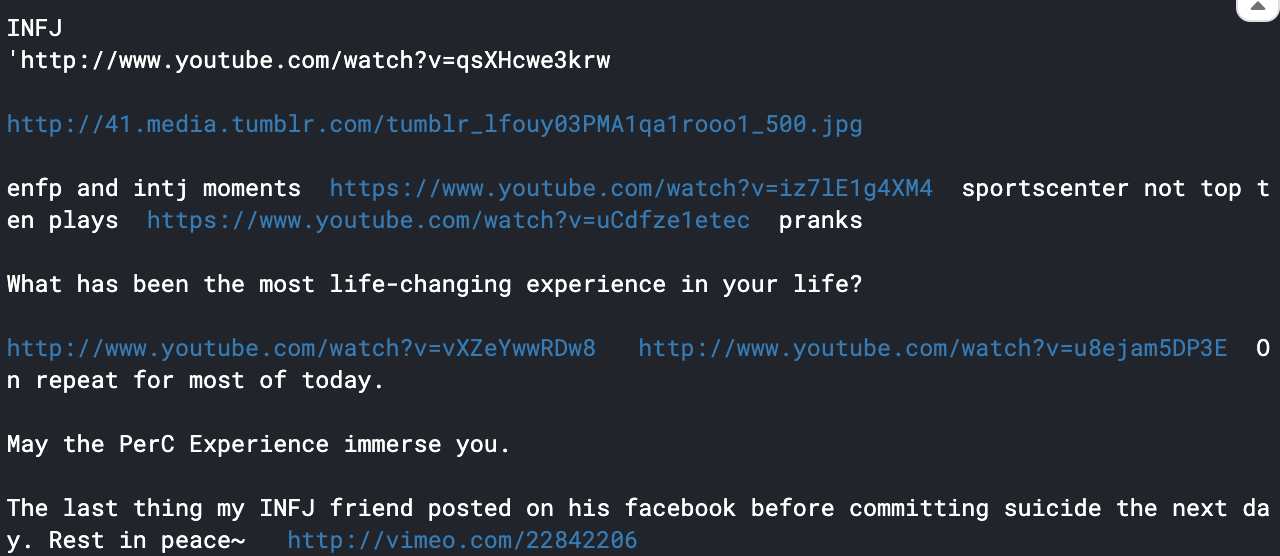
분석할 데이터의 feature는 2개로 하나는 mbti유형(type)이고 나머지 하나는 그 유형에 해당하는 사람이 쓴 글(posts)이다.

<text data 분석 방법>

1. 사람들이 쓴 글(posts)을 다루기 쉽게 전처리한다.

1. 모두 알파벳 소문자로 바꾼다.
2. url은 삭제한다. ex) http:// , youtube
3. ‘|||’ 별로 나눈다.
4. mbti 유형별로 문장들을 정리한다.

아래 그림은 하나의 sample에 들어있는 type과 post 내용이다.



2. 사람들이 쓴 글을 단어 단위로 분석, 등장 빈도수로 정렬하여 단어사전을 생성한다. 가장 높은 빈도수를 기록한 단어 부터 id을 붙인다. id은 0부터 1씩 증가한다. 만약, test\_data에 train\_data에서 만든 단어사전에 포함되지 않는 단어(OOV: Out Of Vocabulary)가 들어올 경우는 NULL값으로 처리된다.

3. 모든 글들의 길이의 평균과 최대 길이를 분석하여 분석할 글의 최대 길이(max\_len)를 정한다.

4. 만약 분석할 글의 길이가 max\_len보다 짧으면 나머지는 0으로 채우고, 더 길면 초과분은 삭제한다.

5. 각각의 mbti 유형에 해당되는 정제된 글들로 모델을 학습시킨다.

6. 그 후 학습된 모델을 활용하여 사용자가 입력한 글을 분석하여 어떤 mbti 유형에 가장 가까운지 구한다.

**4.솔루션 평가 방법**

1. 기존의 데이터를 train\_data와 test\_data로 나눈다. 비율은 9대 1로 한다.

2. Classification 모델의 학습은 train\_data로 이루어지며 최종 솔루션 평가는 test\_data의 정확성으로 평가한다. Softmax, cross-entropy를 적용한다.

(모델은 Rnn이나 Logistic regression중 하나를 이용하여 좀 더 나은 모델을 선택한다.)

3. 실제 팀원의 MBTI 유형과 팀원이 작성한 영문 텍스트를 수집하여 모델에 적용한 후 성능을 확인해본다.