Week1-4 과제

## 한국 스트리밍 서비스 (왓\*, 쿠\*플레이, 티\*)에서 시청자가 영화를 보고 남긴 리뷰를 긍정과 부정으로 나누어 볼 수 있는 대시보드를 만들려고 한다. **리뷰 긍부정 판별 모델**을 만들려고 할 때, NLP 리서처/엔지니어로서 어떤 의사 결정을 할 것인지 각 단계에 맞춰 작성해보자. (단, 수집된 리뷰 데이터의 개수가 1,000개 미만이라고 가정하자.)

대시 보드 예시.

| 긍정 | 부정 |
| --- | --- |
| ID: 맛있는 우유 GT 고소한 저지방  REVIEW: 킬링타임으로 보기 좋은 영화입니다! 개인적으로 강추 | ID: 갤럭시 버즈2  REVIEW: 너무 별로네요 |
| ID: 동전 저금통 REVIEW: 아 역시 흥행보증수표 배우 OOO이 나오니 재미가 없을수가 없네 | ID:호가든 맥주 컵 REVIEW: 이 배우를 쓰고도 이것밖에 못 하다니… |

### 문제 정의

풀고자 하는 문제를 정의하세요. 또한 데이터 생성 시 고려해야할 사항이 있다면 무엇인지 설명하세요. (예, 만약 긍정 리뷰가 부정 리뷰보다 많은 경우 어떻게 해야 할까?, 길이가 정말 긴 리뷰는 어떻게 전처리 해야 할까?)

—-

* 우리가 직면한 Task는 리뷰어의 글이 어떤 성격을 띄고있는지(긍정, 부정) 판별(Classifiaction) 하는 모델을 제작해야하며 한국어임을 감안하여 모델을 선정 시 한국어로 Pre-trained 된 모델을 가져 올 수 있도록 해야합니다. 또한 데이터가 불균형 한 경우 다운 샘플링을 통하여 맞춰주거나, Augmentation기법을 사용하여 데이터의 수를 맞춰주는 방법(Up-Sampling)을고려할 것 같습니다. 보통 리뷰를 달 때 길이를 정해두나 너무 길이가 길어지는 경우 전체적으로 평균적인 리뷰의 길이를 찾아서 그만큼 맞춰서 데이터를 잘라내는 방법을 취하여 아웃라이어를 제거하는 형식을 가져갈 것 같습니다.

### 오픈 데이터 셋 및 벤치 마크 조사

리뷰 긍부정 판별 모델에 사용할 수 있는 한국어 데이터 셋이 무엇이 있는지 찾아보고, 데이터 셋에 대한 설명과 링크를 정리하세요. 추가적으로 영어 데이터셋도 있다면 정리하세요.

—-

* 데이터 셋 소개

# Naver sentiment movie corpus v1.0

* + - 링크 : <https://github.com/e9t/nsmc/>
    - 데이터는 id와 리뷰 댓글과 긍정인지 부정인지를 나타내는 이진값이 존재합니다(긍정 1, 부정 0)
    - 학습 데이터 : 150,000
    - 검증 데이터 : 50,000
    - 테스트 데이터 : 0

## Classify the sentiment of sentences from the Rotten Tomatoes dataset

* + 링크 : <https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data>
  + 데이터는 **PhraseId, SentenceId, Phrase, Sentiment** 로 이루어져 있으며 우리는 Phrase(구문)와 , Sentiment(감정)만 사용해도 괜찮을 것 같습니다. Pharase는 리뷰이며 Sentiment는 0~4사이의 수치로 표현했으며 5점척도를 기준으로 하였습니다

0 - negative

1 - somewhat negative

2 - neutral

3 - somewhat positive

4 - positive

* + 학습 데이터 : 156,060
  + 검증 데이터 : 66,292
  + 테스트 데이터 : 0

### 모델 조사

Paperswithcode(https://paperswithcode.com/)에서 리뷰 긍부정 판별 모델로 사용할 수 있는 SOTA 모델을 찾아보고 SOTA 모델의 구조에 대해 간략하게 설명하세요. (모델 논문을 자세히 읽지 않아도 괜찮습니다. 키워드 중심으로 설명해 주세요.)

—-

* SMART-RoBERTa(Large Parameter)
  + 주요 키워드
    - RoBERTa는 BERT의 Duplicate study임
    - BERT 모델을 개선하기위해 Layer를 길게 늘리고 배치를 크게 만들어줫습니다.
    - NSP(next sentence prediction)을 삭제하였으며
    - 긴 시퀸스를 학습합니다.
    - Dynamic Masking을 적용하엿습니다.

### 학습 방식

* 딥러닝 (Transfer Learning)  
  사전 학습된 모델을 활용하는 (transfer - learning)방식으로 학습하려고 합니다. 이 때 학습 과정을 간략하게 서술해주세요. (예. 데이터 전처리 → 사전 학습된 모델을 00에서 가져옴 → …)

1. 데이터 전 처리(불용어 제거, Sampling, Augmentation, Remove Outlier etc.)
2. Huggingface를 이용하여 사전 학습 모델 및 vocabulary파일 로드
3. 학습에 적합한 데이터로 변환(LabelEncoding, Padding etc.) 및 DataLoader 생성 후 DataLoader에 변환 된 데이터를 태움
4. Task에 맞는 모델 레이어 함수 생성(Classification)
5. 모델 선언 및 모델에 사전 학습 모델을 태움
6. 학습
7. 모델 평가

* (Optional, 점수에 반영 X) 전통적인 방식  
  Transfer Learning 이전에 사용했던 방식 중 TF-IDF를 이용한 방법이 있습니다. TF-IDF를 이용한다고 했을 때, 학습 과정을 간략하게 서술해주세요.

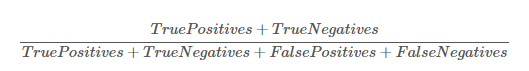
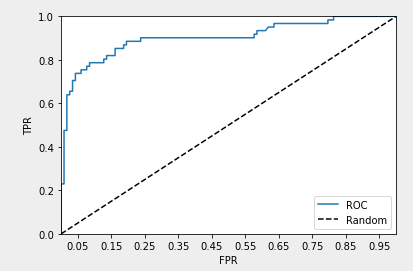
—-

1. 데이터 전 처리(불용어 제거, Sampling, Augmentation, Remove Outlier etc.)
2. Tokenizer 단어를 Vector화
3. Vector화 된 단어의 TF-IDF 계산
4. Logistic Regression 모델 선언
5. 학습
6. 모델 평가

### 평가 방식

긍부정 예측 task에서 주로 사용하는 평가 지표를 최소 4개 조사하고 설명하세요.

—-

* Accuracy
  + Acuuracy는 정확도를 뜻하며 지문의 정답이 맞는지 확인하여 퍼센트 비율을 내놓습니다.
  + 식은 다음과 같습니다.
  + 
* AUC
  + AUC는 ROC curve곡선 아래의 영역을 나타냅니다. 1에 가까울수록 좋은 수치로 판단합니다. 주로 아래의 그림과 같이 표현하고 곡선 밑 면적을 구해서 Score를 냅니다.
  + 
* Precision
  + Precision은 정밀도라고 하며 예측한 True값 중실제로 True값을 예측한 데이터를 뜻합니다. 수식은 아래와 같습니다.
  + 
* F1 Score
  + F1 Score는 Precision과 Recall을 조합하여 하나의 통계치로 반환하며 일반적인 평균계산이 아닌 조화평균을 계산합니다. 수식은 아래와 같습니다.
  + 