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.)

1. 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?)
풀어야하는 문제는 감성 분석이다. 예측하고 싶은 각각의 문장이 긍정 혹은 부정인지 분류하는 문제라고 생각 할 수 있다.

모델을 학습하는 경우 데이터는 긍정과 부정이 비율이 비슷하게 사용하여 train / test 나눠서 모델을 훈련 시킨다.

모델을 학습 시키기 위하여 중요한 부분이 문서의 임베딩 방법이다. 리뷰 별로 임베딩을 해주게 되고 임베딩 된 벡터를 사용하여 모델을 학습시키게 되는데 학습 시키려는 모델에 따라서 맞는 임베딩 방법을 선택한다. 리뷰의 길이 역시 임베딩 문제에 해당한다. 각 리뷰 별로 길이가 다르기 지만, 모델에 입력되는 데이터는 길이가 동일해야한다. 일반적으로는 학습에 사용되는 데이터를 가장 많이 표현할 수 있는 데이터의 길이를 사용하여 최대로 긴 문장의 길이를 정하는 방식을 사용한다.

2. 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.

네이버 영화 리뷰:
한국어로 된 네이버의 영화 리뷰 데이터
데이터 구성은 아이디, 문장, 레이블로 구성
총 200,000의 데이터 중 150,000개의 학습 데이터와 50,000의 테스트 데이터로 구성
긍정과 부정 레이블의 비율은 50:50으로 구성
긍정 / 부정의 리뷰로만 구성되어 있고 중립적인 리뷰 없이 구성된 데이터
https://github.com/e9t/nsmc/

네이버 쇼핑 리뷰:
네이버 쇼핑에서 제품별 후기를 별점과 함께 수집한 데이터
데이터는 탭으로 분리
첫째 열에는 별점(1~5) 표기
두번째 열에는 문장이 표기
긍정 / 부정으로 분류하기 위하여 중립적인 3점의 데이터는 제외
긍정 / 부정의 레이블 구성은 50:50 동등
총 200,000개의 리뷰로 구성
https://github.com/bab2min/corpus/tree/master/sentiment

한국에 스팀 리뷰:
게임 유통 서비스인 Steam의 각종 게임에 달린 한국어 리뷰를 데이터
게임 커뮤니티 특성 상 비속어 및 은어가 많이 사용된 것이 특징
데이터는 탭으로 분리
첫번째 열은 긍정 / 부정(1=긍정, 0=부정)을 나타냄
두번째 열은 리뷰 텍스트가 위치
긍정과 부정의 비율이 1:1에 가깝도록 샘플링
총 100,000개의 데이터

https://github.com/bab2min/corpus/tree/master/sentiment

3. 모델 소개

학습 방법론:
일반 적인 감정 이진 분류 문제일 경우 워드 임베딩을 통하여 리뷰를 표현하고 LSTM을 사용하여 분류하는 모델을 구성
LSTM을 사용(모델은 긍정 / 부정 선택지 중 하나를 예측하는 이진 분류 문제)
출력층에 로지스틱 회귀를 사용
활성화 함수로는 시그모이드 함수
손실 함수로 크로스 엔트로피 함수 사용

그러나 최근 논문의 추세를 보면 문서의 표현방식도 다양해지고 모델도 어텐션이나 트렌스포머를 이용한 모델이 주로 사용

모델:
NB-weighted-BON + dv-cosine

조사한 모델은 paperswithcode에서 IMDb의 감성 분류 분제에서 가장 좋은 성능을 보이는 모델이다.

기존의 방법과 차이를 보이는 점은 문서표현에서 기존의 임베딩 방법과 달리 코사인 유사도를 사용하여 각 문서 별로 길이가 같은 벡터로 표현한다.

Sentiment Classification using Document Embeddings trained with Cosine Similarity 논문의 내용을 참고

4. 평가지표

이진 분류 문제에서 평가지표는 정확도를 사용한다. 모델을 통하여 예측한 값이 실제의 레이블과 일치하는 비율을 통하여 모델의 성능을 검증하는 방법이다.
Accuracy (%): TP + TN / TP + TN + FP +FN