

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 모델을 만들려고 할 때, **NLP** 리서처/엔지니어로서 어떤 의사 결정을 할 것인지 각 단계에 맞춰 작성해보자. (단, 수집된 리뷰 데이터의 개수가 **1,000개** 미만이라고 가정하자.)

대시보드 예시

긍정	부정
ID : REVIEW :	ID : REVIEW :
ID : REVIEW :	ID : REVIEW :

1. 문제 정의 - **Sentiment Analysis**

풀고자하는 문제를 정의하세요. 또한 데이터 생성 시 고려해야할 사항이 있다면 무엇인지 설명하세요.

(예 : 만약 긍정 리뷰가 부정 리뷰보다 많은 경우 어떻게 해야할까?, 길이가 긴 리뷰는 어떻게 전처리 할까?)

Sentiment Analysis

텍스트에 들어있는 의견이나 감성, 평가, 태도 등 주관적인 정보를 컴퓨터를 통해 분석하는 과정
데이터 불균형?

수집한 데이터에서 불균형이 일어날 수 있다.

긍정 리뷰가 부정 리뷰보다 압도적으로 많거나, 반대의 경우도 발생할 수 있다.

Data Augmentation

수집된 리뷰 데이터수가 1000개 이하면 불균형이 일어날 가능성은 적다.

Data Augmentation을 통해 데이터의 균형을 맞춰야 할 것이다.

특정 단어를 유의어로 교체하는 방식으로 데이터 양을 늘려 균형을 맞춘다.

임의로 단어를 삽입한다.

문장 내에서 임의의 두 단어의 위치를 바꾼다.

임의로 단어를 삭제하여 균형을 맞춘다.

참고논문 : <https://arxiv.org/pdf/1901.11196v2.pdf>

길이가 긴 문장? 짧은 문장?

한국어의 경우, 단어들 간의 구분이 명확하지 않기 때문에 토큰화 작업이 까다롭다.

konlpy와 같은 형태소 분석기를 통해 토큰화 한다.

단어를 토큰화하고 정수 인코딩을 통해 단어마다 고유한 정수를 부여
길이가 다른 문장을 파악하고, 패딩을 통해 길이를 맞춰준다.

Transfer Learning

BERT와 같이 아주 큰 데이터셋을 이용해 Pre-trained된 모델의 가중치를 가져와 우리가 해결하고 하는 과제에 맞게 fine-tunning하여 사용하는 것이다.

결과적으로 비교적 적은 수의 데이터를 가지고도 우리가 원하는 task를 해결할 수 있는 딥러닝 모델을 만들 수 있다.

2. 오픈 데이터셋 및 벤치 마크 조사

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

NSMC(한국어)

Naver Sentiment Movie Corpus

15만개의 train data, 5만개의 test data

링크 : <https://github.com/e9t/nsmc>

id	document	label
고유한 id번호	리뷰 내용	0 : 부정, 1 : 긍정

Dataset Preview		
Subset	Split	Go to dataset viewer
default	train	
id (string)	document (string)	label (class label)
9976970	아 대병.. 진짜 짜증나네요 목소리	0 (negative)
3819312	흠...포스터보고 초당영화를....오버연기조차 가볍지 않구나	1 (positive)
18265943	너무재미있었다 그래서 보는것을 추천한다	0 (negative)
9845819	교도소 미야기구면 ..솔직히 재미는 없다..평점 조정	0 (negative)
6483659	사이문풀그의 악살스런 연기가 들보였던 영화! 스파이더맨에서 놀러보이기만 했던 커스틴 턴스트가 너무나도 이쁠보였다	1 (positive)
8483919	막 걸음마 한 3세부터 초등학교 1학년생인 8살동영화.ㅋㅋㅋ...불반개도 아파옴.	0 (negative)
7797314	원작의 긴장감을 제대로 살려내지 못했다.	0 (negative)
9443947	불 반개도 아깝다 옥나온다 미용경 길을우 연기생활이몇년인지..정말 빨로 해도 그것보단 낫겠다 납치.강금민반복반복..미드라마는 가족도없다 연기..	0 (negative)
7156791	액션이 없는데도 재미 있는 몇안되는 영화	1 (positive)

source : <https://huggingface.co/datasets/nsmc>

imdb dataset

imdb의 영화리뷰 데이터셋

25,000 train data , 25,000 test data

링크 : <https://huggingface.co/datasets/imdb>

text	label
review 내용	0 : neg(부정) , 1: pos(긍정)

Dataset Preview

Subset: plain_text | Split: train

text (string)	label (class label)
This is the worst thing the TMNT franchise has ever spawned. I was a kid when this came out and I still thought it was deuce, even though I...	0 (neg)
Sometime in 1998, Saban had acquired the rights to produce a brand-new Ninja Turtles live-action series. Naturally, being a fan of the TMNT...	0 (neg)
This is the biggest insult to TMNT ever. Fortunately, officially Venus does not exist in canon TMNT. There will never be a female turtle, thi...	0 (neg)
I did not like the idea of the female turtle at all since 1987 we knew the TMNT to be four brothers with their teacher Splinter and their...	0 (neg)
I cannot stay indifferent to Lars van Trier's films. I consider 'Breaking the Waves' nothing less than a masterpiece. I loved 'Dancer...	0 (neg)
This film is terrible. You don't really need to read this review further. If you are planning on watching it, suffice to say - don't...	0 (neg)

End of preview (truncated to 100 rows)

source : <https://huggingface.co/datasets/imdb>

GLUE Benchmark : SST

Stanford Sentiment Treebank

GLUE Benchmark에서 sentiment analysis에 사용되는 데이터셋

Rotten Tomatoes로 부터 수집한 10,000건의 리뷰이다.

다른 리뷰에 비해 길다고 한다.

링크 : <https://huggingface.co/datasets/sst>

추가로 KLUE,KoQuad에는 조사해보았지만, sentiment analysis에 대한 지표가 없다.

KLUE : <https://klue-benchmark.com/>

KorQuad : <https://korquad.github.io/>

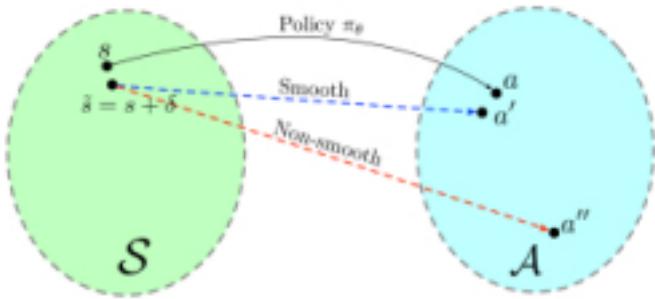
3. 모델조사

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

Keyword : Smoothness-inducing regularization

효과적으로 model의 complexity를 관리하기 위한 방법



source : <https://deepai.org/publication/deep-reinforcement-learning-with-smooth-policy>

input에 미세한 noise를 부과했을 때, output의 분포가 noise를 부과하지 않은 output의 분포와 많이 떨어지지 않도록 제어한다고 한다.

Keyword : Bregman proximal point optimization

모델의 급격한 update를 방지하기 위한 방법

기존 parameter를 이용해 산출된 output과 새롭게 update될 parameter를 이용해 산출된 output의 분포차이를 줄이도록 제어한다고 한다.

4. 학습방식

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 → ...)

Transfer Learning(전이학습)

사전학습(Pre-trained)된 모델의 가중치로 초기화된 모델을 나만의 task로 추가 학습시키는 것을 말한다.

보통의 딥러닝 모델보다, 더 적은 양의 학습 데이터를 사용하고 속도도 빠르다.

데이터 로드

데이터 전처리

수집한 데이터 셋을 확인한다.

이 과정에서 데이터의 개수나 크기, 중복유무, NULL값 처리, 공백, 불용어 제거 등을 수행 한다.

수집한 데이터를 토대로 tokenizing을 진행한다.

인코딩을 통해 토큰에 할당한다.

패딩을 통해 문장의 길이를 맞춰준다.

train - validation - test set 분리

모델 학습

pre-trained model의 모듈을 불러와서 학습

test set을 통해 모델의 학습을 평가한다.

TF-IDF

TF : Term Frequency

특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값

이 값이 높을수록 문서에서 중요하다고 생각할 수 있다.

IDF : Inverse Document Frequency

단어 자체가 문서 내에서 자주 사용되는 경우, 그 단어가 흔하게 등장한다는 것을 의미한다. 이것을 DF라 하고, 이 값의 역수가 IDF
TF와 IDF를 곱한 값을 나타낸다.

데이터 로드 및 데이터 전처리

데이터 토큰화

토큰화된 데이터를 토대로 TF-IDF 벡터화를 진행
train-validation-test set 분리

모델학습

DecisionTree, Logistic Regression (Binary Classification)
teset set을 통해 모델의 학습을 평가한다.

5. 평가 방식

긍부정 예측 task에서 주로 사용하는 평가 지표를 최소 4개 조사하고

설명하세요. **Accuracy**(정확도)

가장 직관적으로 모델의 성능을 나타낼 수 있는 평가지표

전체 표본 중 정확히 분류된 표본의 수

data가 불균형하게 되면, 한 쪽으로 치우쳐진 편중(bias)이 나타날 수 있다.

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) <i>Type II Error</i>	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) <i>Type I Error</i>	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

source : <https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html>

F1-Score

Precision과 Recall의 조화평균을 의미한다.

F1-Score가 1에 도달할 때 최적의 Precision과 Recall을 의미
높을수록 좋은 모델을 의미한다.

$$F1 Score = 2 \times \frac{recall \times precision}{recall + precision}$$

source :<https://inside-machinelearning.com/en/recall-precision-f1-score-simple-metric-explanation-machine-learning/>

Precision

모델이 맞게 분류한 것중에 실제 값이 맞는 경우

$$Precision = \frac{True Positive}{True Positive + False Positive}$$

source :<https://pinatadata.com/theory/precision-vs-recall/>

Recall

본래 가진 실제 Positive 중에서 모델이 얼마나 Positive를 잘 분류하였는가?

$$Recall = \frac{True Positives}{True Positives + False Negatives}$$

source :<https://lawtomated.com/accuracy-precision-recall-and-f1-scores-for-lawyers/>

Reference

<http://dsba.korea.ac.kr/seminar/?mod=document&uid=1462>

<https://choice-life.tistory.com/40>

<https://ichi.pro/ko/tf-idfleul-sayonghan-leseutolang-libyue-daehan-gamjeong-bunlyu-73628754371661>

<https://arxiv.org/pdf/1911.03437v5.pdf>

<https://github.com/e9t/nsmc>

<https://huggingface.co/datasets/imdb>

<https://gluebenchmark.com/>