RNN 모델과 ELMO 모델을 이용한

스팸 문자 분류 및 분석

기본적이면서도 단순한 분류지만 다양한 모델을 활용하여 분석한 프로젝트

AI 07 강동우

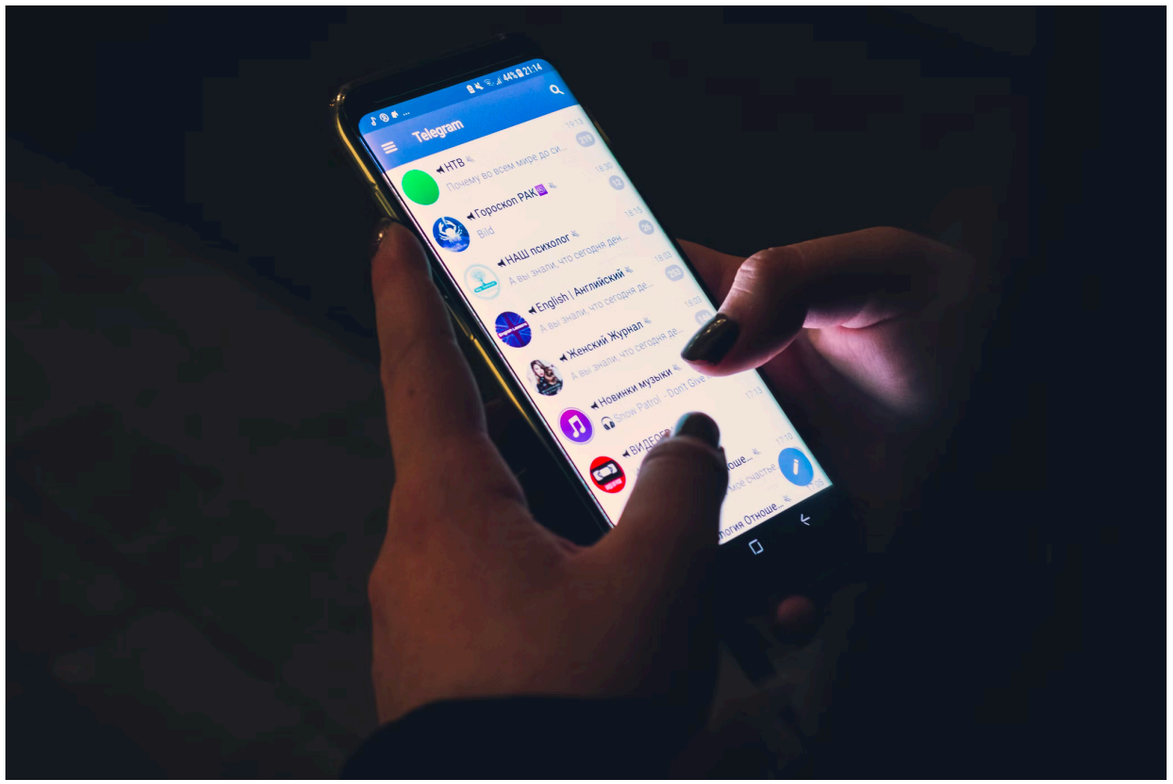
Context

목차 안내

1. Intro
2. Data Preprocessing
3. Visualization
4. Modeling
5. Outtro

Intro

프로젝트의 배경, 주제를 선정한 이유.



SMS Spam Collection Dataset

kaggle

Intro

프로젝트의 배경, 주제를 선정한 이유.

스팸 문자의 확실한 분류

SMS Spam Collection Dataset

kaggle

Intro

프로젝트의 배경, 주제를 선정한 이유.

다양한 모델을 활용하여
성능을 높일 수 있지 않을까?

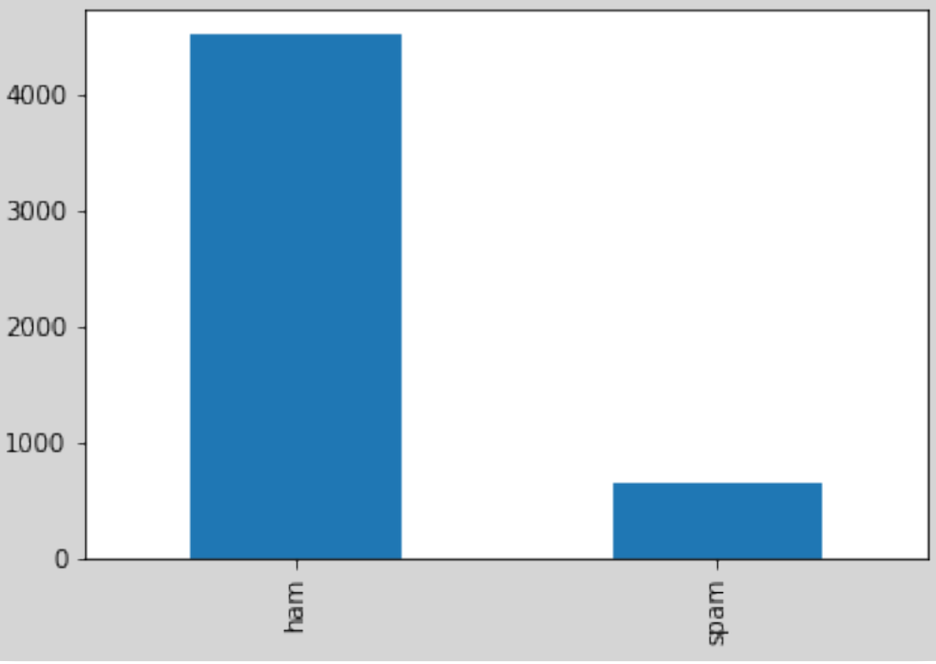
딥러닝 뿐만아니라 머신러닝을 이용하면
어떤 모델이 더 좋은 성능을 낼 수 있을까?

SMS Spam Collection Dataset

kaggle

Data Preprocessing

데이터 전처리



`isnull().values.any()`

Null 값이 있는 데이터 확인 후 제거

`.nunique()`

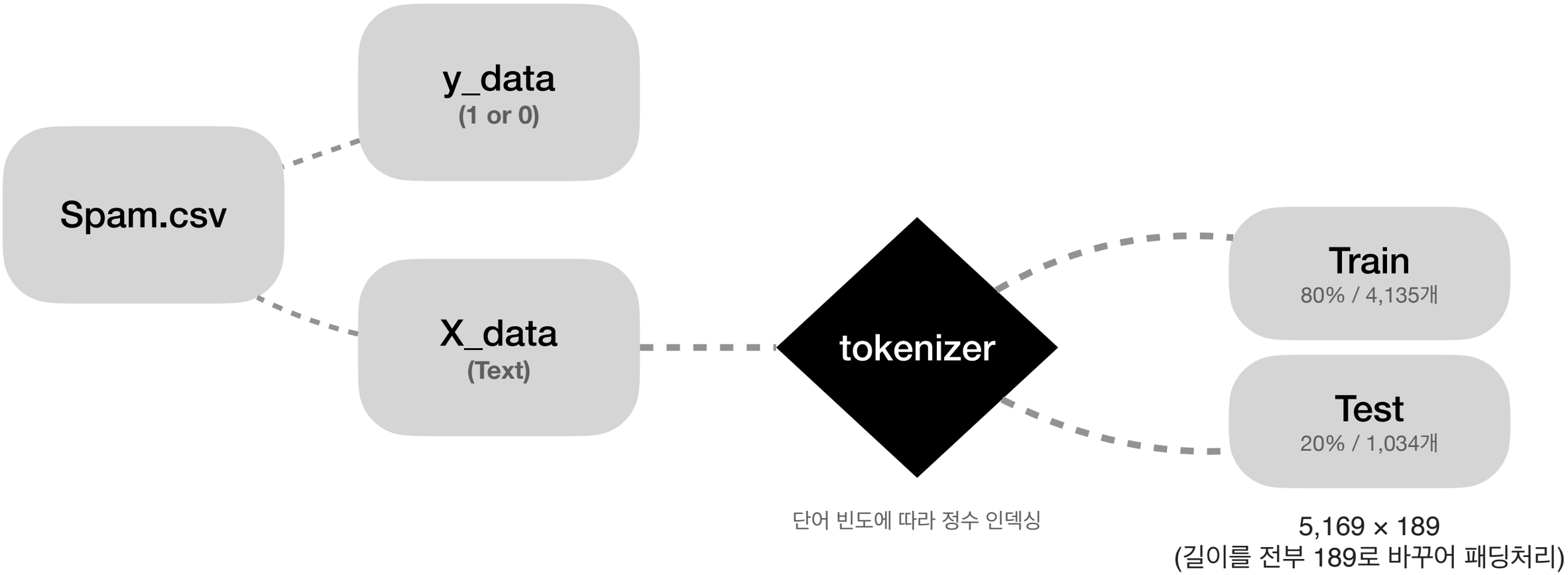
중복이 있는 데이터 제거

`Columns modify`

학습을 효율적으로 하기 위해 0과 1로 바꿈

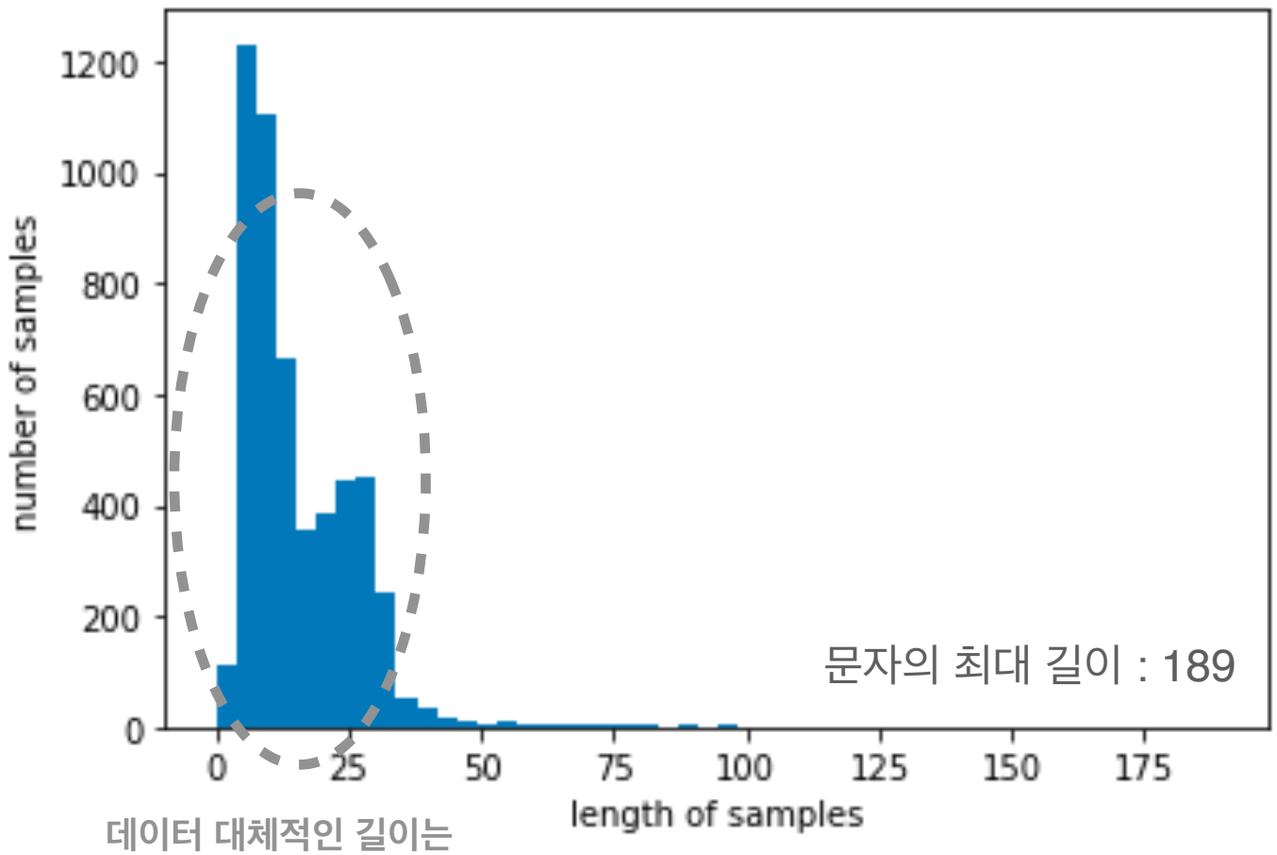
Data Preprocessing

데이터 전처리



Visualization

문자의 최대길이와 이에 따른 데이터 수 시각화

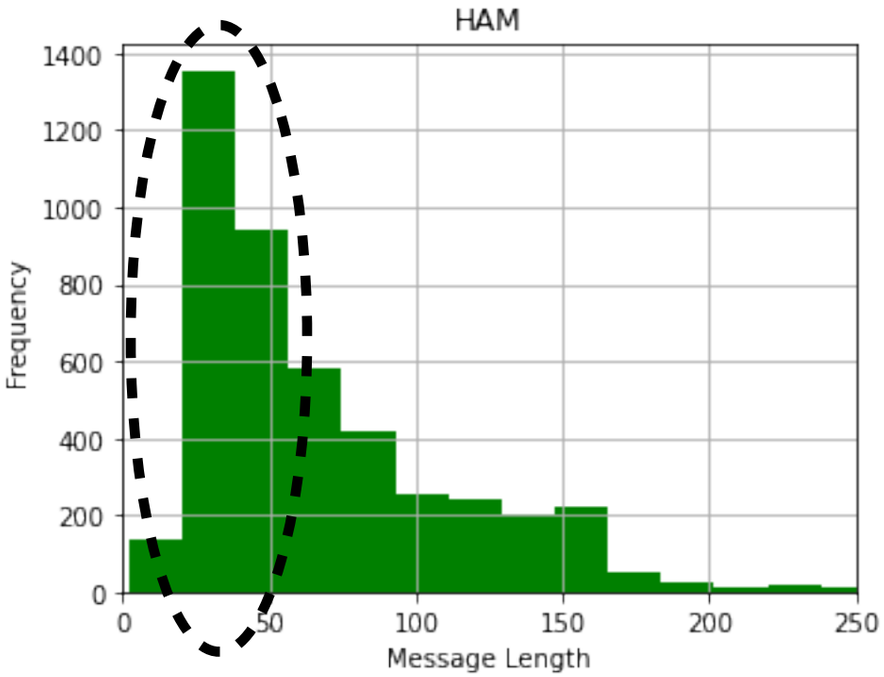


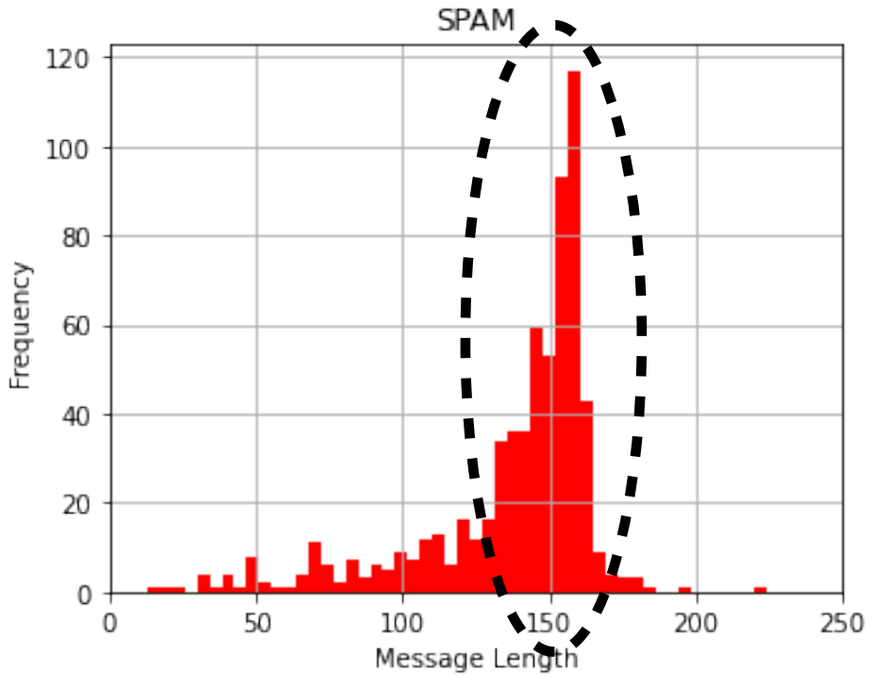
데이터 대체적인 길이는

50 이하라는 것을 알 수 있음.

Visualization

스팸문자와 비스팸 문자의 메세지 빈도수 시각화





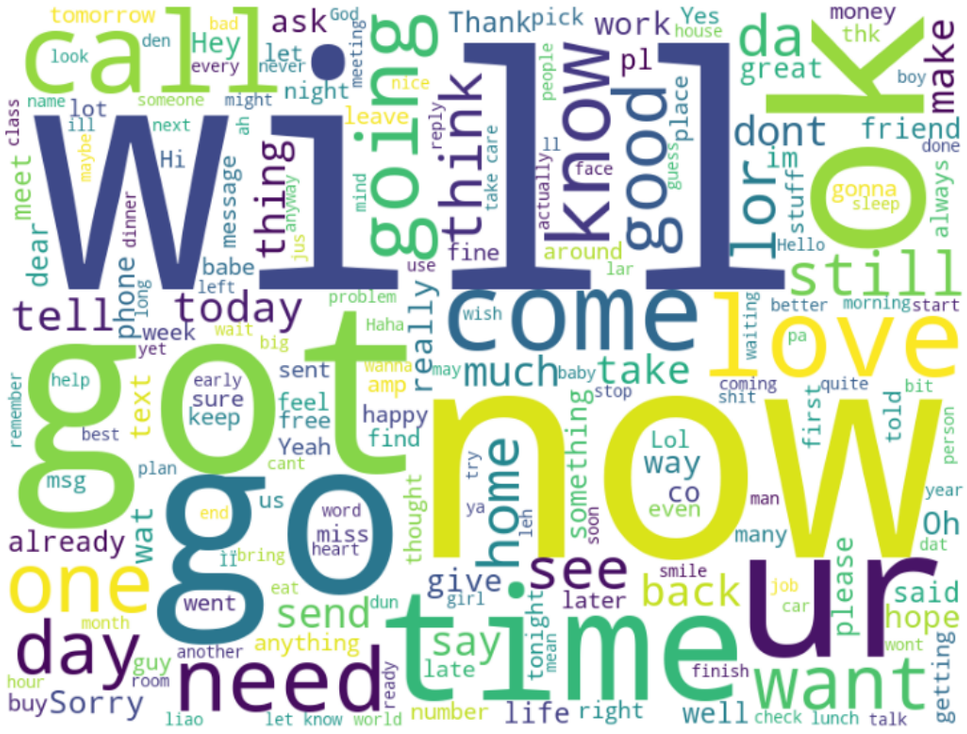
Visualization

전체 데이터 워드 클라우드 시각화

Visualization

전체 데이터 워드 클라우드 시각화

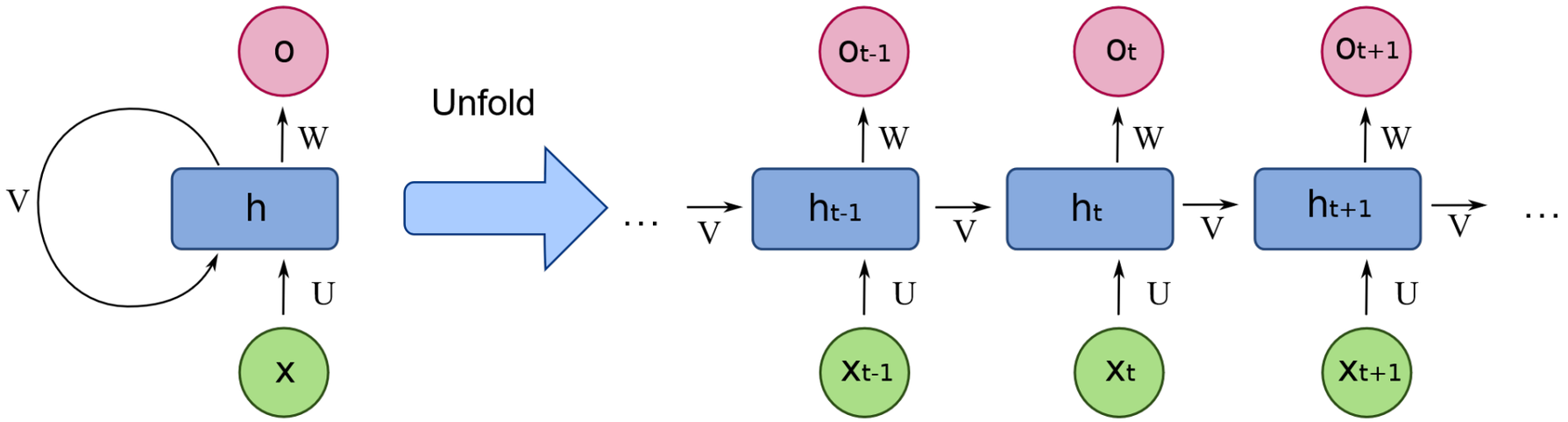
스팸 문자 데이터



일반 문자 데이터

Modeling_RNN

단순한 RNN모델(Vanila RNN)을 이용한 스팸 문자 분류



https://en.wikipedia.org/wiki/Recurrent_neural_network

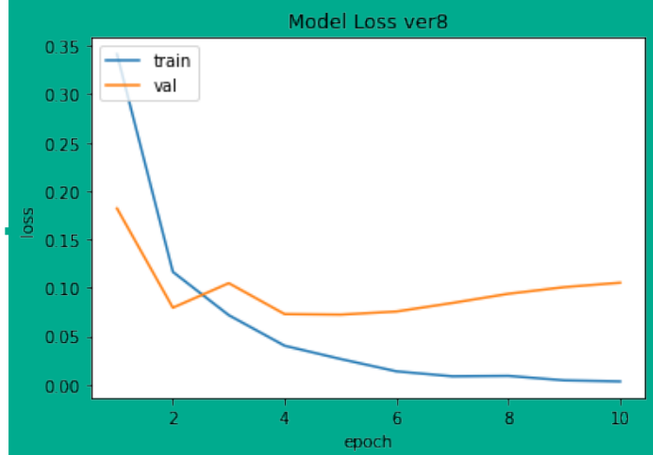
Modeling_RNN

단순한 RNN모델(Vanila RNN)을 이용한 스팸 문자 분류

Epoch	10
Batch-size	64
Layer	32-32
Embedding	32
Optimizer	rmsprop

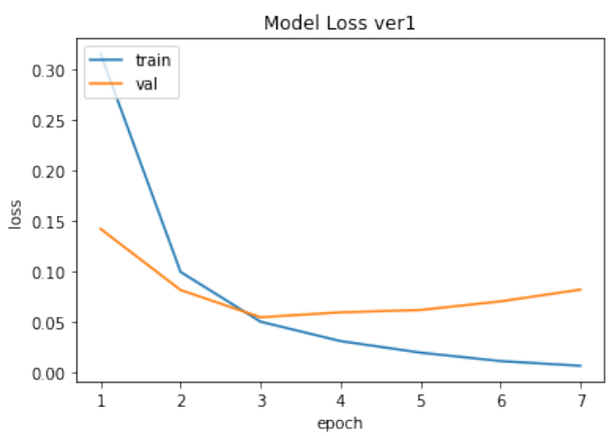
0.9816

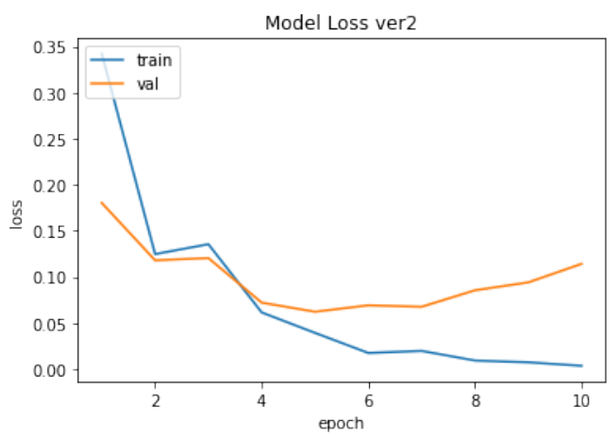
8번의 학습을 통한 최대 정확도

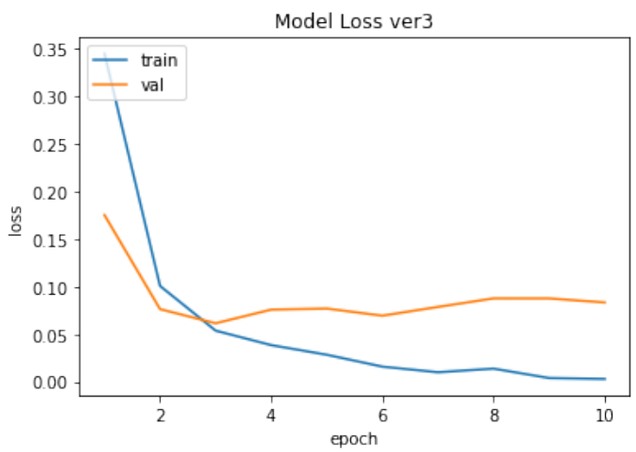


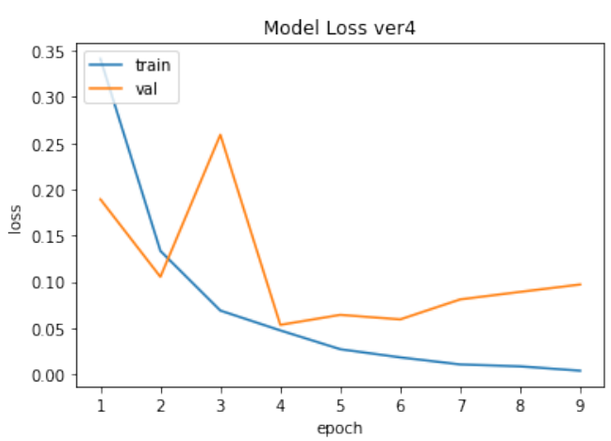
Modeling_RNN

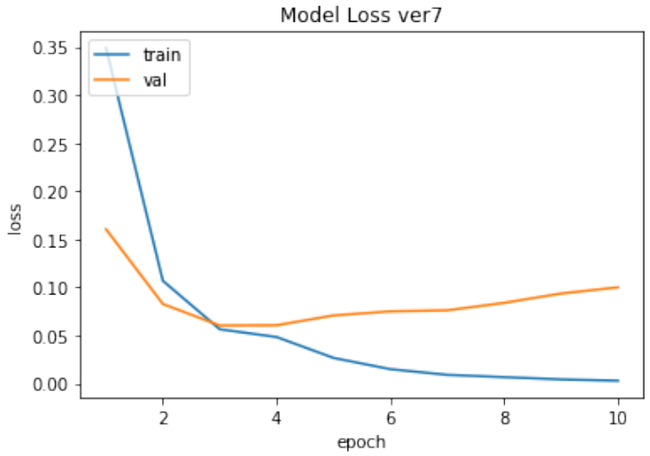
단순한 RNN모델(Vanila RNN)을 이용한 스팸 문자 분류

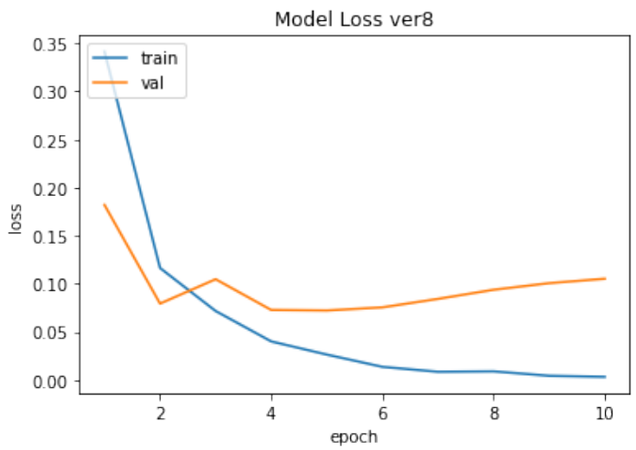












Modeling_Ensemble

Weak한 학습기 여러개로
강한 학습기 만듬

Boosting

0.932

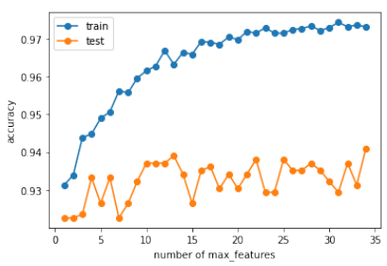
모델 정확도

1.1s

모델 시간 측정

n_estimator : 200
max_features : 30

하이퍼 파라미터



배깅을 사용한 결정트리
투표를 통해 예측

Random Forest

0.935

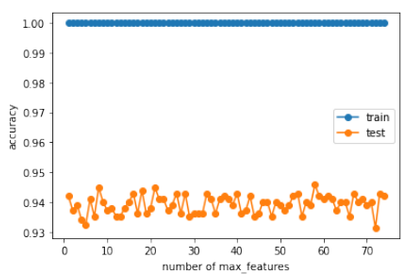
모델 정확도

1.2s

모델 시간 측정

n_estimator : 75
max_features : 55

하이퍼 파라미터



중복허용,샘플링

Bagging

0.926

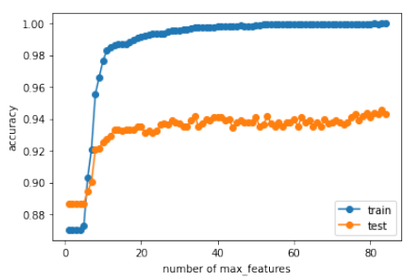
모델 정확도

0.6s

모델 시간 측정

n_estimator : 80
max_features : 15

하이퍼 파라미터



GBM 기반 수행시간 , 과적합 규제 개선

XGBoost

0.939

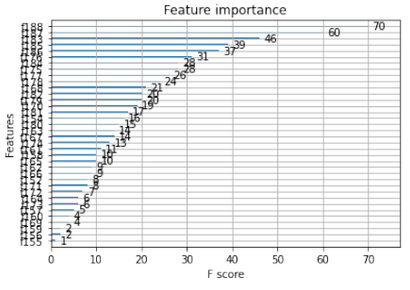
모델 정확도

43.7s

모델 시간 측정

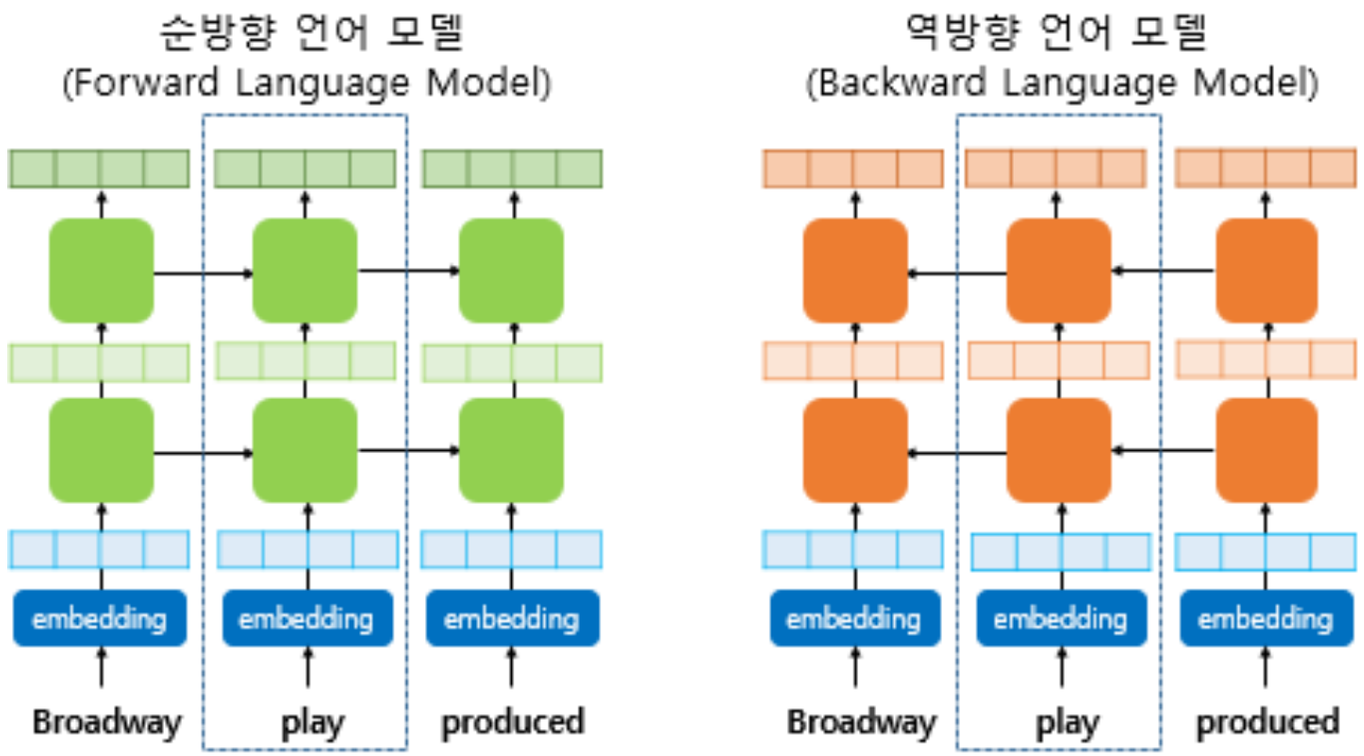
n_estimator : 200
max_features : 30

하이퍼 파라미터



Modeling_ELMo

단순한 ELMo 모델을 이용한 스팸 문자 분류



https://wikidocs.net/33930

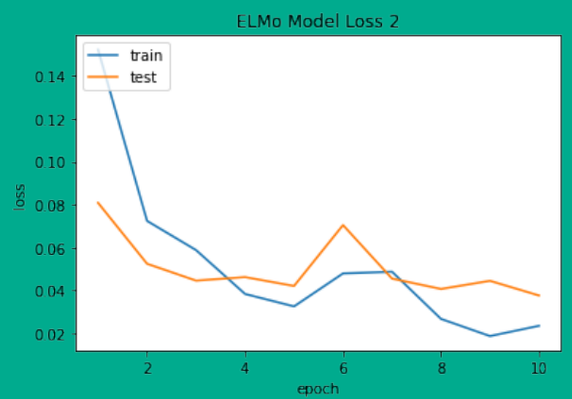
Modeling_ELMo

ELMo 모델을 이용한 스팸 문자 분류

Epoch	10
Batch-size	64
Layer	256
Optimizer	adam

0.9930

3번의 학습을 통한 최대 정확도



Outtro

회고

난제는 과적합

머신러닝과 딥러닝을 사용해본 결과

감사합니다.

참고 자료

https://www.kaggle.com/faressayah/natural-language-processing-nlp-for-beginners
https://wikidocs.net/22886
https://www.kaggle.com/andreshg/nlp-glove-bert-tf-idf-lstm-explained
https://velog.io/@changhtun1/ensemble