201702081 최재범 10주차 보고서

■ 설명

문장 데이터를 처리하기 위해, train 데이터와 test 데이터의 토큰화 과정을 거쳐 공통된 단어 사전을 구축하고, 단어 별 빈도수를 희소행렬로 표현하였다. 공통된 단어사전을 사용하였기 때문에, 특징값의 차원이 같게 나왔다.

이후 Logistic Regression 분류기를 사용하여 train 데이터를 학습시키고, test 데이터에 대하여 성능을 측정하였다.

■ 실행결과

```
훈련 데이터의 문서 수 :  25000
테스트 데이터의 문서 수  :  25000
클래스별 샘플 수 (훈련 데이터) :  [12500 12500]
클래스별 샘플 수 (테스트 데이터) :  [12500 12500]

TRAIN, TEST 의 총 특징 개수 (어휘사전의 크기) :  26966
X_TRAIN :
 <25000x26966 sparse matrix of type '<class 'numpy.int64'>'
    with 2149958 stored elements in Compressed Sparse Row format>
X_TEST :
 <25000x26966 sparse matrix of type '<class 'numpy.int64'>'
    with 2077852 stored elements in Compressed Sparse Row format>

훈련 세트 점수 : 0.997
테스트 세트 점수 : 0.855
```

train 데이터를 학습시켰기 때문에 그 기준으로 측정했을 땐 점수가 아주 높게 나왔고,

새로운 test 데이터 기준으로 측정했을 땐 상대적으로는 점수가 낮았지만 꽤 많이 적중하였다.