201702081 최재범 11주차 보고서

■ 설명

문장 데이터를 처리하기 위해, 단어사전을 구축하고 희소행렬로 표현하였다 (Bag of Words). 이 때 TF-IDF Vectorizer 를 사용하여, 다른 문서보다 특정 문서에 자주 나타나는 단어일 경우 높은 가중치를 줌으로써 연관성이 높은 것이 반영되도록 하였다. L2 정규화와 불용어 처리를 하기 위해 이를 파라미터로 넣었다.

이후 Linear SVM을 사용하여 5-cross validation 을 통해 성능을 측정하였다.

■ 실행결과

```
훈련 데이터의 문서 수 :  737
클래스별 샘플 수 :  [101 124 265 147 100]

교차 검증 평균 점수 : 0.9878
```

Fold를 5로 한 교차검증의 평균 점수는 위와 같이 나왔다.