3주차 패키지

지금부터 데이터셋을 만들어 분류 모델링을 진행하고자 합니다.

0. 데이터셋을 불러오세요.

1. 해당 데이터셋을 만드세요.

1-1) 단어가 등장했으면 1 아니면 0으로 만드는 count vectorization 데이터셋

힌트: CountVectorizer 함수를 이용해보세요

1-2) 단어의 등장에 가중치를 주는 Tfi-idf 데이터셋

힌트: TfidfVectorizer 함수를 이용해보세요

1-3) 데이터셋을 shuffle하여 20%의 test set을 만드세요

2. 시각화

2-1) 시각화를 통해 class의 분포를 확인하세요

3. 아래의 모델을 통해 다중분류 모데링을 진행하세요

위에서 만든 두 가지 데이터셋을 모두 사용하세요.

3-1) 다항분류 나이브베이즈 모형

Validation set을 만들어 alpda에 대한 튜닝을 진행하고 이를 시각화 하세요.

3-2) SVM 모델

3-3) KNN 모델

Validation set을 만들어 n_neighbors에 대한 튜닝을 진행하고 이를 시각화 하세요.

3-4) Test set을 통해 최고의 성능을 보이는 모델을 보이세요.