기계학습 기말 대체과제

**Wine Review 데이터에 대한 감성 분석**

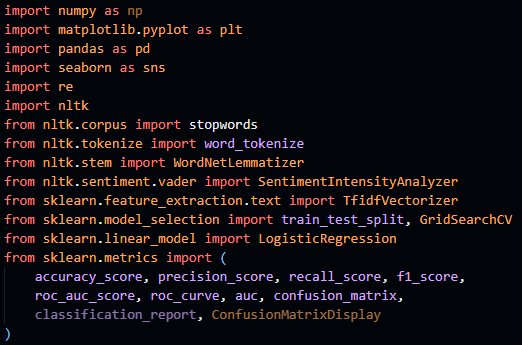
ICT융합공학부 소프트웨어전공

201820934

배지용

데이터: wine\_review.csv

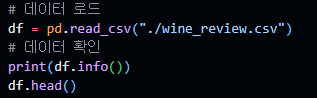
**필요한 라이브러리 import**

****

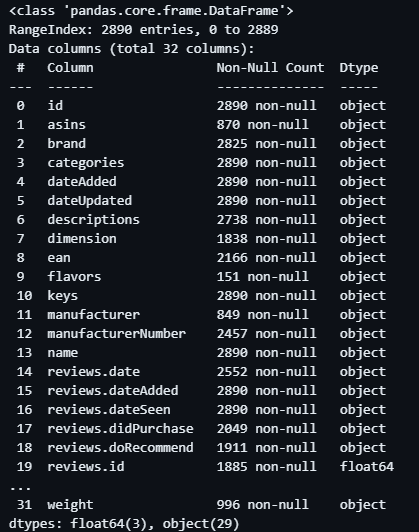
**데이터 로드&확인**

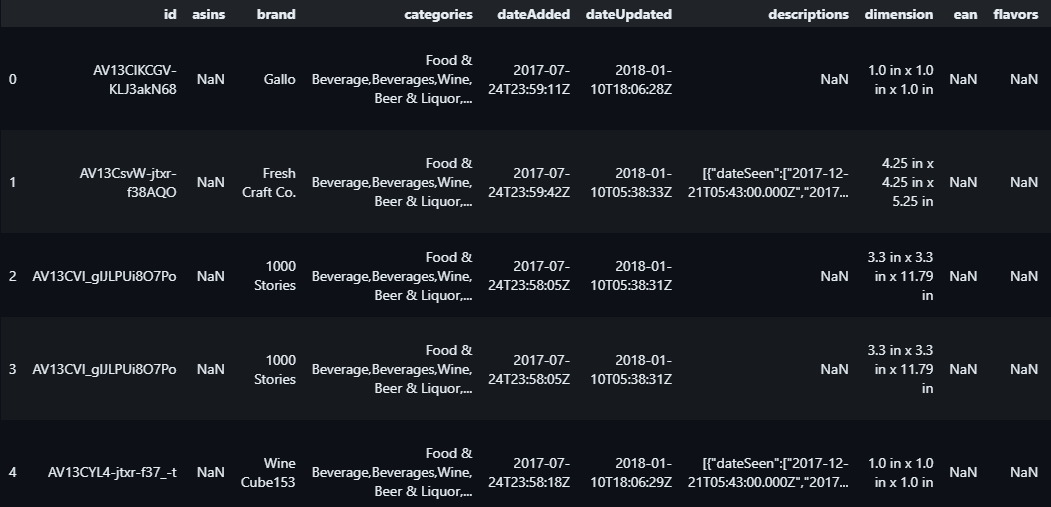
* **제공된 데이터를 로드 및 확인**

코드



결과



****

**데이터 전처리**

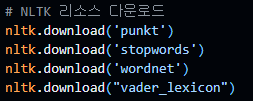
* 리뷰 제목 결측 값 빈 문자열로 대체
* 중복된 리뷰 텍스트 제거
* 리뷰 텍스트의 결측 값 제거

코드



**NLTK 리소스 다운로드**

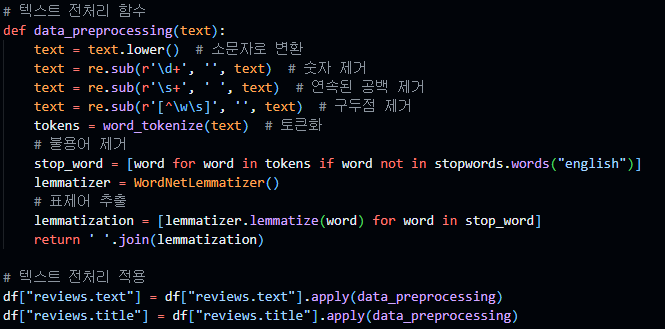
코드



**텍스트 전처리 함수(리뷰 텍스트, 리뷰 제목에 적용)**

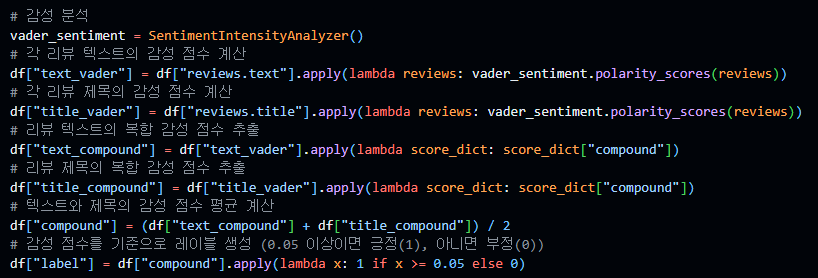
* 소문자로 변환
* 숫자, 연속된 공백, 구두점 제거
* 토큰화
* 불용어 제거
* 표제어 추출
* 리뷰 텍스트와 리뷰 제목에 적용

코드



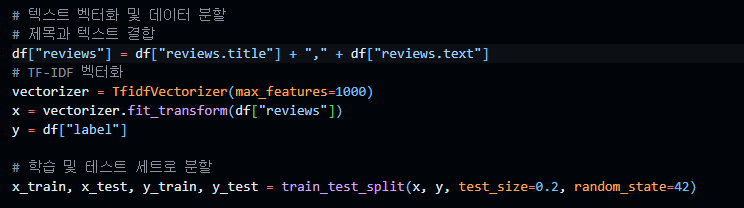
감성분석

* 각 리뷰 텍스트, 리뷰 제목 감성 점수 계산
* 리뷰 텍스트, 리뷰제목 복합 감성 점수 추출
* 텍스트와 제목 감성 점수 평균 계산
* 감성 점수 기준으로 레이블 생성



텍스트 벡터화 및 데이터 분할

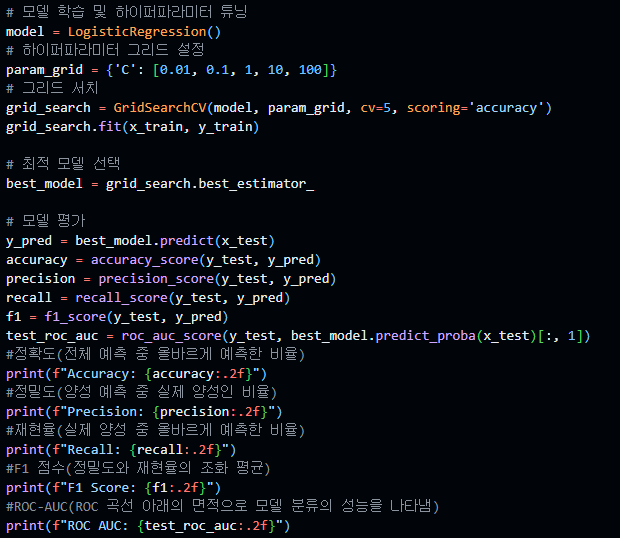
* 제목과 텍스트 결합
* 학습 및 테스트 세트로 분할 (20%)



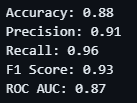
모델학습 및 하이퍼파라미터 튜닝(그리드 서치 사용)

* 하이퍼파라미터 그리드 설정
* 그리드 서치
* 최적 모델 선택
* 모델 평가 (정확도, 정밀도, 재현율, F1 점수, ROC-AUC)

코드



**결과**

****

정확도 (Accuracy): 모델이 정확히 예측한 비율

정밀도 (Precision): 모델이 Positive로 예측한 것 중에서 실제 Positive인 비율

재현율 (Recall): 실제 Positive인 것 중에서 모델이 Positive로 예측한 비율.

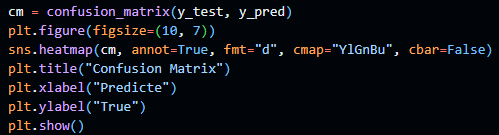
F1 score: 모델의 정밀도(Precision)와 재현율(Recall)의 조화 평균을 나타내는 지표

ROC-AUC : ROC곡선 아래의 면적을 의미하며, 이는 모델의 분류 성능을 나타내는 지표 중 하나

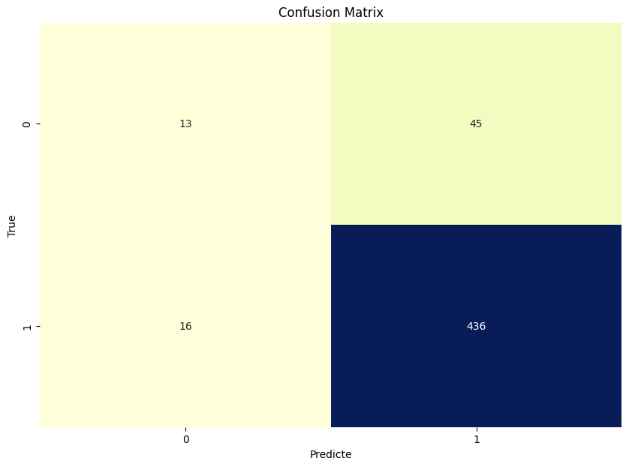
정확도, 정밀도, 재현율, F1점수, ROC-AUC를 봤을 때 전체적으로 수치가 높게 형성 돼있어 모델이 예측하고 분류하는 성능은 준수한 편으로 보인다.

**혼동 행렬**

코드



결과



위의 혼동 행렬을 보고 분석할 수 있는 내용은 아래와 같다.

 436명이 실제로 Positive이고 모델이 이를 정확히 Positive로 예측함. (True Positive).

 16명이 실제로 Positive이지만 모델이 이들을 Negative로 잘못 예측함. (False Negative).

 13명이 실제로 Negative이고 모델이 이를 정확히 Negative로 예측함. (True Negative).

 45명이 실제로 Negative이지만 모델이 이들을 Positive로 잘못 예측함. (False Positive).

잘못 예측한 것이 총 61개 정확히 예측한 것이 총 449개

TP = 436

FN = 16

TN = 13

FP = 45

 정확도 **(Accuracy)**: 모델이 정확히 예측한 비율.

Accuracy = TP+TN/TP+TN+FP+FN = 436+13/436+13+45+16 = 약0.88

 정밀도 **(Precision)**: 모델이 Positive로 예측한 것 중에서 실제 Positive인 비율

Precision = TP/TP+FP = 436/436+45 = 약0.91

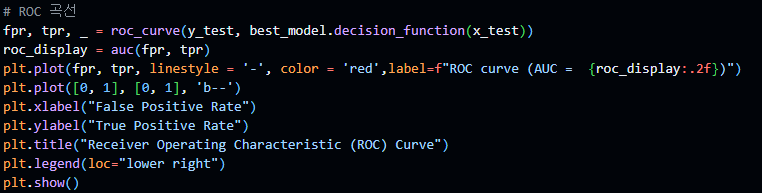
 재현율 **(Recall)**: 실제 Positive인 것 중에서 모델이 Positive로 예측한 비율. Recall = TP/TP+FN = 436/436+16 = 약0.96

F1 score: 모델의 정밀도(Precision)와 재현율(Recall)의 조화 평균을 나타내는 지표

F1 score = 2×(정밀도\*재현율)/(정밀도+재현율)​ = 약 0.93임을 계산할 수 있다.

**ROC 곡선**

코드



결과

텍스트, 라인, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

ROC곡선의 AUC(곡선 아래 면적)은 0.5~1의 값을 갖고 1에 가까울수록 분류 성능이 높음을 나타낸다. 해당 모델은 ROC-AUC가 0.87로 분류 성능이 준수한 것을 확인할 수 있다.