DACON 수상작리뷰
이동욱00 | Private 1위 (점수 : 1) | Resnet50 & 10-Fold & Adam - DACON

[영화 리뷰 감성 분석 AI 해커톤]

◈ 주제 및 문제점

👉 네이버 영화 리뷰 데이터를 통해 긍정/부정 분류

- 평가기준: accuracy

◈ 데이터

- train.csv : id, document(리뷰 내용), label (target)
- test.csv : id, document(리뷰 내용)
- sample_submission.csv : id, label (target)

◈ 코드 리뷰

(1) 데이터, 패키지 로드

🔷 import rich : 출력을 보기 좋게 꾸며주는 라이브러리

(2) 테스트 정규화

🔷 두칸 이상 공백 → 한 칸 공백 변환 / str.replace

🔷 정규 표현식

- [^A-Za-z가-힣]
  영문 대소문자, 한글, 공백을 제외한 모든 문자"를 찾음.

(3) 피처 벡터화 & 모델 학습

🔷 **Pipleline 사용**

- TF-IDF 기반 벡터화
- K-FOLD 사용
- 사용한 모델: 나이브베이즈, SGD, RF, SVC, ADA, LGBM 2개, XGB, KNN 2개 👇 각 모델의 정확도

Model Comparison Table

Model Name	Accuracy
naive_bayes	0.884
SGD	0.870
rfc	0.826
SVC	0.862
ada	0.769
lgbm	0.839
lgbm2	0.836
lgbm3	0.837
xgb	0.783
knc1	0.801
knc2	0.793

🔷 Pipeline 후, **스태킹** 적용

- from sklearn.ensemble import StackingClassifier

```
stack_models = [(name, get_pipe(model, name)) for name, model in models]

stacking = StackingClassifier(stack_models)
acc = return_kfold_accuarcy(stacking)
rich.print(acc)
```

◈ **배울점**

- rich 라는 출력에 사용되는 새로운 패키지를 알게됨.
- 과제로 배운 텍스트 분석을 실제로 데이터에 적용하는 과정을 알게되었음.
- 텍스트 전처리를 많이 하는것보다 간단하게 하는 것이 성능이 더 좋을 수도 있음.