**기말 대체 과제 보고서**

**[기계학습]**



**학번 : 201920831**

**이름: 김건호**

**1. 감성 분석의 기본 이해**

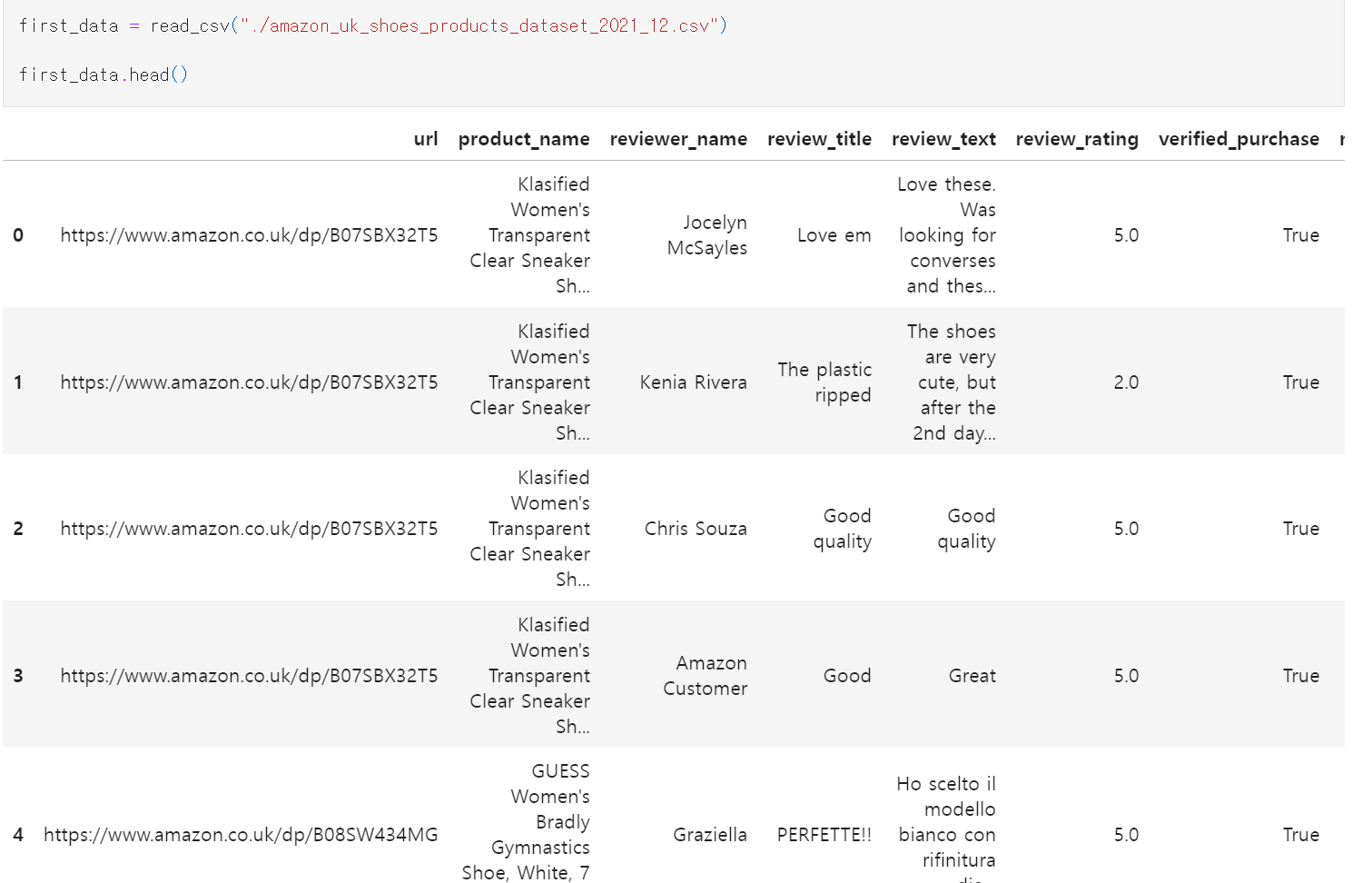
감정 분석은 텍스트 데이터에서 감정, 의견, 또는 태도를 식별하고 추출하는 것입니다. 감정분석은 자연어 처리(NLP)의 한 분야로, 주로 사람들이 작성한 텍스트에서 긍정적, 부정적, 중립적 감정을 자동으로 감지합니다.

**2. 데이터 전처리**

학번은 201920831입니다.

amazon\_uk\_shoes\_products\_dataset\_2021\_12.csv 데이터를 활용하여 진행합니다.

1) 데이터 가져오기 및 확인



2) 데이터 파악

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

데이터의 Columns과 내용을 보고 데이터를 파악합니다.

3) 선택 및 결측값 제거 텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

필요하다고 생각되는 데이터만 남기고 결측값을 제거합니다.

4) 전처리 함수텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

NLTK를 사용해 전처리 작업 수행.

노이즈를 제거하기 위한 함수들을 사용하였습니다.

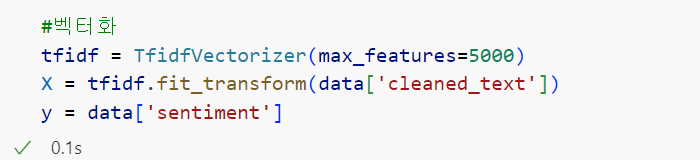
5) 감성 레이블 생성텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

리뷰 평점을 기준으로 3 이상 긍정 < 1 >, 이하 부정 < 0 > 으로 분류합니다

**3. 모델 학습**

1) 데이터 벡터화



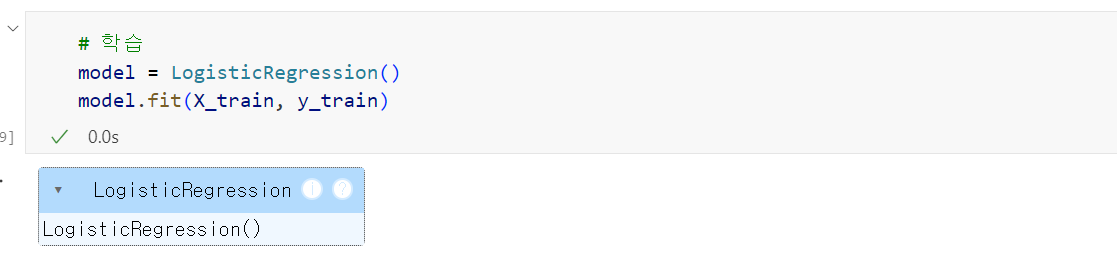
단어의 중요도를 계산해 텍스트 데이터를 숫자로 바꿉니다.

2) 데이터 분할



데이터셋을 학습과 테스트로 분할합니다.

3) 학습



로지스특 회귀 모델을 사용하여 모델을 학습합니다.

4) 예측텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

아까 분할한 테스트로 모델이 감성을 예측하도록 합니다.

5) 그리드텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

그리드 서치를 사용하여 하이퍼 파라미터 튜닝을 진행합니다.

교차 검증은 5로 설정하였습니다. 그리드 서치가 선택한 최적의 하이퍼 파라미터는 c = 10이었습니다.

모델의 정확도는 86.8% 이었습니다.

6) 최적의 학습

텍스트, 폰트, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명

최적의 모델을 사용해 학습을 진행했습니다.

**4. 평가**

텍스트, 폰트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

1. Accuracy (정확도): 0.86의 수치로, 모델이 상당수의 경우 올바르게 예측했다는 것을 알 수

있습니다.

1. Precision (정밀도): 0.89의 수치로, 모델이 긍정을 예측할 때 대부분 올바르게 예측했다는 것을 알 수 있습니다.
2. Recall (재현율): 0.95의 수치로, 모델이 긍정을 올바르게 예측했습니다.
3. F1 Score: 0.92의 수치로, 모델이 긍정과 부정 예측 모두에서 균형 잡힌 성능을 보이고 있다는 것을 알 수 있습니다.
4. ROC AUC (Positive Reviews): 0.87로, 긍정 리뷰에 대한 예측 성능이 우수했습니다.

**5. 시각화 및 해석**

1) Confusion

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명

긍정리뷰에서는 0.96으로 긍정으로 예측을 했지만

부정리뷰에서는 0.4로 부정리뷰를 부정으로 예측한 모습을 보입니다.

2) ROC텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명라인, 텍스트, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명

ROC 곡선은 대각선 곡선보다 위에 있으면 성능이 우수한 것인데,

역시 부정적인 리뷰를 분류하는 데 있어서 문제가 있었는지, 좋은 분류기 보다 기울기가 낮다고 생각되는 모습입니다.

**6. 결과**

F1 Score가 높게 나와서 기대했지만 부정리뷰 부분이 상당히 잘못 예측하고 있는 모습을 보이고 있습니다.

이러한 문제는 여러가지 이유가 있는데 그 중 오분류가 문제라고 생각됩니다.

1. 별점 3점 이상을 줬지만 부정적인 리뷰가 있었던 것

2. 사람마다 별점에 대한 기준이 다른 것 (3점이 낮다고 생각, 3점이면 높다고 생각)

그렇기 때문에 정확한 분류가 모델의 성능을 향상시키는데 도움이 될 것입니다