**기계학습 보고서**

**제품 review 데이터의**

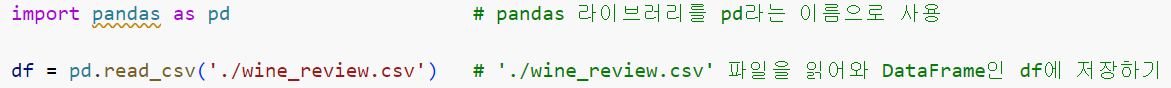
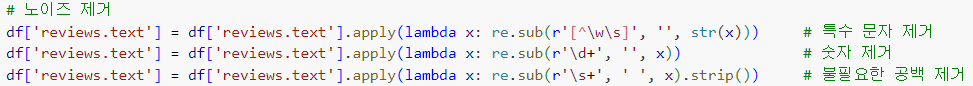
**감성 분석**

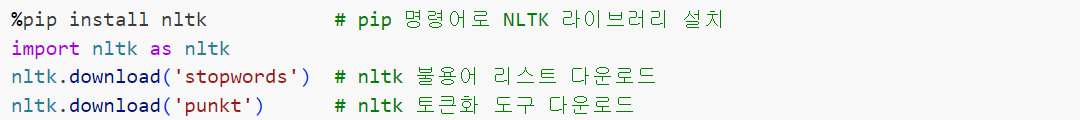
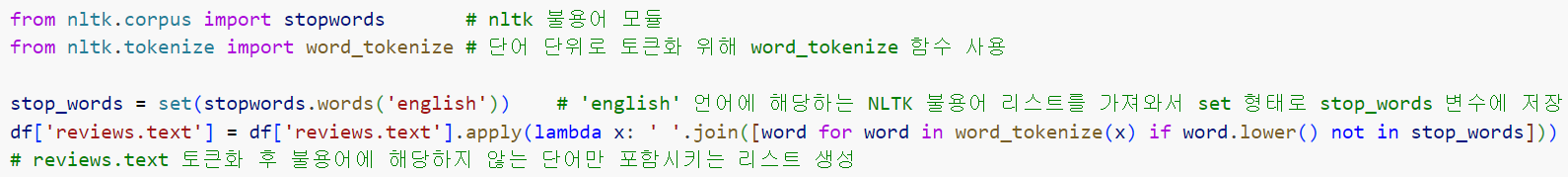
텍스트, 로고, 엠블럼, 원이(가) 표시된 사진

자동 생성된 설명

|  |  |
| --- | --- |
| 학과 | 컴퓨터공학과 |
| 학번 | 202121856 |
| 이름 | 오예나 |
| 제출일 | 2024. 06. 15 |

**감성 분석의 기본 이해**  
감성 분석(Sentiment Analysis)이란 텍스트에 들어있는 의견이나 감성, 평가, 태도 등의 주관적인 정보를 컴퓨터를 통해 분석하는 과정이다. 여기서 얘기하는 "텍스트"는 사람의 의견이 담긴 소비자 리뷰, 설문조사 응답, 채팅 등이 될 수 있고, 그 의견이 긍정적인지 부정적인지 혹은 중립적인지를 판단하는 것이다. 입력은 텍스트이며, 출력은 감정으로 보통 긍정/부정/중립 중 하나로 선택될 수 있다. 더 세밀한 분류를 위해 스케일을 갖는 숫자가 될 수도 있다. 예를 들어 1에서 5까지 5점 척도를 갖고 숫자가 커지면 긍정이라고 해석하는 것.

1. **데이터 전처리**  
   - 라이브러리 설치 후 데이터 로드  
     
     
   - 데이터를 확인해 구조 확인하기  
     
     
   - 데이터가 방대해 중간에 생략되어 나왔기 때문에 열들을 다시 확인  
     
     
   - 데이터 내의 빈 데이터를 없애는 과정 – 결측치 제거  
     
     
   - 정규식을 활용해서 노이즈 제거  
     
     
   - 노이즈 제거 후 결과  
   텍스트, 스크린샷, 폰트, 문서이(가) 표시된 사진

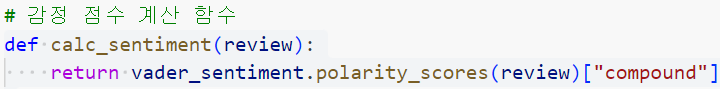
   자동 생성된 설명  
     
   - 텍스트 불용화 및 토큰화  
     
   

‘id': 식별자 / 'asins': ASIN 번호 / 'brand': 브랜드 / 'categories': 카테고리 / 'dateAdded': 추가된 날짜 / 'dateUpdated': 업데이트된 날짜 / 'descriptions': 설명 / 'dimension': 차원 / 'ean': EAN (유럽 제품 번호) / 'flavors': 맛 / 'keys': 키 / 'manufacturer': 제조사 / 'manufacturerNumber': 제조사 번호 / 'name': 이름 / 'reviews.date': 리뷰 작성 날짜 / 'reviews.dateAdded': 리뷰 추가된 날짜 / 'reviews.dateSeen': 리뷰 확인된 날짜 / 'reviews.didPurchase': 구매 여부 / 'reviews.doRecommend': 추천 여부 / 'reviews.id': 리뷰 ID / 'reviews.numHelpful': 도움이 된 수 / 'reviews.rating': 리뷰 평점 / 'reviews.sourceURLs': 리뷰 소스 URL / 'reviews.text': 리뷰 텍스트 / 'reviews.title': 리뷰 제목 / 'reviews.userCity': 사용자 도시 / 'reviews.userProvince': 사용자 지방 / 'reviews.username': 사용자 이름 / 'sizes': 사이즈 / 'sourceURLs': 소스 URL / 'upc': UPC (미국 제품 코드) / 'weight': 무게

불용화 및 토큰화를 진행하기 위해 nltk를 활용한다.  
영어를 기반으로 감정 분석을 진행하기 때문에 English에 해당하는 불용화 리스트 사용  
x: 각 행의 텍스트 데이터  
word\_tokenize(x): 텍스트를 단어 단위로 토큰화하여 리스트로 반환  
[word for word in word\_tokenize(x) if word.lower() not in stop\_words]: 토큰화 된 각 단어에 대해 소문자로 변환한 후, 불용어(stop\_words)에 포함되지 않는 단어들로 이루어진 리스트를 생성   
' '.join(...): 생성된 리스트를 다시 공백을 기준으로 하나의 문자열 합침  
df['reviews.text'].apply(...): 이 과정을 모든 행에 적용하여 'reviews.text' 열을 업데이트  
  
- NLTK를 사용하여 텍스트 데이터에서 스테밍(stemming)을 적용하는 작업을 수행  
텍스트, 폰트, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명  
스테밍: 어형이 변형된 단어로부터 접사 등을 제거하고 그 단어의 어간을 분리해 내는 것을 의미  
PorterStemmer = 스테밍, WordNetLemmatize = 표제어 추출도구 nltk.download('wordnet')는 표제어 추출을 위해 필요한 데이터를 다운로드  
과정을 거치면 리뷰 텍스트의 모든 단어들이 기본 형태로 바뀌어서, "running", "runs" 같은 단어들이 모두 "run"으로 통일되어 컴퓨터가 분석하기 더 쉬워짐  
  
- 표제어 추출을 통해 텍스트 정규화  
  
리뷰 텍스트를 표제어 추출해서 단어들을 기본 사전 형태로 바꾸는 작업을 진행

1. **모델 선택 및 학습**   
   - 감정 분석(Sentiment Analysis)은 텍스트에 담긴 감정(긍정, 부정, 중립)을 자동으로 판단  
   - VADER는 특히 소셜 미디어 텍스트 같은 짧고 간단한 문장에서 감정을 분석하는 데 효과적이고, 단어의 긍정적, 부정적, 중립적 점수를 계산해준다.  
   텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

   자동 생성된 설명  
   SentimentIntensityAnalyzer는 VADER 감정 분석을 도와주는 도구, nltk.download('vader\_lexicon')는 VADER가 사용하는 감정 사전, VADER 감정 분석기를 초기화 후 도구를 사용해서 텍스트의 감정을 분석  
     
     
   calc\_sentiment라는 함수는 리뷰 텍스트를 받아서 VADER 감정 분석기를 사용해 감정 점수를 계산  
   vader\_sentiment.polarity\_scores(review): 리뷰 텍스트의 감정 점수 계산  
   ["compound"]: 그 중에서 종합적인 감정 점수인 compound 점수를 가져옴. 이 점수는 -1에서 1 사이의 값으로, -1은 매우 부정적, 1은 매우 긍정적을 의미  
     
     
   apply(calc\_sentiment): 각 리뷰 텍스트에 대해 calc\_sentiment 함수를 적용해 감정 점수 계산, 계산된 감정 점수를 sentiment\_score라는 새로운 컬럼에 추가

- 감정 점수 해석  
텍스트, 폰트, 스크린샷이(가) 표시된 사진

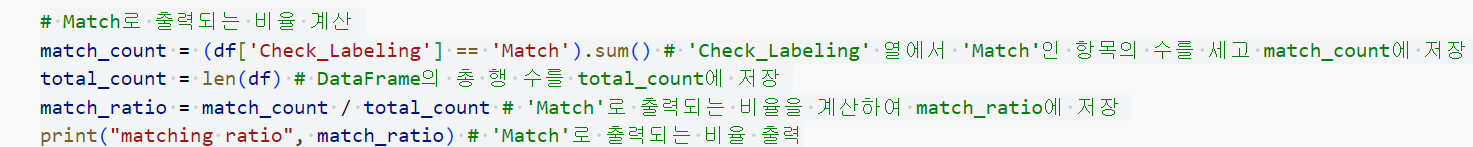
자동 생성된 설명

sentiment\_score 값: 이 값은 -1에서 1 사이의 값으로, 텍스트의 감정 강도를 나타냅니다. 1에 가까울수록 긍정적인 감정을, -1에 가까울수록 부정적인 감정을 의미합니다. 0에 가까운 값은 중립적 감정을 나타냅니다.

* 첫 번째 리뷰: fantast white wine occa  
  감정 점수: 0.0000  
  해석: 감정 분석기는 이 리뷰를 중립적인 텍스트로 판단. 단어들이 충분히 긍정적인 감정을 표현하지 않아서 점수가 0으로 나왔음
* 세 번째 리뷰: given wine delight surpri find flavor delici tannin balanc easi finish  
  감정 점수: 0.5994  
  해석: 이 리뷰는 긍정적인 감정을 가지고 있다. "delight", "surpri", "delici", "balanc", "easi" 같은 긍정적인 단어들이 포함되어 있어 높은 점수를 받았음

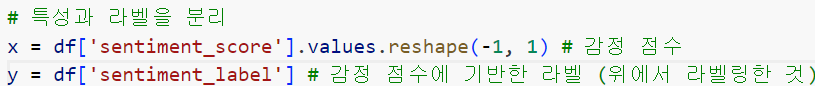
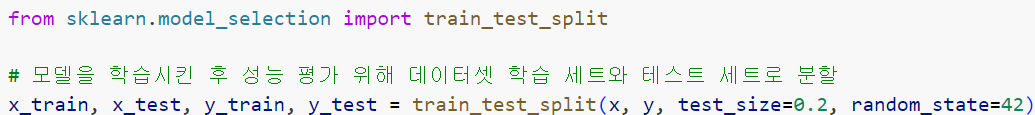
- 함수 정의: 리뷰의 별점과 라벨링된 감정 점수가 일치하는지 확인하는 함수  
텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명  
review\_rating: 리뷰의 별점 (1~5) / sentiment\_label: 감정 점수에 따라 라벨링된 값 (긍정: 1, 부정: 0)  
리뷰의 별점이 2 이상이면 긍정적(1), 2 미만이면 부정적(0)이라고 가정하고, 이와 감정 점수가 일치하는지 확인  
리뷰 별점이 긍정적이고 감정 점수도 긍정적이면 "Match", 부정적이고 감정 점수도 부정적이면 "Match", 그렇지 않으면 "Mismatch"를 반환  
  
  
  
- 감정 점수 라벨링 및 일치 여부 확인: 감정 점수에 따라 라벨링  
  
감정 점수가 0 이상이면 긍정적(1), 0 미만이면 부정적(0)으로 라벨링  
  
- 감정 점수 라벨링 및 일치 여부 확인: 리뷰 별점과 라벨링된 감정 점수 비교  
  
각 리뷰에 대해 check\_labeling 함수를 적용해서 리뷰 별점과 라벨링된 감정 점수가 일치하는지 확인하고, 그 결과를 Check\_Labeling 열에 저장  
  
- 결과 출력 및 일치 비율 계산: 결과 출력  
  
텍스트, 스크린샷, 번호, 영수증이(가) 표시된 사진

자동 생성된 설명  
  
- 해석  
  
matching ratio 0.7674740484429066  
**높은 일치 비율**: 76.7%의 리뷰가 일치하는 것은 감정 점수가 리뷰 별점과 상당히 일치한다는 것을 보여줍니다.  
이는 VADER 감정 분석기가 리뷰 텍스트에서 긍정적 또는 부정적 감정을 잘 파악하고 있다는 것을 의미할 수 있습니다.  
**일치하지 않는 비율**: 약 23.3%의 리뷰에서 일치하지 않는 부분은 감정 분석기가 텍스트의 뉘앙스를 완전히 파악하지 못하거나 리뷰어가 별점을 매길 때 텍스트와 일치하지 않는 감정을 표현했을 가능성을 나타냅니다.  
예를 들어, 리뷰어가 긍정적인 단어를 많이 사용했지만 최종적으로 부정적인 별점을 주는 경우가 있을 수 있다.

- 불일치 항목 (Mismatch) 분석  
텍스트, 스크린샷, 메뉴, 번호이(가) 표시된 사진

자동 생성된 설명  
**Review 16**: 별점: 2.0, 감정 점수: -0.0829 (부정)  
내용: "im sure bad bottl intent made effervesc dont want bubbl drink pinot grigio goe well mozzarella string chees cinemasin"  
해석: 리뷰어는 와인이 좋지 않았다고 표현했지만, 별점은 2점을 주었습니다. 감정 분석기는 이를 부정적으로 인식했습니다.  
**Review 69**: 별점: 1.0, 감정 점수: 0.0000 (중립)  
내용: "tast either old spoil stick bottl"  
해석: 리뷰어는 와인이 상했거나 오래된 것 같다고 표현했습니다. 그러나 감정 점수는 중립으로 나왔습니다.  
  
**전체적인 경향**감정 분석기(VADER)는 텍스트에서 부정적인 단어를 찾는데, 리뷰어의 의도와 다르게 부정적으로 인식할 수 있습니다.  
리뷰어가 별점을 줄 때 감정 분석 결과와 다르게 매길 수 있습니다. 예를 들어, 텍스트는 긍정적이지만 리뷰어가 별점을 낮게 줄 수 있습니다.  
일부 리뷰에서는 감정 분석기가 텍스트의 뉘앙스를 제대로 파악하지 못할 수 있습니다.

1. **평가:** 적절한 지표를 사용하여 모델의 성능을 평가  
   - 머신러닝 모델을 훈련시키기 위한 준비 작업  
     
   x 변수: sentiment\_score 열의 값을 가져와서 모델의 입력 데이터로 사용 .  
   이 값을 2차원 배열로 변환하기 위해 reshape(-1, 1)을 사용합니다. 이는 Scikit-learn과 같은 머신러닝 라이브러리에서 요구하는 형식입니다.  
   y 변수: sentiment\_label 열의 값을 가져와서 모델의 타겟 데이터로 사용합니다. 이 값은 감정 점수에 기반하여 라벨링한 값입니다.  
     
     
   train\_test\_split 함수를 사용하여 데이터를 학습 세트와 테스트 세트로 분할  
   데이터셋을 80% 학습 데이터와 20% 테스트 데이터로 나누고, random\_state를 42로 설정하여 데이터 분할 시 일관된 결과를 얻도록 설정  
     
     
   LogisticRegression 모델을 사용하여 하이퍼파라미터 튜닝을 하기 위해 GridSearchCV를 사용  
   GridSearchCV는 하이퍼파라미터의 조합을 시도하여 최적의 모델을 찾는 방법  
     
   텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

   자동 생성된 설명  
   로지스틱 회귀 모델을 초기화하고, GridSearchCV를 사용하여 주어진 하이퍼파라미터 그리드(param\_grid)를 기반으로 모델을 학습하며, 5-겹 교차 검증을 수행  
     
   - 그리드 서치를 통해 찾은 최적의 하이퍼파라미터를 출력  
     
     
   그리드 서치를 통해 찾은 최적의 하이퍼파라미터는 C=100입니다. 이는 로지스틱 회귀 모델에서 규제 강도를 조절하는 하이퍼파라미터입니다. C 값이 클수록 모델은 데이터에 더 적합하게 학습되지만, 과적합의 위험도 증가할 수 있습니다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명텍스트, 폰트, 영수증, 화이트이(가) 표시된 사진

자동 생성된 설명  
**종합적인 해석**모델은 전반적으로 매우 높은 정확도와 재현률을 보여줍니다. 특히 재현률이 1.0으로 완벽하게 실제 긍정을 예측하는 것을 보여줍니다.  
정밀도 역시 매우 높으며, F1 점수는 정밀도와 재현률의 균형을 나타내고 있습니다.  
ROC-AUC 점수가 0.9352로 중간 정도의 성능을 보이고 있으며, 이는 클래스 간 잘 구분된다는 것을 나타냅니다.  
  
텍스트, 영수증, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명  
- 분류 보고서 해석  
높은 정확도 (Accuracy: 0.99)와 매우 높은 재현률 (Recall)을 보여준다.  
클래스 1 (긍정적인 리뷰)의 경우에는 거의 모든 샘플을 정확하게 예측하고 있다. (재현률 1.00).  
클래스 0 (부정적인 리뷰)의 경우에는 재현률이 낮지만 (0.87), 정밀도는 매우 높다. (1.00).  
F1 점수는 모든 클래스에서 높은 값을 보여 모델이 각 클래스를 잘 예측하고 있음을 나타낸다.  
ROC-AUC 점수는 0.9352로, 클래스 간의 구분능력이 중간 정도이다.  
종합적으로, 이 모델은 주어진 데이터셋에서 높은 성능을 보이며 긍정과 부정 리뷰를 효과적으로 분류하는 데 적합하다고 평가할 수 있다.  
- 모델의 일반화 성능을 더 정확히 평가하기 위해 교차 검증을 사용하여 계산  
  
텍스트, 폰트, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명  
텍스트, 폰트, 스크린샷, 대수학이(가) 표시된 사진

자동 생성된 설명  
- 해석  
모든 평가 지표가 매우 높은 수치를 보이므로, 이 모델은 데이터셋에서 매우 우수한 성능을 보입니다.  
특히, 재현률과 ROC-AUC 점수가 1.0으로 완벽한 예측을 나타내므로, 긍정적인 리뷰와 부정적인 리뷰를 정확하게 구분하는 데 뛰어난 능력을 보여줍니다.

1. **시각화 및 해석:** 결과를 시각화하고 해석  
   텍스트, 스크린샷, 폰트이(가) 표시된 사진

   자동 생성된 설명

텍스트, 라인, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명  
ROC(Receiver Operating Characteristic) Curve는 모든 분류 임계값에서 분류 모델의 성능을 보여주는 그래프로 x축이 FPR(1-특이도), y축이 TPR(민감도)인 그래프이다.  
즉 민감도와 특이도의 관계를 표현한 그래프이다.   
ROC Curve는 AUC(Area Under Curve : 그래프 아래의 면적)를 이용해 모델의 성능을 평가한다.  
AUC가 클수록 정확히 분류함을 뜻한다. 즉, 위의 그래프에서 A = 0.94인 모델은 95%의 확률로 제대로 분류하고 있음을 뜻한다.  
#seaborn 사용

텍스트, 스크린샷, 폰트이(가) 표시된 사진

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

자동 생성된 설명  
위의 분류 모델의 성능 지표의 결과를 바탕으로 seaborn을 사용하여 혼동 행렬 그래프를 작성한 것이다.   
True Positive (TP): 실제 Positive인 데이터 중에서 모델이 Positive로 정확히 예측한 경우가 47개입니다.  
False Positive (FP): 실제 Negative인 데이터 중에서 모델이 잘못하여 Positive로 예측한 경우가 7개입니다.  
False Negative (FN): 실제 Positive인 데이터 중에서 모델이 잘못하여 Negative로 예측한 경우가 0개입니다.  
True Negative (TN): 실제 Negative인 데이터 중에서 모델이 Negative로 정확히 예측한 경우가 524개입니다.

\*\*False Negative (FN)\*\*이 0으로 나타나 있습니다. 이는 모델이 실제 Positive인 데이터를 하나도 놓치지 않고 모두 Positive로 정확히 예측했음을 의미합니다.  
\*\*False Positive (FP)\*\*가 상대적으로 작은 수인 7개로 나타나 있습니다. 이는 모델이 실제 Negative인 데이터를 Positive로 잘못 분류한 경우가 적다는 것을 나타냅니다.  
\*\*True Negative (TN)\*\*의 값이 상대적으로 큰 수인 524개로 나타나 있습니다. 이는 모델이 실제 Negative인 데이터를 대부분 정확하게 Negative로 예측했음을 의미합니다.