DAV 6150 모듈 10 할당

나이브 베이즈 텍스트 분류

***** 이 과제를 위해 3명 이하의 소규모 그룹으로 작업할 수 있습니다 ***.**

나이브 베이즈 분류기는 텍스트 데이터 분류(예: 감성 분석)에 있어 그 효율성이 널리 인정받고 있습니다. 지금까지 살펴본 바와 같이, 많은 조직에서 기존 고객과 잠재 고객의 의견을 파악하기 위해 감성 분석 알고리즘을 활용하고 있습니다. 예를 들어 아마존, 트립어드바이저, 부킹닷컴, 월마트, 옐프 등의 기업은 고객이 제공하는 온라인 제품/서비스 리뷰에 감성 분석 알고리즘을 적용하여 대중이 경쟁 제품과 서비스를 어떻게 인식하는지 더 잘 이해합니다.

모듈 10 과제의 과제는 영화 리뷰의 감정을 측정하기 위해 나이브 베이즈 감정 분류기를 구축하는 것입니다. 작업할 데이터 집합은 다음 사이트에서 가져옵니다: http://www.cs.cornell.edu/people/pabo/movie-review-data/. 특히 긍정적인 영화 리뷰 1000개와 부정적인 영화 리뷰 1000개로 구성된 극성 데이터 세트 v2.0으로 작업하게 됩니다. 각 영화 리뷰는 웹 사이트 게시물에서 캡처한 자유 형식 텍스트의 형태입니다. 이 과제를 완료하려면 상당한 양의 사전 처리 기술을 사용하여 분류 모델 내에서 사용할 수 있도록 리뷰의 내용을 준비해야 합니다(예: 구두점, 중단 단어 제거 등).

과제를 시작하려면 다음과 같이 하세요:

1) review_polarity.tar.gz 파일을 로컬 환경에 다운로드하고 압축을 풉니다. 압축 파일에는 부정적인 영화 리뷰 1000개가 들어 있는 **negative** 디렉터리와 긍정적인 영화 리뷰 1000개가 들어 있는 **pos** 디렉터리 두 개가 있습니다.

2) **부정** 및 **긍정** 디렉토리를 DAV 6150 Github 리포지토리에 로드합니다. 디렉터리 자체가 리뷰 분류를 위한 레이블 역할을 하므로 디렉터리의 콘텐츠를 분리하여 보관해야 합니다.

3) 그런 다음, 주피터 노트북을 사용해 새 Github 디렉터리에서 각 개별 영화 리뷰의 콘텐츠를 읽고 해당 콘텐츠를 2000개의 영화 리뷰에 포함된 모든 가능한 단어를 포괄하는 용어-문서 매트릭스 내에서 적절한 레이블(즉, POS/NEG 또는 적절한 프록시)이 붙은 항목으로 변환하는 알고리즘을 구축합니다. 용어 문서 매트릭스를 구성하는 동안 리뷰에서 구두점이나 중단 단어를 제거해야 합니다. 용어-문서 행렬의 구성과 적절한 레이블을 관리하는 방법은 데이터 과학/파이썬 실무자가 결정할 수 있습니다.

4) 다음으로, **양수에서** 가장 자주 발생하는 30개의 단어에 대한 빈도 분포를 플롯합니다.
   리뷰. 줄거리에서 어떤 인사이트를 얻을 수 있나요?

5) **부정적인** 리뷰에서 가장 자주 등장하는 30개의 단어에 대한 빈도 분포를 플롯합니다. 이 플롯에서 어떤 인사이트를 얻을 수 있나요?

6) 이제 2000개의 개별 영화 리뷰 각각에 대한 용어-문서 행렬 항목을 성공적으로 구성하고 적절하게 레이블을 지정했으므로, 용어-문서 행렬에 포함된 벡터의 75%를 무작위로 샘플링하여 모델 훈련 데이터 하위 집합에 포함시키고 나머지 25%의 벡터는 모델 테스트 데이터 하위 집합에 남겨둡니다. 데이터를 분할하는 방법은 데이터 과학/파이썬 실무자가 결정할 수 있습니다.

7) 데이터를 훈련 하위 집합과 테스트 하위 집합으로 나눈 후, 훈련 하위 집합을 사용하여 Python에 대한 지식을 바탕으로 나이브 베이즈 텍스트 분류기를 구성하고 훈련합니다. 나이브 베이즈 분류기를 구현하는 방법은 데이터 과학/파이썬 실무자가 결정할 수 있습니다. 분류기의 응답 변수는 관련 영화 리뷰가 긍정적인지 부정적인지를 나타내는 각 용어-문서 행렬 항목에 추가한 이진 플래그여야 한다는 점을 기억하세요.

8) 새로 학습된 모델을 생성한 모델 테스트 데이터 하위 집합에 적용하고 결과에서 도출한 분류 성능 지표(예: 정확도, 감도, 정밀도 등)에 대해 논의합니다.

9) 나이브 베이즈 분류기가 결정한 가장 유익한 특징 30개를 찾아서 표시하고 토론합니다. 이러한 특징을 분석하여 어떤 결론을 도출할 수 있나요?

10) 적절한 단계를 거쳐 나이브 베이즈 분류기에서 사용할 수 있는 형식으로 변환한 후 이전에 보이지 않던 약간 부정적인 리뷰에 분류기를 적용합니다(분류기에서 사용하기 위해 이전에 보이지 않던 리뷰를 어떻게 변환할지는 데이터 과학/파이썬 실무자가 결정할 수 있습니다):

"이 영화에서 마음에 들지 않는 부분이 몇 가지 있었어요. 제가 가장 강하게 기억하는 것은 다음과 같습니다. 기발하게 가짜로 보이는 북극곰 분장을 한 남자(뉴욕의 헤라클레스보다 더 웃긴), 믿을 수 없을 정도로 웃는 엑스트라, 마약 중독자 출신 화성인, 아주 천천히 조심스럽게 대사를 낭송하는 아역 배우, 산타가 '납치되었다'는 신문 헤드라인, 거대한 로봇이죠. 이 영화에서 가장 매력적이지 않은 연기는 마더 클로스와 엘프들이 '화성인'의 무기에 의해 '얼어붙는' 장면일 것입니다. 그들은 과장된 공포감을 드러내는 것 같았어요. 아마도 1960년대에는 이런 연기 스타일이 선호되었던 걸까요?"

11) 분류자가 이 리뷰에 어떤 분류를 지정했나요? 예상했던 것과 일치하나요? 분류가 정확한가요? 그렇지 않다면 분류기가 예상대로 작동하지 않는 이유는 무엇인가요?

이 과제의 결과물은 Jupyter 노트북입니다. 이 노트북에는 적절한 형식의 마크다운 셀에 포함된 Python 코드 셀과 설명 내러티브의 조합이 포함되어야 합니다. 노트북에는 최소한 다음 섹션이 포함되어야 합니다(각 섹션의 관련 Python 코드 포함):

1) **소개(5점)**: 문제 요약 + 문제 해결을 위해 취할 단계를 설명합니다.

2) **데이터 준비(25점)**: 제공된 데이터를 학기-문서 행렬 내에서 적절하게 레이블이 지정된 개수 벡터로 로드하고 변환하기 위해 수행한 단계를 설명하고 보여주세요. 이 섹션에는 데이터 준비에 사용된 Python 코드가 포함되어야 합니다.

3) **빈도 분포도(10점)**: 단어 수 빈도 분포도를 설명하고 제시합니다. 이 섹션에는 플롯을 만드는 데 사용

된 Python 코드가 포함되어야 합니다.

4) **나이브 베이즈 모델 훈련(25점)**: 카운트 벡터를 훈련 및 테스트 하위 집합으로 분리한 과정을 포함하여 나이브 베이즈 분류기 모델링 작업을 설명하고 제시하세요. 영화 리뷰 데이터에 사용할 특정 유형의 나이브 베이즈 분류기를 어떻게 결정했는지 설명하세요. 이 섹션에는 훈련 + 테스트 하위 집합을 생성하고 분류기를 훈련하는 데 사용된 Python 코드가 포함되어야 합니다.

5) **모델 테스트(25점)**: 테스트 하위 집합에 모델을 적용하고 그 결과를 토론합니다. 모델이 예상대로 잘 작동했나요? 그렇지 않다면 그 이유는 무엇일까요?

관찰되었습니다. 모델에서 도출한 분류 성능 메트릭의 맥락에서 토론의 틀을 잡아야 합니다. 나이브 베이즈 분류기가 식별한 가장 유익한 30가지 특징을 식별하여 얻을 수 있는 인사이트에 대해 토론하세요. 이전에 본 적이 없는 영화 리뷰(위에 제공된)에 모델을 적용합니다. 모델이 해당 리뷰에서 얼마나 잘 수행했나요?

6) **결론(10점)**

주피터 노트북의 결과물은 출판물 수준의 전문 문서와 유사해야 하며, 명확하게 표시된 그래픽, 고품질 서식, 명확하게 정의된 섹션 및 하위 섹션 헤더, 맞춤법 및 문법 오류가 없어야 합니다. 또한 Pythion 코드에는 간결한 설명 주석이 포함되어야 합니다.

제공된 M10 과제 캔버스 제출 포털에서 Jupyter 노트북을 업로드합니다. 노트북을 저장할 때는 **이름_성_M10_assn**"(예: J_Smith_M10_assn)과 같은 명명법을 사용하여 저장해야 합니다. ***소규모 그룹은 Jupyter 노트북을 시작할 때 모든 그룹 구성원의 신원을 확인하고 각 팀원은 Canvas 내에서 팀 작업의 사본을 제출해야 합니다.***