Assignment #6

Due date: 5월 29일 (금)

Sentiment Analysis on Movie Review Dataset

"**imdb.csv**" 파일은 영화 리뷰 사이트인 IMDB로부터 추출한 10,000개의 영화 리뷰 정보를 포함하며 다음 2개의 column으로 구성되어 있다.

- **review**: 리뷰 텍스트
- **sentiment**: "negative" or "positive"

데이터셋 중에서 5,028개는 평점 7점 이상을 기록한 **positive** 리뷰이며, 나머지 4,972개는 평점 4점 이하를 기록한 **negative** 리뷰이다. 평점 5~6점을 기록한 리뷰는 데이터셋에 포함되지 않았다. 그리고 영화 하나당 최대 30개의 리뷰를 포함하고 있다. 본 과제에서는 영화 리뷰 텍스트로부터 리뷰의 positive/negative 여부를 판별하기 위한 모델을 만들어본다.

1. Positive 리뷰 텍스트와 negative 리뷰 텍스트의 word cloud를 각각 그려보자. 둘을 비교했을 때 어떠한 차이점을 발견할 수 있는가?

2. 전체 10,000개의 리뷰 텍스트를 대상으로 corpus를 생성하고 bag-of-words 기법을 적용하기 위해 적절한 preprocessing을 수행해보자.

    A. 강의노트에서 다룬 모든 preprocessing 단계를 순서대로 수행한다.

    B. 모든 preprocessing 단계가 끝난 후, "movi"와 "film"을 모든 텍스트에서 삭제하자. 이들은 대부분의 텍스트에서 등장하기 때문에 feature로써 의미가 약하다고 할 수 있다. "movie"는 stemming에 의해 "movi"로 변환되었음에 유의하자.

    C. 원 텍스트와 preprocessing 후의 텍스트 사이에 어떤 변화가 있는지 리뷰 텍스트의 예를 들어 비교해보자.

3. Document-Term Matrix (DTM) 와 TF-IDF matrix를 생성하자. 그리고 출현 빈도 수가 낮은 단어(term)를 DTM과 TD-IDF matrix에서 제외한다. 이때 출현 빈도 수가 낮은 단어를 삭제한 기준은 무엇이며, 단어의 수가 얼마나 줄어들었는가?

4. 첫 5,000개의 데이터를 training set으로, 나머지 5,000개의 데이터를 test set으로 설정한 후, training set을 사용하여 리뷰텍스트의 positive/negative 여부를 판별하기 위한 predictive model을 만들어보자.

    A. Test set에 대한 classification accuracy를 높이기 위한 모델을 자유롭게 만들어본다. 이때 지금까지 학습한 모델을 최소 2개 이상 만들어보고, 분석 과정과 결과를 report하자. 사용하는 모델, 모델에 포함되는 파라미터에 대한 튜닝, 모델에 포함되는 feature의 수, DTM/TF-IDF 사용 여부 등이 classification accuracy에 영향을 미칠 수 있다. [**주의**: 모델을 수립할 때에는 test set을 사용할 수 없다.]

    B. 최종적으로 선택한 모델은 무엇이며 이 모델의 training set accuracy와 test set accuracy는 얼마인가?

[채점 기준]

1~3.	텍스트 프로세싱	2점
4A.	모델 수립 및 분석	4점
4B.	Test set accuracy	3점