[싹캠퍼스 최종 프로젝트]

호텔 예약 사이트 리뷰 기반 추천 시스템 개발 프로젝트

- **프로젝트 기간 : 2022.01.17. ~ 2022.02.10.**
- **프로젝트 팀 구성 : 자연어초밥 (우문주(팀장), 원천표)**

Ⅰ. 프로젝트 주제 선정 배경과 목표

1. 프로젝트 주제 선정 배경

리뷰를 참고하여 실제로 호텔을 선택하는 사용자 비율이 84% 이상이라는 논문 결과가 있을 정도로 상당히 많은 소비자들이 직접 호텔을 이용한 사람들이 남긴 리뷰를 기반으로 묵을 호텔을 선택하는 것으로 보인다. 그러나 호텔 예약 사이트에서 특정 지역의 호텔을 검색하면, 사이트의 자체적인 알고리즘에 따른 순서대로 호텔이 추천되는 것을 볼 수 있는데, 이것이 어떤 기준으로 나열되는 것인지는 알려진 바가 없어 신뢰하기 어렵다고 판단된다. 그리고 사용자는 리뷰를 작성할 때, 평점과 글을 동시에 작성하게 되는데, 리뷰는 텍스트 자체로 긍정과 부정이 판단될 수 있으나, 숫자로 된 평점은 주관적으로 매겨지므로 리뷰 텍스트에 비해 신뢰도가 떨어질 수밖에 없다. 그리고 통계적으로 부정적이고 자세한 리뷰보다는 긍정적이고 짧은 리뷰가 상대적으로 많이 등록되어 있기 때문에, 호텔 선택에 있어서 실패하지 않으려면 부정적인 리뷰들을 잘 파악하여 선택하는 것이 바람직하다고 판단했다.

따라서 우리는 리뷰 텍스트를 키워드 기반으로 직접 감성분석하여 이를 통해 새롭게 키워드 라벨링을 한 후, 부정적인 키워드에 중점을 두어, 실패하지 않을 확률이 높은 안전한 호텔을 추천을 하는 알고리즘을 만들기로 했다.

2. 프로젝트 목표

이번 프로젝트에서는 '야놀자'와 '여기어때', 두 호텔 예약 사이트 내 리뷰 데이터를 크롤링해서 DB에 저장하고, '야놀자' 데이터를 전처리하여 평점 기반 감성분석 모델과 키워드 기반 감성분석 모델을 만들어서 '여기어때' 데이터를 가지고 예측하여 정확도를 확인 및 비교한다. 평점 기반 모델은 리뷰 작성자가 직접 작성한 별점을 기반으로 레이블링하고, 키워드 기반 모델은 평점 기반 모델과 달리 긍정/부정 키워드 기반으로 키워드 평점을 매겨 레이블링을 하는 것을 큰 차이점으로 하며, 이 결과 리뷰 텍스트의 감성을 더 잘 반영하는 키워드 감성분석 모델이 더 정확도가 높을 것으로 예상한다.

Ⅱ. 데이터 수집과 저장(Data Scraping and Storing)

모델링에 사용할 '야놀자' 사이트와 예측에 사용할 '여기어때' 사이트는 json, html을 url 파싱하여 호텔 이름, 호텔 등급, 호텔 키, 호텔 사이트 키, 리뷰 텍스트, 리뷰 키, 호텔 시설 상세사항(와이파이 가능여부, 피트니스, 사우나, 어메니티, 레스토랑 여부)을 가능하면 1, 불가하면 0으로 인코딩하여 수집하였다. 이렇게 수집한 데이터는 MARIA DB와 AWS를 이용해 저장하고 util.py의 코드를 활용해서 DB에서 불러와서 모델링에 사용했다.

Ⅲ. 리뷰 텍스트 전처리(Review Text Preprocessing)

1. 1차 전처리 작업

1) 한글/공백 제외 구두점/이모티콘/영어 제거 (only_ks())
2) null값 제거 (remove_null())
3) 자모음 중복 제거 (no_repeat())
4) 공백 있는 행 제거 (remove_whitespace())
5) 중복행 제거 (remove_dupli())
6) 맞춤법 검사 (spell_check())

2. 2차 전처리 작업

1) 훈련 데이터와 테스트 데이터로 나누기

훈련 데이터는 전체 데이터의 2/3, 그리고 나머지 1/3은 테스트 데이터로 분리한다.

2) 레이블링 하기

① 평점 기반 모델

리뷰평점 1~4점까지는 부정(0)으로, 5점을 긍정(1)로 라벨링한다.

② 키워드 기반 모델

부정리뷰가 긍정리뷰에 비해 현저히 적은 관계로, 부정 키워드에 가중치(w)를 매긴다. 그리고 키워드 평점을 내기 위한 n을 계산하는 식은 아래와 같다.

{(한 리뷰의 부정 키워드 개수 / 한 리뷰의 전체 키워드 개수) * w} * 100

n이 50점 이상이면 키워드평점 1점, 40점 이상이면 키워드 평점 2점, 30점 이상이면 키워드 평점 3점으로 여기까지가 부정(0)으로 라벨링 되고, n이 20~30점 사이면 키워드 평점 4점, 그 아래는 키워드 평점 5점을 주어 긍정(1)으로 라벨링한다.

3) 토큰화와 불용어 제거

형태소 분석기 Mecab을 사용해서 토큰화 작업을 수행하고, 조사나 의미없는 단어를 불용어 리스트로 만들어 제거한다. Counter()를 사용하여 긍정리뷰와 부정리뷰의 빈도수 높은 토큰들을 확인하는 과정도 여기서 진행된다.

4) 정수 인코딩

훈련 데이터에 대한 단어 집합을 만들고, 고유한 정수를 부여하여 인코딩한다. 등장횟수가 매우 적은 단어들은 배제하고 최종 단어 집합을 만들고, 더 큰 숫자가 부여된 숫자들은 OOV 변환한다.

5) 패딩

가장 길이가 긴 리뷰와 전체 데이터의 길이 분포를 파악하고 서로 다른 길이의 샘플들의 길이를 동일하게 맞춰준다.

Ⅳ. 모델링 하기(Modeling)

모델은 다 대 일 구조의 LSTM을 선택했고, 임베딩 벡터 차원은 100, 은닉 상태의 크기는 128로 설정했다. 그리고 이는 긍정인지 부정인지 둘 중의 하나를 선택하는 이진 분류 문제 수행 모델이므로, 출력층에서 로지스틱 회귀를 사용, 활성화 함수로는 시그모이드 함수를 사용했다. 그리고 손실함수로는 크로스엔트로피 함수를 사용한다. 배치 크기는 64로 진행하고 15에포크 정도만 수행하는 것으로 설정했다.

또한 얼리 스타핑(EarlyStopping) 기능을 추가해서 validation loss가 4회 증가하면 과적합 징후로 판단하여 에포크를 조기종료하도록 했고, 모델 체크포인트(ModelCheckpoint) 기능을 사용하여 validation accuracy가 이전보다 나아질 경우에만 모델을 저장하기로 설정한다. 검증 데이터는 훈련데이터의 20% 정도만 분리해서 사용하도록 설정하여 한다.

Ⅴ. 모델로 리뷰 예측하기

1) 평점 기반 모델로 리뷰 예측하기

전체 리뷰 개수가 111,172개인데, 이 중 긍정으로 판단한 리뷰 개수는 77,269개, 부정으로 판단한 리뷰 개수는 33,903개로, 원래 라벨링과 일치하는 샘플은 86,750개, 불일치하는 샘플은 24,422개이며, 따라서 정확도는 78.03%를 기록하였다.

2) 키워드 기반 모델로 리뷰 예측하기

시간 관계상 전체 리뷰 중 50,000개 샘플을 뽑아서 예측해 본 결과, 키워드기반 라벨링과 일치하는 샘플은 49,034개, 불일치하는 샘플은 966개이며, 따라서 정확도는 98.07%를 기록하였다.

Ⅵ. 추천 시스템 만들기

더 정확도가 높은 키워드 기반 모델로 각 리뷰마다 긍정과 부정 결과를 표시해서 통계를 낸 후, 한 호텔의 전체 리뷰를 분모로, 부정 리뷰를 분자로 놓고 100을 곱한 네거티브 점수가 가장 낮은 순으로 정렬한다. 그리고 사용자가 지역 검색 시, 네거티브 점수가 최하위 순으로 5개를 추천하는 시스템을 구성한다.

● 역할분담

우문주(팀장)	원천표
- 프로젝트 전체적인 큰 계획 - 크롤링 ('여기어때' 사이트) - 발표 PPT, 프로젝트 보고서 작성 - 모델 구조 만들기	- 구체적인 코딩 구현 - 크롤링 ('야놀자' 사이트) - 코드 리뷰 및 수정 - 코드와 datasheet 정리

-