H040-4-4139-01 텍스트 및 오피니언 마이닝

Team Project 최종 보고서

오늘의 집 리뷰 토픽 모델링 분석

TM-WDM (김기수, 김재윤, 김태영)

**1. 서론**

**1.1 주제 선정 배경**

최근 상품 설명 요약문, SNS 글, 사용자 리뷰 등과 같이 단문 형태의 텍스트 데이터의 양이 증가하고 있다. 이러한 단문 데이터는 사용자의 의견이 함축되어 있기 때문에 중요성이 날로 높아지고 있다. 그 중 리뷰 데이터는 쉽고 빠르게 수집할 수 있기 때문에, 이와 관련된 주제를 설정하여 프로젝트를 진행하고자 하였다. 11주차에 걸쳐 배운 내용들을 충분히 활용하고, 이를 통해 유의미한 인사이트를 도출하는 것을 큰 목표로 설정하였다.

**1.2 구체적인 주제 선정**

학기 시작부터 팀원 미팅을 통해 추천 시스템으로도 발전시켜볼 수 있는 참신한 주제를 고민하였다. 사용자의 리뷰 데이터는 다양한 분야에서 수집할 수 있는데, 그 중 대학생의 많은 공감을 불러올 수 있는 자취생을 타겟으로 한 주제를 고민하였다. 코로나 종식이 다가옴에 따라, 대면 수업 증가로 자취생의 비율이 늘어날 것이라고 예상할 수 있으며, 그에 따라 ‘오늘의 집’ 리뷰에서 찾을 수 있는 단점을 보완하고 장점을 극대화할 수 있는 형태의 프로젝트를 진행하기로 하였다. 자취 관련 기구를 파는 사람의 입장은 판매자이다. 판매자의 입장에서 어떤 부분을 보완해서 제작자에게 알려줄 지, 반응이 좋은 부분은 계속 이어나갈 수 있도록 도와주는 것을 목표로 하였다.

**2. 내용**

**2.1. 데이터 분석 프로세스**

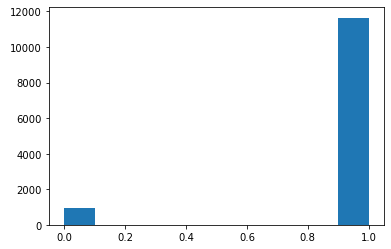
해당 프로젝트의 분석 프로세스를 요약하면 아래와 같다.

먼저 ‘오늘의 집’에서 자취와 관련된 제품의 댓글 및 평점 데이터를 크롤링한 후, 노이즈 캔슬링, 토크나이징, 텍스트 벡터화 등의 전처리 과정을 수행하였다. 이후 리뷰 데이터로 감성 분석을 진행하여 댓글의 평점을 Rescoring 해주고, 토픽 모델링을 진행하여, 각 제품별 토픽을 정리했다. 정리된 토픽을 바탕으로 인사이트를 도출하며 프로젝트를 정리하였다.

**2.2. 데이터 수집 및 전처리**

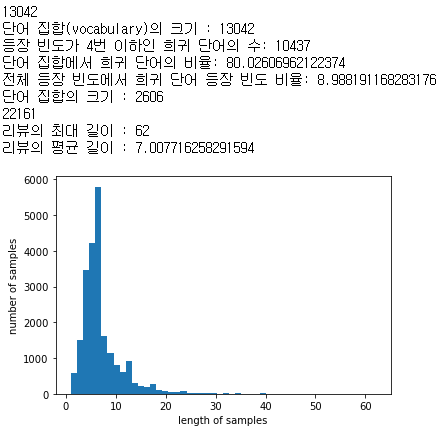
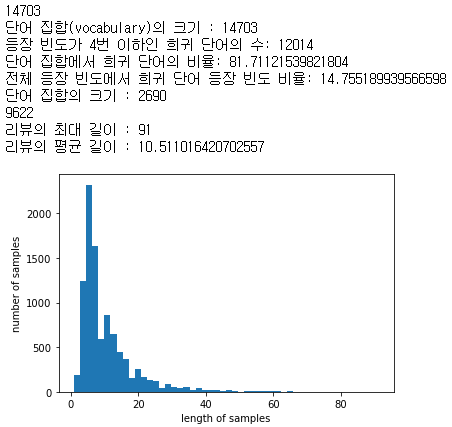
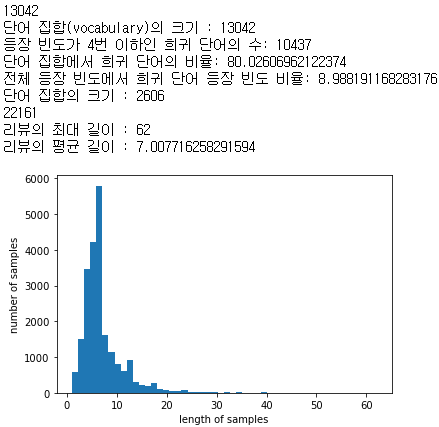
데이터는 ‘오늘의 집’ 사이트에서 ‘자취’ 라는 키워드를 검색한 후 리뷰 많은 순으로 정렬시켜 각각 다른 카테고리를 갖는 5개 정도 제품의 평점 및 리뷰를 데이터 프레임 형태로 수집하였다. 이 때 5개의 제품에는 세부적인 선택사항이 존재했기 때문에 20개 정도의 각각 다른 제품에 대한 데이터 수집이 이루어졌다고 볼 수 있다.

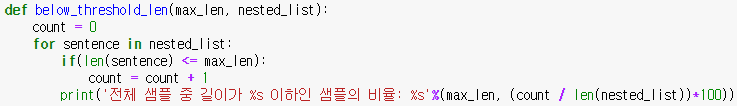
반복되는 작업(리뷰 가져오기, 리뷰 내 페이지 넘기기 등)을 자동화하기 위해 우리는 “Selenium”을 사용했고, 텍스트 형태의 데이터에서 html 태그를 이용하여 원하는 객체 추출을 위해 “BeautilfulSoup”을 사용했다.



리뷰 데이터를 수집하는 과정 중 대부분의 리뷰가 위와 같이 4점 이상의 평점을 갖는 것으로 보여졌다. 이는 추후 감성 분석을 진행하기 위해 미리 확인하는 것이 필요하다고 판단했으며, 4점 이상의 평점을 갖는 리뷰 데이터와 그 미만의 평점을 갖는 리뷰들의 불균형을 위와 같이 확인할 수 있었다. 이를 충분히 반영하여 평점을 재설정하는 과정이 필요하다고 판단할 수 있다.

수집된 리뷰 데이터의 결측 및 중복 데이터를 확인 후 제거하였고, 분석에 필요한 한글 외에 특수 문자 등을 삭제해주었다. 이후 py-hanspell을 활용하여 맞춤법을 수정하였다. 다음으로는 konlpy의 형태소 분석기인 Okt를 사용하기로 하였으며, morphs보다 주요 품사들만 태깅 하여 추출할 수 있는 pos가 더 좋은 토픽 모델링으로 이어질 수 있을 것이라 판단하여 Okt.pos를 통해 유의미한 품사의 단어들만 추출하였다.



이후 텍스트를 벡터화를 진행하기 전, 위와 같이 희귀 단어 빈도를 확인하고 벡터화 할 크기를 재조정 함으로써 리뷰의 평균 길이에 변화를 주었다. 최종적으로 pad\_sequences를 통해 벡터화를 최종 진행하며 전처리 과정을 마무리하였다.

**2.3. 감성분석**

모든 리뷰에 대해서 객관성을 부여하는 과정이 필요했기 때문에 모든 리뷰에 대한 감성분석을 진행했다. 감성 분석을 위한 학습 데이터셋은 AI hub, DACON 등 다양하게 찾아보고 고민한 뒤 ‘Naver Movie Sentiment Corpus’로 최종 선택하였다. 해당 데이터셋은 20만개의 리뷰 데이터(Train-15만 개, Test-5만 개)를 포함하고 있기에 학습 데이터셋으로 충분하다고 판단하였다. 총 3개의 모델을 사용하여 학습을 진행시켰으며, 결과는 아래와 같았다.

|  |  |
| --- | --- |
| Model | Accuracy |
| Logistic Regression (TF-IDF) | 약 82% |
| Logistic Regression (Count Vectorizer) | 약 81% |
| **LSTM(Tokenizing)** | **약 85%** |

Accuracy가 85%로 가장 높게 나타난 LSTM을 수집한 데이터셋에 최종 활용하기로 결정하였다. 학습 시간이 오래 걸리더라도 감성 분석 시 정확도를 높이는 것이 중요했기 때문에 hidden units 파라미터에 대해서 128로 설정해주었다. 추가적으로 embedding\_dim은 512로 설정하였다.

수집한 리뷰 데이터에 대해서 감성 분석을 진행하였고, 새로운 변수 ‘sentiment\_score’를 만들어 결과에 따른 감정 점수를 따로 부여하였다. 이렇게 생성된 ‘sentiment\_score’만으로 토픽 모델링을 진행하는 것이 좋을지, 혹은 기존 점수에도 가중을 두고 분석하는 것이 좋을지에 대해서 다양하게 고려하였다. 우선 sentiment\_scorer과 기존의 평점 차이가 무척 크게 나타나는 것을 확인할 수 있는데, 사용자의 평점 의견이 더 중요하다는 판단 하에 3:1의 가중치를 주어 Rescoring을 진행하였다.

**2.4. 토픽 모델링**

Rescoring 된 평점 별 어떤 토픽을 주제로 하고 있는지 확인하기 총 4개의 구간으로 아래와 같이 분할 하였다.

|  |  |
| --- | --- |
| 기준 | 보정된 점수에 따른 평균값 |
| 2점 이하 (very bad) | 1.506 |
| 2점 초과 3점 이하 (bad) | 2.678 |
| 3점 초과 4점 이하 (normal) | 3.640 |
| 4점 초과 (good) | 4.487 |

토픽 모델링은 gensim 라이브러리를 통해 LDA 토픽 모델링을 진행하였고, pyLDAvis를 통해 시각화 하였다. LDA 모델을 생성할 때 파라미터(num\_topics(토픽 수), passes(반복 훈련 횟수))를 조금씩 변경 시키면서 결과 차이를 확인했고, 몇 번의 과정을 통해 가장 적합한 파라미터 설정에 따른 결과를 얻어낼 수 있었다. 그렇게 하여 토픽은 총 2개로 나타내기로 하였다. 카테고리 별 최종 토픽 결과는 아래와 같이 나타났다.

1. 선반 / 수납장 / 거울

|  |  |
| --- | --- |
| 기준 | Topic Modeling |
| verybad | 교환, 배송, 환불, 하자, 반품, 센터, 문의, 검수, 상태, 주문, 불량 |
| bad | 조립, 디자인, 수납, 별로, 좋아요, 배송, 나사, 구멍, 불량, 귀찮아서, 하자 |
| normal | 좋아요, 조립, 깔끔, 책상, 디자인, 튼튼, 최고, 배송, 서랍, 빠르고, 여자 |
| good | 튼튼, 쉽다, 디자인, 구매, 빠르다, 이쁘다, 화이트, 처음, 완성, 여자 |

1. 가구

|  |  |
| --- | --- |
| 기준 | Topic Modeling |
| verybad | 배송, 조립, 교환, 제품, 디자인, 택배, 별로, 귀찮음, 문의 |
| bad | 생각, 가격, 조립, 좋아요, 만족스럽다, 괜찮다, 디자인, 환불 |
| normal | 배송, 조립, 디자인, 빠르다, 예쁘다, 만족스럽다, 좋다, 튼튼하다 |
| good | 튼튼, 쉽다, 디자인, 구매, 빠르다, 이쁘다, 화이트, 처음, 완성, 여자 |

1. 청소용품(밀대, 빨래 바구니)

|  |  |
| --- | --- |
| 기준 | Topic Modeling |
| verybad | 별로, 배송, 반품, 불량, 조립, 교환, 제품, 밀대, 않고 |
| bad | 좋아요, 배송, 청소, 구매, 교환, 가볍고, 닦이고, 편하다 |
| normal | 걸레, 배송, 가격, 조립, 편하다, 닦인다, 자주, 가볍다 |
| good | 청소, 걸레질, 편하다, 사세요, 가볍다, 배송, 구매, 물걸레, 분사 |

교수님께 최종 결과 발표 피드백을 받은 후 카테고리마다의 다른 토픽의 결과를 내는 것이 중요하다고 판단됐다. 이에 따라, 위에서 진행한 카테고리와는 차이가 자취 용품을 선정하여 추가적으로 데이터를 수집하였다. 토픽 모델링 결과는 아래와 같았다.

1. 욕실용품(욕실화)

|  |  |
| --- | --- |
| 기준 | Topic Modeling |
| verybad | 미끄럽다, 하자, 넘어질, 다이소, 가격, 배송, 불량, 물이, 바닥, 최악 |
| bad | 냄새, 구멍, 상태, 부분, 화가, 구멍, 걸렸어요, 욕실 |
| normal | 좋아요, 배송, 깔끔하고, 가성, 가볍고, 만족합니다 |
| good | 만족합니다, 이뻐요, 가볍고, 디자인, 구매, 쿠션, 욕실, 좋아요 |

다른 제품들과는 다른 느낌의 Topic이 결과로 나타난 것을 확인할 수 있었다. 하지만, 여전히 배송과 관련된 불만 리뷰가 존재함을 알 수 있었다. 다른 카테고리와 비교하여 보았을 때 기준에 따른 Topic 결과가 크게 다르게 나타나지 않는다. 오늘의 집의 사용자들은 공통적으로 택배, 배송 측면에서 불만을 가지고 있음으로 판단된다. 카테고리 별 차이를 보일 수 있는 색상, 디자인과 관련된 특정 단어들에 대해서만 토픽 모델링을 진행할 수 있는 방법을 생각해보아야 할 것이다.

**3. 기대 효과 및 결론**

**3.1 프로젝트 결과 및 기대 효과**

아래의 표는 각 카테고리들에 대해서 공통적으로 나타난 결과에 대한 분석 내용이다.

|  |  |
| --- | --- |
| 기준 | 분석 결과 |
| verybad | 제품에 대한 불만으로 인한 반품&교환과 관련한 언급이 주를 이룬다. |
| bad | very bad의 내용을 포함하는 동시에 제품에 대한 만족도 존재한다. |
| normal | 긍정적인 평가가 주를 이루었다. 내용은 편안함, 청결 등이 언급됐다. |
| good | 부정적인 평가는 보이지 않는다. 제품 자체에 대한 만족이 대부분이다. |

각 카테고리에서 공통적으로 나타난 토픽들의 빈도가 무척 높았다. 제품에 대한 불만보다는 배송 과정에서 오는 불만이 더 크다는 사실을 알 수 있었다. 제품 자체에 대한 불만 또한 배송 과정에서 발생하는 불량 등에 대한 내용이 빈번했다.

이에 따라, 포장과 관련한 부분, 제품을 던지는 비도덕적 행위와 같이 배송 과정에 있어서 충분한 교육이 필요할 것이라고 판단된다.

텍스트 마이닝에 초점을 맞추어 프로젝트를 진행하였지만, 프로젝트 주제 자체는 추천 시스템으로의 연결 가능성이 충분하다. 협업 필터링과 콘텐츠 기반 필터링을 활용하여 추천 시스템으로의 구현이 가능할 것이다. 가구별 클러스터를 생성한다면 신규 고객에게 원하는 카테고리를 입력 받음으로써 추천 시스템에서 대두되는 cold start 문제를 어느정도 해결할 수 있을 것으로 판단된다.

**3.2 한계점 및 개선할 점**

1) 리뷰 평점은 4점 이상의 값을 가지는 경우가 대부분이었고, 리뷰의 내용과는 상반된 점수를 갖는 경우가 존재했기 때문에, 이를 충분히 고려한 Rescoring 과정이 필수적으로 필요했다. 기존 평점과 리뷰의 감성 점수를 통해 재 측정하는 과정은 정성적인 부분이었기 때문에, 이를 적절한 점수가 되도록 고민하는 것에는 어려움이 따랐다.

2) LSTM 모델의 정확도가 낮다고 할 수는 없지만, 리뷰에 대해서 적절한 감정 점수를 내지 못하는 경우도 존재하였다. 이 부분에 대한 추가적인 고민이 필요하다고 느꼈으며, 현재로서는 사용자 정의 사전을 주기적으로 추가함으로써 정확성을 높이는 방향을 고민하고 있다.

3) 카테고리 별 평점 분석을 목표하는 결과물로 두고 시작하였지만, 대부분의 카테고리에서 등장하는긍/부정 토픽 단어들은 유사하게 나타났다. 카테고리 각각의 개별적인 특징을 보여줄 수 있는 토픽 모델링을 진행하는 것이 필요하기 때문에, 이 부분을 고민해봐야 할 것이다.

4. 참고 문헌

[1] [Kisooofficial/TextMiningProject: This is the Text Mining Team Project which I had study in 2022-1 (github.com)](https://github.com/Kisooofficial/TextMiningProject)

[2] [한국어 불용어 리스트 100개 (tistory.com)](https://bab2min.tistory.com/544)

[3] [딥 러닝을 이용한 자연어 처리 입문 - WikiDocs](https://wikidocs.net/book/2155)