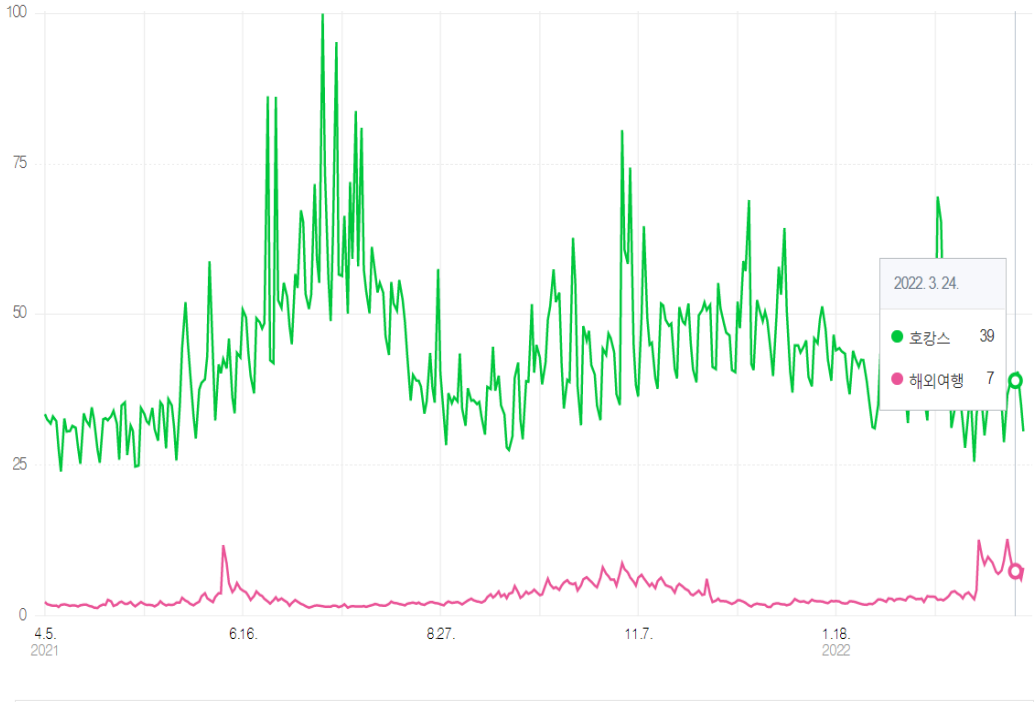
20195262 장예진

[주제]

호텔 리뷰 텍스트 분석

1. 연구배경



코로나 이후 재택근무의 시행, 사회적 거리두기 등으로 실내에 머무르는 시간이 증가함에 따라 서울 호텔, 호캉스 키워드는 예전에 비해 증가하였다. 이전에 인기있던 해외여행 키워드는 최근 호텔+바캉스의 줄임말인 호캉스라는 강자 키워드에 확연히 밀리고 있는 모습을 확인할 수 있다. (출처: 네이버 데이터랩 검색어트렌드) 완벽한 호캉스를 위한 호텔을 예약하기 위해 가장 중요하게 여기는 요소는 바로 “사용자들의 호텔 리뷰”이다. 그렇지만 리뷰의 대부분은 ‘굿’,’잘 쉬다 갑니다.’등 예약에 크게 도움이 되지 않는 리뷰 들이 많은 것을 확인할 수 있었다. 이는 고객의 불필요한 스크롤을 하게 만드는 요소이며, 이를 해결하기 위해 여러 호텔 예약사이트의 리뷰를 수집하여 그에 대한 중요도를 추출하는 텍스트 마이닝 방법을 제시한다.

2. 관련연구

(데이터 수집)

웹 크롤링

Python 기반의 web crawling 라이브러리인 selenium을 이용해 크롤링 대상인 페이지에 동적인 동작을 곁들여 크롤링을 진행한다. Chrome Webdriver 라는 가상의 브라우저 프로그램(일종의 웹 터스트 도구)과 연동하여 크롤링을 구현한다.

(데이터 전처리)

형태소 분석 및 추출

- 전처리 과정은 Python 기반의 한글 자연어 처리 패키지(Konlpy)를 활용하였으며, 영문(홈페이지 주소 등), 숫자, 구어체, 오타, 문장부호, 이모티콘 등 분석과정에는 제외해야 할 요소들을 정제하는 작업을 진행한다. (정규표현식을 통해서도 가능)
- 아울러, 형태소 분석기를 통해 데이터 중 주요 한글 품사를 바탕으로 '명사'만을 추출한다.

불용어 제거

- 형태소 분석을 거쳐 추출한 명사 중 형태소이지만 별다른 뜻이 없는 지시어('이것', '저것'등)를 stopword함수를 사용하여 처리한다.

-최소분석 단위 설정을 통해 내용 컬럼에서 15단어가 넘지 않는 내용은 분석에서 제외한다.

워드 클라우드 (WordCloud)

-형태소 분석 및 추출을 통해 각 단어의 빈도수를 파악하고 그에 따른 변수를 지정한다. 파악된 단어의 빈도수를 활용하여 높은 단어 일수록 크게 빈도수가 낮은 단어일수록 작게 표현하는 시각화 기법이다.

TF-IDF

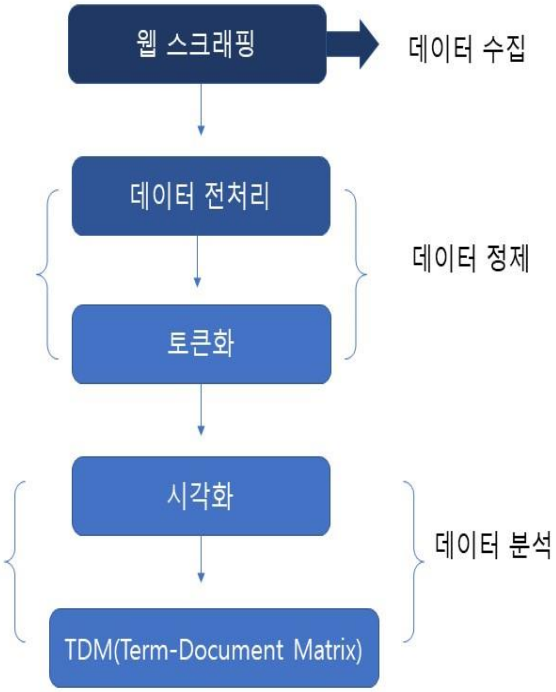
-일반적으로 단어 인덱싱을 하는 부분은 python 패키지에서 자동적으로 수행되기는 하지만 기본인 counter 기반부터 접근하여 빈도들을 dictionary에 먼저 반환해 주었다.

- Term frequency 띄어 쓰리고 구분되어 있는 단어들의 집합 documents를 입력하여 CounterVectorizer 를 사용하여 쉽게 document-term matrix를 구할 수 있다.
- tf-idf패키지인 TfidVectorizer패키지를 사용하여도 같은 결과 값을 도출한다. 이를 통해 단어의 중요도를 산출하였고 선택된 100개의 단어를 추출하여 단어의 중요도, 밀집도를 파악 가능하다.

텍스트 마이닝(Text mining)

- 이메일, 웹 페이지, 텍스트 문서 등의 비정형 형태의 텍스트 데이터에서 의미 있는 새로운 정보를 추출하는 것을 목적으로 하는 빅데이터 분석방법으로 문서에서 표현된 단어들을 바탕으로 내용을 분석하여 작성자의 의도를 파악할 수 있으며 문서 분류 및 군집화, 개념 및 요소 추출, 요소 관계 모델, 감정 분석, 문서 요약 등에 주로 활용되고 있다.

3. 제안방법론



텍스트 마이닝 기술을 이용해 호텔 리뷰를 분석하여 보다 효율적이고 효과적으로 호캉스의 질을 높일 수 있는 시장 세분화의 개념을 도입하였다. 즉, 제안한 방법론은 텍스트 마이닝 분야에서 시장 세분화의 개념에 부응하는 기술들이라 할 수 있는 범주화와 정보 추출 기법의 사용을 포함한다. 특히, 통계적으로 보다 견고한 분석 결과를 도출할 수 있도록 형태소 분석, 불용어 처리, 토큰화를 통한 품사 선정과 단어의 원형 추출을 포함하였다. 제안한 방법론의 타당성을 확인하기 위해 호텔의 고객리뷰가 많은 야놀자 웹사이트에 제시된 실제 온라인 고객리뷰들을 데이터 분석에 활용하였다.

4. 실험내용

먼저. 네이버 데이터랩 검색어 순위를 통해 호텔예약 플랫폼 1위인 야놀자의 서울 호텔 5성급 호텔 4곳을 임의로 선정하였다. 데이터 수집 즉, 웹 스크래핑(크롤링)과정에서는 야놀자 플랫폼의 호텔리뷰 페이지에 접속해 python library에 내장된 chrome driver, Selenium 모듈을 이용한 크롤링을 통해 1465개의 리뷰 데이터를 수집하였다. data_전처리전.csv를통해 전처리 전의 raw data set을 확인할 수 있다.

In [70]: df

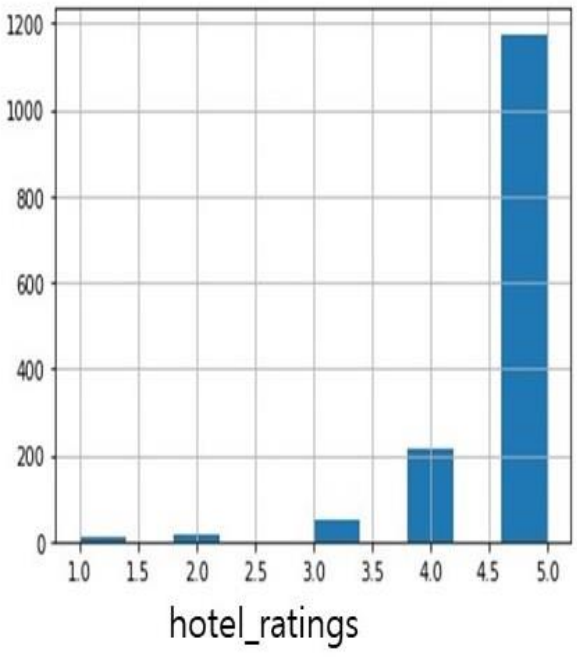
Out [70]:

	hotel	ratings	review	real_date	length
0	그랜드 인터컨티넨탈 서울 파르나스	5	12월에 리오픈해서 그런지 엄청 깨끗하고호텔리어 분들이 너무 친절 하셨어요.조식도 ...	2020. 12. 20	68
1	그랜드 인터컨티넨탈 서울 파르나스	5	침대도 좋고 시설 자체가 전부 청결하고 좋았습니다.어메니티도 마음에들어서 한세트 더...	2022. 04. 19	102
2	그랜드 인터컨티넨탈 서울 파르나스	5	일단 방 사이즈는 크지는 않았지만, 지내기엔 부족하지 않았습니다. 테헤란로뷰 였지만...	2022. 04. 18	226
3	그랜드 인터컨티넨탈 서울 파르나스	5	너무 깔끔하고 역시 좋았습니다	2022. 04. 18	16
4	그랜드 인터컨티넨탈 서울 파르나스	5	일단 코엑스에 위해서 너무 좋았고룸 컨디션또한최고였습니다	2022. 04. 18	31
...	...	...	...	...	...
1460	포시즌스 호텔 서울	5	좋았습니다	2020. 09. 08	5
1461	포시즌스 호텔 서울	5	겁나 깨끗함 친절함	2020. 09. 03	10
1462	포시즌스 호텔 서울	5	아주조아용	2020. 08. 18	5
1463	포시즌스 호텔 서울	5	아주조아용	2020. 08. 18	5
1464	포시즌스 호텔 서울	5	역시나 너무 좋았어요 호캉스는 역시 포시즌스가 최고 입니다 룸 컨디션이 너무 좋아서...	2020. 08. 17	310

1465 rows × 5 columns

각 호텔의 리뷰에서 개행문자(₩n)를 제거하고, 의미 없는 자음, 모음, 특수문자, 이모티콘을 제거하는 전 처리 작업을 진행하였다.

데이터 정제, 즉, 데이터 전처리 과정에서는 형태소 분석기인 Knolpy를 이용해 크롤링 과정에서 추출한 data를 활용하여 형태소 분석과 명사추출, 불용어 제거, 토큰화, 해당 데이터의 review 부분에서 명사만을 추출하여 각 명사의 빈도계수를 측정하였다. '호텔','저','앞'등 토큰화를 통해 특정 품사를 추출하고, 분석에 의미 없는 불용어를 제거하여 명사나 유의미한 형태의 용어들 만을 남겨 변수에 저장해 두었다.



>>>그림1)리뷰를 남기는 사람의 대부분이 평균적으로 4,5점의 별점을 남긴다. 별점으로는 크게 해당 호텔 예약을 좌우하지 못하는 것을 확인 가능

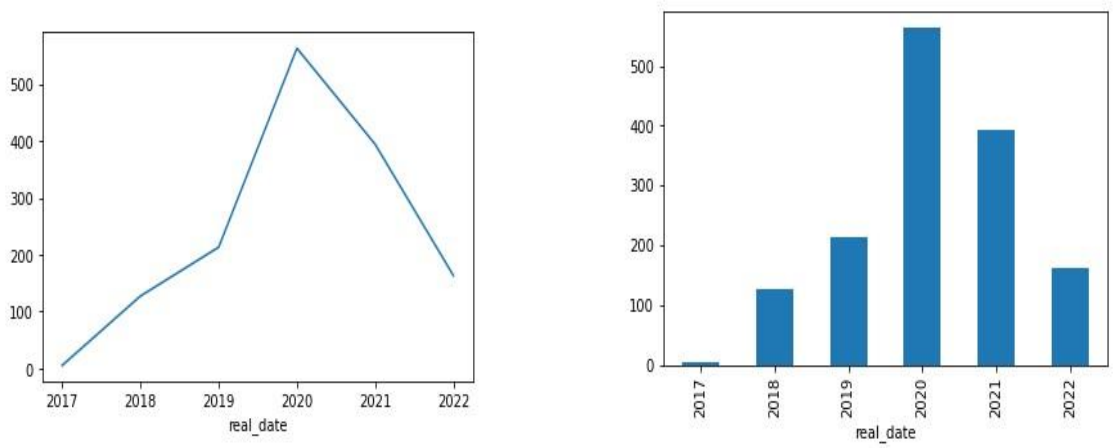


그림2,3) 해당 호텔의 리뷰를 남긴 날의date를 년, 월, 일로 표현한 것(str)을 datetime으로 변환하여plotting하였다. line, bar plot 두 가지를 비교해보았을 때 코로나가 발생하기 이전의 2017,2018,2019년에 비해 감염률이 치솟았던 2020,2021년에 호텔 리뷰의 개수가 월등히 많은 것을 확인할 수 있다.

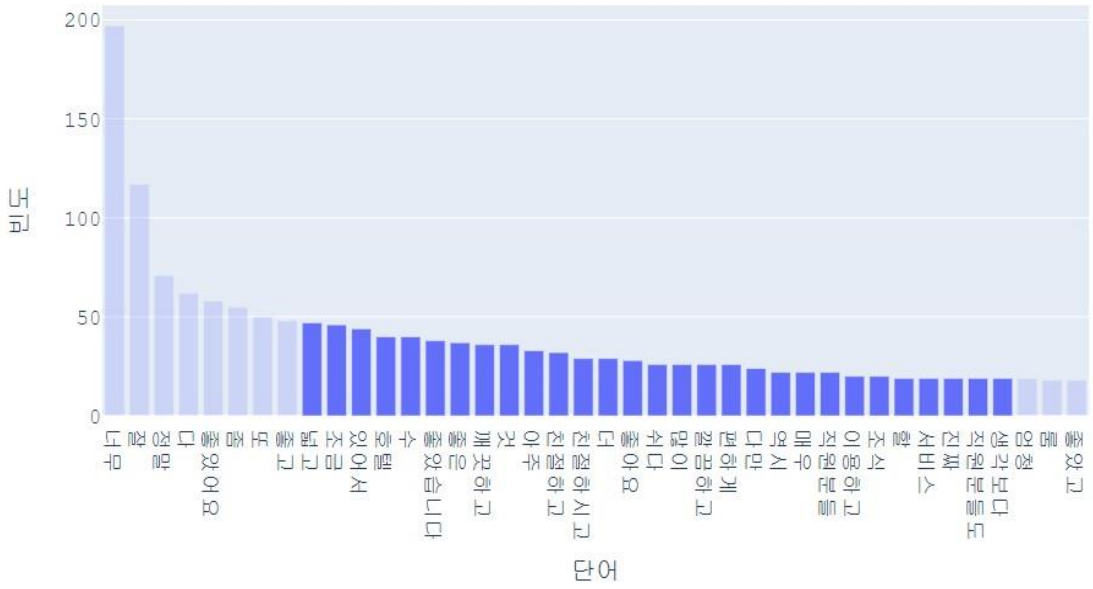


그림4) 단어 빈도수를 파악하여 그것을 시각화 한 그래프이다. 단어 빈도수가 높을수록 높은 중요도를 나타낸다고 가정했을 때, (너무, 잘, 정말, 다, 좋았다. 등의 상투적인 조사를 제외(불용어 제거하기 이전)) 서비스, 직원분들, 깔끔, 편하다 등의 서비스 관련 단어들이 고객들의 리뷰에서 크게 눈에 띄는 것을 확인하였다.

그림4) 전체 말뭉치의

그림5)top30단어의 wordcloud

저장된 변수를 활용하여 워드클라우드를 생성하고, 해당 data에서 column을 추출하여 비교 대상을 line plot, bar plot 등으로 시각화를 진행하였다. 내장된 TF-IDF함수를 통해 단어의 가중치를 확인하여 호텔 리뷰에서는 어떤 키워드가 중요하게 여겨지는지를 확인할 수 있었다. TF-IDF의 텍스트벡터화 (TF-idfVectorizer, CounterVectorizer)를 이용하여 리뷰 내의 특정 단어의 중요도를 판단할 수 있었다.

```
['갑니다', '깔끔하고', '깨끗하고', '너무', '넓고', '정말', '좋아요', '좋았습니다', '좋았어요', '호텔']
[[0.        0.        0.        ... 0.        0.        0.       ]
 [0.        0.        0.        ... 1.        0.        0.       ]
 [0.        0.        0.        ... 0.652289  0.        0.       ]
 ...
 [0.        0.        0.        ... 0.        0.        0.       ]
 [0.        0.        0.        ... 0.        0.        0.       ]
 [0.        0.        0.        ... 0.        0.59893857 0.       ]]
```

➔ Tf-idf의 결과에서는 wordcloud의 단어 가중치와 비슷하게 넓은, 깨끗, 깔끔의 키워드에서 높은 수치를 나타내는 것을 확인할 수 있다.

5. 결론

소셜 미디어의 발달로 소비자 들은 정보 공유와 확산이 빠르게 이루어지고 있고 호텔리뷰는 호텔 이미지와 평판에 영향을 미치기 때문에 지속적인 호텔 품질 제공과 브랜드 충성도 유지, 신규 고객 유치를 위해서 해당 텍스트에 나타난 고객의 리뷰를 통해 중심 키워드를 파악하는 것이 중요하다.

분석 결과, 소비자들은 생성한 온라인 리뷰에서 접근성이나 호텔의 여러 상품보다 “깔끔함, 깨끗함, 청결, 직원들의 서비스” 키워드의 중심성이 높았고 이것이 지켜졌을 때의 호텔의 만족도가 더욱 높은 것을 확인할 수 있었다.

호텔의 리뷰 상품과 관련된 속성이 고객(소비자)에게 중요한 평가대상으로 다루어지고 있음을 알 수 있다. 고객에 최종 구매결정에 영향을 끼치기 위해 온라인 리뷰 활용에 실무적인 시사점을 제공하였다. 아울러 텍스트마이닝 이론을 접목하여 도출된 호텔 리뷰의 현황을 설명할 수 있다는 측면에서 사회적 유용성 등 이론적 시사점을 함께 제시하고 있다.