경영전문석사학위 논문

호텔리뷰데이터를 활용한
호텔 키워드 추출

２０２２년 １ 월

서울과학종합대학원대학교

권 정 현

경영전문석사학위 논문

호텔 리뷰데이터를 활용한 호텔 키워드 추출

２０２２년２월

서울과학종합대학원대학교

권 정 현

호텔리뷰데이터를 활용한 호텔 키워드 추출

지도교수 장 중 호

이 논문을 경영학 석사 학위논문으로 제출함
2022년 2월
서울과학종합대학원대학교
권 정 현

권정현의 석사 학위논문을 인준함
2022년 1월

위 원 장 _____ 김보영 _____ (인)

위 원 _____ 문달주 _____ (인)

위 원 _____ 장중호 _____ (인)

초 록

대부분의 소비자들을 상품을 구매할 때 리뷰를 참고한다. 공급자 입장에서도 소비자가 리뷰를 확인하고 이에 따라 제품을 구매할 가능성이 높기 때문에 양질의 리뷰를 확보하기를 원한다. 공급자와 소비자 양측에서 리뷰의 중요성은 높으나, 모든 리뷰를 확인하기에는 어려움이 따른다. 따라서 본 연구에서는 호텔리뷰데이터를 바탕으로 각 호텔별 핵심키워드를 추출할 수 있는 토픽모델링 기법인 LDA를 포함하여 Countvectorizer, Textrank, TF-IDF모델로 키워드를 추출하고 각 모델별 성능을 비교한다. 선행연구의 경우 감성분석 또는 네트워크 분석을 활용하여 호텔산업 전반에서 특징이 되는 키워드 또는 감정단어를 분석했으나, 본 연구에서는 각 호텔별 특징이 되는 키워드를 추출한다는 점에서 차이를 보인다. 분석에 사용하는 데이터는 데일리호텔에서 186개 호텔을 대상으로 약 9만 건의 데이터를 수집하였고, 전처리 및 후처리를 진행하여 102개 호텔 6만건의 데이터를 가지고 분석을 진행했다. 모델별 성능은 uniqueness를 비교하여 TF-IDF모델이 가장 높은 성능을 보였다. 호텔별 높은 uniqueness의 영향을 주는 특성은 리뷰별 길이보다는 리뷰의 개수가 더 큰 영향을 미쳤다. 본 연구 결과를 통해서 호텔별 전체리뷰를 확인하지 않고도 키워드로 호텔의 특징을 알 수 있게 되어 공급자와 수요자 모두의 편의성을 도모할 수 있을 것이다.

핵심용어 : 호텔리뷰, TF-IDF, LDA, Textrank, 키워드 추출

목 차

표 목 차/그 림 목 차

제 I 장 서 론

1.연구의 배경 및 목적

많은 소비자들이 제품을 구매할 때 공급자가 제공하는 정보 외에 다른 소비자가 제품을 이용한 후기나 리뷰를 참고한다. 엠브레인 2017 소비자 리뷰 영향력 조사에 따르면 대부분의 소비자(86.9%)들은 소비자 리뷰의 필요성에 공감하고, 69.3%는 소비자 리뷰를 신뢰한다고 답했다. 또한 전체 78.6%는 제품을 구매 시 항상 소비자 리뷰를 확인한다고 답했다. 따라서 소비자들에게 리뷰는 제품 구매를 결정하는 데 중요한 역할을 한다.

공급자입장에서는 어뷰징이 아닌 신뢰성 있는 리뷰만을 소비자에게 제공하여, 제품 구매를 유도하려고 한다. 호텔업계에서는 보다 양질의 리뷰를 확보하기 위해서 각각의 정책을 실시하고 있다. 실제 이용자만 후기를 남길 수 있도록 수정하여, 어뷰징을 방지하고 100자 이상의 리뷰를 남기는 경우에 포인트를 제공하는 방법 등을 사용하고 있다. 야놀자의 “바른 후기”는 숙소 이용 후 14일 이내에 후기 작성이 가능하고 최대 2500포인트를 제공한다. 여기어때의 “리얼리뷰”는 100자 이상의 후기를 남기면 1000포인트를 제공하는 운영정책을 변경하여 고품질의 리뷰를 확보하려는 시도를 하고 있다. 데일리호텔 또한 “트루리뷰”를 통해 실이용객의 후기를 확보하려는 노력을 보이고 있다.

공급자와 소비자 모두에게 리뷰는 중요한 역할을 하고 있으나, 모든 리뷰데이터를 읽어보는 것은 쉽지 않다. 호텔과 예약플랫폼에 따라 다르지만 많은 리뷰가 등록되어 있는 호텔의 경우 4천 건 이상의 리뷰가 등록되어 있어 모든 리뷰를 읽어보기 어렵다. 또한 일부 리뷰의 경우 “좋았다”, “감사합니다”와 같은 짧은 리뷰이거나 정보의 가치가 부족한 경우도 존재한다. 따라서 리뷰를 효율적으로 탐색할 수 있다면 소비자입장에서는 호텔을 선택함에 있어 많은 리뷰를 찾아보지 않아도 되어 피로도를 줄일 수 있으며, 호텔종사자의 경우 문제점을 개선하거나, 호텔을 강조하기 위

한 키워드를 찾고자 할 때 보다 쉽게 리뷰를 탐색할 수 있다.

따라서 본 연구의 목적은 호텔별 리뷰데이터를 활용하여 호텔별 핵심 키워드를 추출하고 토픽모델링인 LDA를 포함하여 머신러닝 모델들 간의 성능 비교를 통해 보다 차별화된 모델을 선정하는 것을 목적으로 한다. 본 연구는 6단계로 진행한다. 1단계에서는 연구의 목적 및 관련 이론을 제시한다. 2단계에서는 분석에 필요한 데이터를 확보하기 위해 호텔예약 사이트에서 리뷰데이터를 크롤링한다. 3단계에서는 확보한 데이터를 정제하고 4단계에서는 모델별로 호텔의 특징을 추출한다. 이 단계에서는 모델별 특징 및 결과를 간략히 제시한다. 5단계에서는 최종적으로 사용한 모델의 특징 및 활용방안에 대해 제시한다. 마지막으로 6단계에서는 본 논문의 내용을 요약하고 시사점을 제시한다.

제 Ⅱ 장 이론적 배경

1.토픽모델링

토픽모델링은 문서의 집합의 추상적인 주제를 발견하기 위한 통계적 모델 중 하나로, 텍스트 본문의 숨겨진 의미구조를 발견하기 위해 사용되는 텍스트 마이닝 기법 중 하나이다. (위키백과) 다양한 분야에서 토픽모델링을 활용하여 리뷰, SNS데이터를 분석하여 유의미한 결과를 도출하려고 하고 있다. 차윤정, 이지혜, 최지은, 김희웅(2015)는 최신스마트폰에 대한 트위터 데이터 토픽모델링을 수행하여 마케팅전략 수립을 지원하는 연구를 진행했다. 김광국, 깅용환,김자희(2018)은 모바일 쇼핑앱 리뷰를 분석하여 충성고객과 불평고객을 심층 분석하였다. 박준형,오효정(2017)은 기록관리학 연구동향 분석을 위해 LDA와 HDP 모델을 비교 분석했다.

토픽모델링을 활용한 이전연구는 대부분 연구 동향을 분석하거나, 리뷰, SNS데이터를 분석하였다. 본 연구는 박준형,오효정(2017)과 같이 다양한 토픽모델링을 비교 분석하고 호텔업에 가장 적절한 모델을 활용하는 것을 목적으로 한다.

2.관련연구

호텔관광 분야에서 데이터를 사용하여 호텔 및 호텔리뷰를 분석하려는 시도가 진행되어왔다. 이병철, 변효정(2014)는 실험설계법을 이용하여 소비자리뷰가 관광상품 구매와 관련한 의사결정에 미치는 영향을 검증하려고 시도했다. 점차 빅데이터분석이 대두되면서 정량적데이터를 활용하는 단계에서 그치지 않고 정성적인 텍스트분석을 시도하는 경향이 두드러지고 있다. 특히 감성분석을 중점으로 연구가 진행되어 왔다. 임영희, 김홍범(2019)은 트립어드바이저에서 국내 15개 호텔에 외국인관광객이 남긴 영문리뷰를 대상으로 감성분석을 진행하였으며, Harvard IV 어휘집(lexicon)을 근간으로 감성사전을 구축하여 분석을 진행했다.

곽민정, 최지유, 박소현(2019)는 국내 18개 호텔 4,203개의 국문리뷰를 서비스 품질속성사전을 정의하고 KNU 한국어 감정사전을 기반으로 감성사전을 구축하였다. 박영욱, 정규엽(2021)은 토픽모델링 중 하나인 DMR을 이용해 34개 호텔의 20,094건의 리뷰데이터를 사용하여 감성분석을 진행했다. 김도경, 김인신(2017)은 텍스트 네트워크 분석을 통해 호텔 선택속성에 따른 연결관계를 확인하였다.

기존의 호텔리뷰 분석은 감성 및 네트워크 분석이 중심으로 연구가 이뤄졌으며, 각 호텔의 특징보다 호텔산업전반에서 중요키워드 또는 감정키워드를 도출하는데 연구가 이뤄졌다. 그러나 본 연구에서는 호텔산업 전반이 아닌 각 호텔의 중요키워드를 추출하는데서 차별점을 갖는다. 또한 토픽모델링인 LDA를 포함하여 머신러닝 모델들 간의 성능 비교를 통해 보다 차별화된 모델을 선정하는 것을 목적으로 한다.

제 Ⅲ 장 연구 설계

1.데이터수집 및 데이터 정제

1-1.데이터수집

본 논문에서 사용된 데이터의 수집은 호텔 예약사이트 데일리호텔에서 제주도에 있는 호텔 185개를 대상으로 데이터를 수집하였고 총 89,227개의 리뷰데이터를 수집했다. 이 데이터 중 텍스트의 양이 작거나(5자 이하, 예 : “좋아요”) 한글이 아닌 특수문자, 단자음모음(예 : “ㅋㅋㅋㅋ!”)은 제외하여 102개 호텔, 60,898개의 리뷰데이터를 사용했다.

1-2.데이터클리닝

리뷰데이터 중 빈번하게 발생하는 오탈자나 축약어로 인해 본래 같은 단어이나 같은 단어로 인식 못하는 경우 변환하여 같은 단어로 인식시켰다.(스벅, 스타벅스 등) 어미에 따라 의미가 같음에도 같은 단어로 인식하지 못하는 경우 한 단어로 인식시키기 위해 어간을 추출하여 일반화를 진행했다. 동사, 부사의 경우 한 단어로 의미를 도출해내기 어려우므로 명사, 형용사만 추출하여 분석에 활용했다.

1-3.불용어사전

너무 빈번하게 등장하는 단어(‘방’,호텔‘)의 경우 모든 호텔에서 등장하여 특징을 도출하기 어려우므로, 해당 단어들의 경우 제외하도록 불용어 사전에 넣어 제외하고 분석했다. 추가적으로 각 모델에서 호텔별로 상위 키워드 10개를 추출하고 해당키워드가 40개 호텔(전체 호텔 중 40%)에서 등장하는 키워드인 경우 불용어 사전에 추가하여 재모델링을 실시했다.

1-4.최종사용데이터

최종적으로 사용한 데이터는 5자 이상의 리뷰데이터를 사용하고 초성 및 특수문자는 제거했다. 형태소분석기에 등록되지 않은 단어를 쉽게 넣을 수 있는 Customize-konlpy 모듈을 사용하여 추가 단어를 넣고 Konlpy에서 제공하는 OKT를 사용하여 명사, 형용사를 추출했다.

2.사용 모델

모델에 따른 성능 비교를 하기 위해 countvectorizer, TF-IDF, LDA, TextRank 모델을 통해 데이터를 분석하고 그 결과를 비교하여 호텔리뷰 데이터를 분석하기에 가장 적절한 모델을 도출한다.

2-1.countvectorizer

countvectorizer는 각 문장에서 단어출현횟수를 카운팅하여 벡터화하는 방법이다. 등장한 단어 개수 X 문서의 길이 수의 매트릭스를 생성한다. 본 연구에서는 호텔별로 Countvector를 진행하여 호텔별 빈도가 높은 단어를 추출하되, 다른 호텔에서도 자주 등장하는 단어는 호텔의 특징을 나타낼 수 없다고 판단한다. 따라서 자주 등장하는 단어는 stopword를 지정하여 카운팅하지 않게끔 했다.

2-2.TF-IDF

TF-IDF(Term Frequency-Inverse Document Frequency)는 특정 문서 d에서 특정 단어 t의 등장 횟수를 구하는 tf값과 특정 단어 t가 등장한 문서의 수 df의 역수를 취한 idf를 곱하여 특정 문서에서 자주 등장하는 단어의 중요도가 높다고 판단한다. 보통 TF-IDF는 문서의 유사도를 구하는데 사용되나, 본 연구에서는 각 호텔별 리뷰데이터를 하나의 텍스트로 보고 TF-IDF를 진행한다. TF-IDF의 특성상 특정단어 t가 등장한 문서의 개수인 df값이 커질수록 TF-IDF값이 작아지므로 상대적으로 각 호텔에서만 등장하는 단어의 TF-IDF값이 높아져, 특징을 도출할 수 있다.

2-3.LDA

LDA(Latent Dirichlet Allocation, LDA)는 토픽모델링의 대표적인 모델링으로 주어진 문서에 대하여 어떤 주제들이 존재하는 서술하는 모델이다.(위키백과) 연구자가 k개의 주제를 임의로 지정해야 하므로 연구자의 주관이 많이 개입되는 단점이 존재한다. 호텔별로 LDA를 진행하고 각 리뷰를 분리하는 목적이 아닌 특징을 추출하는 것에 목적이 있으므로, k개

의 주제가 아닌 1개의 주제를 추출한다. 모든 호텔리뷰에서 등장하는 단어를 제외하기 위해 LDA 단어사전은 10번 이하 등장한 단어는 제외하고, 모든 문서의 절반에서 등장한 단어는 배제한다.

2-4.TextRank

Google의 PageRank를 활용한 알고리즘으로 하이퍼링크를 가지는 웹문서에 상대적 중요도에 따라 가중치를 부여하는 방법이다. TextRank는 이를 활용하여 문서내의 단어를 이용하여 Ranking계산하는 알고리즘이다. TextRank는 키워드 추출기능과 핵심문장 추출기능을 제공하는데, 키워드를 추출하기 위해 두 단어의 유사도를 정의하고 문장 내 많은 단어들과 연관이 있는 단어일수록 문서에서 중요한 의미를 가지고 있을 것이라고 추측한다.

제 Ⅳ 장 연구 결과

각 모델의 결과를 비교하기 위해 샘플로 사용할 호텔을 지정하고 결과를 비교하여, 어떤 모델에서의 결과가 가장 유용한지에 대해서 평가한다. 평가 방법으로는 해당 키워드가 다른 호텔과 비교했을 때, 핵심키워드로 적합지 여부를 평가하는 지표인 uniqueness를 사용한다. 산식은 전체호텔의 수(hn)를 호텔별로 키워드 등장빈도(ht)로 나눈 값에 리뷰별 키워드 등장빈도(rt)값을 해당 호텔의 리뷰 수(rn)으로 나눈 값을 곱한다. 해당 키워드가 다른 호텔에서 등장하지 않으면서 해당호텔에서는 자주 나타나는 키워드가 호텔을 대표할 수 있다고 판단한다. 수식은 다음과 같다.

$$uniqueness = \frac{\text{전체 호텔의 수}(hn)}{\text{호텔별 키워드 등장빈도}(ht)} X \frac{\text{리뷰별 키워드 등장빈도}(rt)}{\text{전체 리뷰의 수}(rn)}$$

<그림 1> uniqueness 수식

102개 호텔에 모델결과를 나열하기보다 리뷰의 개수가 가장 적은(104개) 평균개수 (633개) 가장 많은 (3315개) 3개 호텔을 활용하여 각각의 모델을 비교한다. 각 호텔에 대한 부정적인 리뷰가 있어 호텔명은 공개하지 않는다.

1. countvectorizer

countvectorizer를 실행한 결과는 아래와 같다. 키워드별로 살펴보았을 때 1420번 호텔은 방음, 냄새가 상위 키워드로 부각되었고, 29562번 호텔은 수영장, 공항이, 48829번 호텔은 전화, 사람이 중요 키워드로 추출되었다.

<표 1> countvectorizer 결과

code	1_word	1_count	2_word	2_count	3_word	3_count	4_word	4_count	5_word	5_count
29562	수영장	885	공항	533	가깝다	309	접근성	195	방음	150
1420	방음	43	냄새	25	가깝다	23	해변	20	깔끔하다	20
48829	전화	16	사람	12	사진	11	청소	10	오션	9

각각을 자세히 살펴보면, 29562번 호텔의 경우 수영장 키워드가 상위 키워드로 나타났으며, 수영장이 있다는 특징을 가지고 있다. 하지만, 102개 호텔 중에서 수영장을 보유한 호텔이 다소 있으므로 큰 특이점으로 보기는 어렵다.

<표 2> countvectorizer_29652번 호텔 리뷰

********** 수영장 **********
침구는 깨끗했어요~ 조식도 괜찮았어요~ 수영장은 체크아웃 연장 후(시간당 1만원) 10시 30분에 갔는데 사람이 없어서 좋았어요^^ 자쿠지가 있어서 아이랑 자쿠지에서 오래 있었네요: **수영장**이 사진만큼 근사하진 않아요ㅎ

공항에서 가깝고 **수영장**이용할수있어 좋았습니다

저흰 룸만 이용했는데 의외로 **수영장** 이용하시는 투숙객분들이 많아서 놀랐어요. 편의점가깝고 공항도 가까워서 편리했습니다

시설이 깨끗하고 직원도 친절하고 모두 만족스러웠어요. **수영장**,헬스장도 깨끗하고 좋았어요

가성비 최고! 서비스도 젛고 수영장도 잘 이용했습니다~ **수영장**은 가족 단위로 와서 놀때 좋은 것 같네요

1420번 호텔의 경우 방음이 상위 키워드였고 방음이 잘 안된다는 부정적인 리뷰가 중점을 이루고 있어, 방음시설이 약하다는 특징을 보여준다.

<표 3> countvectorizer_1420번 호텔 리뷰

********** 방음 **********
기대 안하고 갔는데 깨끗하고, 친절했어요 **방음** 부분에서는 살짝 아쉬웠어요

아빠와 함께한 여행이었는데, 침대도 2개 있어서 편하게 잘 잤어요:) 깨끗하다는 리뷰 하나 믿고 예약했는데 방, 침대 화장실 깔끔해서 만족했습니다. 가격도 가성비 좋습니다. 다만 창문이 두껍고 **방음**도 괜찮은데 도로가 바로 옆에 있어 자동차

나 오토바이 소음은 어쩔수 없이 들려요. 차로 3분 거리 이내에 검은모래해변이 있어서 좋았습니다.

깔끔하나 현관문 **방음**이 잘안되네요

침구깨끗했어요 약간 **방음**이 안됐어요

너무 방음이 안되서, 옆방 커플 이 바로 옆에서 자는듯했네요... **방음** 심각합니다... 개선이 정말 필요해요

48829번 호텔의 가장 상위 키워드인 “전화”가 들어간 리뷰를 살펴보면 “전화가 필요한 호텔이다”라는 특징을 얻을 수 있었다.

<표 4> countvectorizer_48829번 호텔 리뷰

********** 전화 **********

이런 호텔은 처음입니다. 도착해서 방키를 받아야하는데 자리비움안내와 전**화**번호딱 나와있고 아무도 없길래 **전화**를 수십통을 햇는데 전혀 받지를 않으시더라구요. 막연히 앞에서 계속 한시간 반가량 있다가 인포앞에 있는 키들고 에라 모르겟다 하고 들어갔어요 시간이 늦어서 다른데를 갈수도없고, 취소도 안되고 이런적은 처음입니다. 시설도 그닥이에요 체크아웃 시간에도 아무도 없어요 사진이 진짜 잘나왔네요. 그냥 모텔개조한 호텔입니다 제주도 와서 호텔금액 아깝긴 오랫만이네요 곧 망하실거 같아요

3박 연박했어요. 매일 청소해달라 했는데 첫날 안해주셔서 두번째날은 **전화** 따로 해서 부탁했어요. 그런데도 수건이랑 침구교체, 휴지통비우기가 끝이었고요, 화장실 머리카락도 그대로였어요. 다른 방은 모르겠는데 저희방에 주기적으로 담배냄새가 자꾸 들어와 확인해보니 환풍기가 뜯어져 있더라고요. 전반적으로 청소상태나 시설 관리 상태가 나빴습니다. 이건 조금만 신경쓰면 더 쾌적하게 할수 있는 부분인데 안되어있어서 이용객 입장으로 불쾌했어요. 콘센트도 3개인데 침대를 붙여놓은 벽쪽으로 애매한 위치에 달려있어 너무너무 불편했어요. 안되겠어서 냉장고 있는 뒤편 벽쪽에 콘센트 있는지 보느라 수납장을 살짝 앞으로 당겼다가 경악했네요. 쓰고난 ㅋㄷ이 바닥에 떨어져 있더라고요. 청소좀 제대로 하셨음 하네요. 진짜 너무 더러웠어요.

처음부터 예약한 방안내를 제대로 안해주셔서 문의**전화**를 5통넘게 했어요... 방에 난방도 고장나서 폭설내린 제주에서 벌벌 떨면서 자느라 너무 힘들었던 기억이 있네요 ㅠㅠ

체크인할 방이 본관인지 별관인지 설명되어있지 않아 헛걸음했고, 오션뷰했는데 시티뷰라 **전화**했더니 무인? 숙박이라 대응도 안되네요 무엇보다 방이 비좁아요

모르고 리뷰못보고 그냥갔는데.. 가격싼게전부입니다.. 방사진이랑 아예 다릅니다 오션뷰처럼 찍어노셧는데 3분의1오션뷰?? 멀리보면 살짝보입니다... 1회용품.세면용품등등 없습니다. 카운터에 팔지도않고 사람도없습니다.. 이름은 호텔인데 무인텔입니다. 방에있던거 생수2개.비누1개. 수건. 두루마리휴지1개 끝입니다 나머지 다챙겨가셔야되요 저것이외

필요한것들 카운터에 사람이 하루종일없어서 사장님께**전화**해야합니다.. 돌아오는 대답은 근처 편의점가서 사서쓰세요 술드시거나 차없으시는분들은 비추천합니다 방바닥보일로 35도로 하루종일틀었는데 차갑고건조합니다. 보일러안돌아가요 천장히터 풀로틀었는데 안따뜻해집니다. 바닷가에있는 3만원짜리모텔이더 난거같은..

각 호텔코드별로 uniqueness를 측정한 결과는 아래와 같다. uniqueness가 클수록 호텔의 특징을 잘 나타낸다고 할 수 있다. 48829호텔의 경우 다른 호텔에서 등장하지 않는 전화, 사람 키워드가 등장했으므로 uniqueness가 높게 측정되었고 29562번 호텔의 키워드 수영장, 공항은 다른 호텔에서도 등장하여 29562번 호텔만의 특징이라고 보기는 다소 어렵다.

<표 5> countvectorizer Uniqueness

code	1_word	1_uni	2_word	2_uni	3_word	3_uni	4_word	4_uni	5_word	5_uni
29562	수영장	0.71	공항	0.62	가깝다	0.38	접근성	0.38	방음	0.55
1420	방음	0.85	냄새	0.15	가깝다	0.14	해변	3.38	깔끔하다	0.4
48829	전화	11.77	사람	9.81	사진	8.83	청소	0.69	오션	0.26

countvectorizer는 단순 단어의 등장 빈도만을 측정하기 때문에 다른 호텔에서 빈도수가 낮은 단어의 경우 큰 특징으로 추출이 가능하다는 장점을 가지고 있다. 하지만, 리뷰의 개수가 많고 유사한 리뷰가 많을수록 특징을 잡아내지 못하는 단점을 가지고 있다.

2. LDA

LDA의 결과는 다음과 같다. countvectorizer의 결과와 다소 유사한 결과를 보여주고 있다. 29562번 호텔은 공항, 가깝다가 상위에 위치했으며, 1420는 방음, 가성비, 48829는 전화 괜찮다가 상위에 위치했다.

<표 6> LDA결과

code	1_word	1_lda	2_word	2_lda	3_word	3_lda	4_word	4_lda	5_word	5_lda
29562	공항	0.032	가깝다	0.018	아쉽다	0.012	접근성	0.012	괜찮다	0.012
1420	방음	0.011	가성비	0.009	전체적	0.008	괜찮다	0.007	냄새	0.006
48829	전화	0.005	괜찮다	0.004	사람	0.004	가성비	0.004	사진	0.004

가장 상위키워드가 들어있는 리뷰를 살펴보면 29562번 호텔은 “공항” 키워드가 가장 높았고 공항이 가까워서 좋았다는 리뷰가 주를 이뤄, 공항 + 가깝다가 상위 키워드로 나온 것으로 보인다.

<표 7> LDA_29652번 호텔 리뷰

********* 가깝다 **********

침구 푹신푹신 좋았구요 일회용품 없다고했는데 기본 샴푸린스바디등은 있습니당 전체적으로 깔끔하고 좋았으며 공항까지 택시타고4천원 안나왔어요ㅋㅋ **가깝고**좋았습니다

공항에서 **가깝고** 수영장이용할수있어 좋았습니다

방이나 침구류 등 청결상태는 좋았습니다. 다만 옆방과의 방음이 너무 안좋았습니다. 창문쪽에 옆방과의 틈이 작게 있는데 그 사이로 옆방에서 말하는 소리가 또렷하게 들려서 (무슨 말을 하는지 알 수 있을 정도) 방을 바꿔주셔서 그나마 잘 쉴수 있었습니다. 저층의 경우 창을 닫아도 자동차 소음이 비교적 잘들립니다. 주차 공간은 낮에는 넉넉한 편이지만 외출 후 밤에들어갔더니 자리가 없어서 라인 없는곳에 주차했습니다. 위치는 공항과 매우 **가까워** 좋습니다.

---- 저흰 룸만 이용했는데 의외로 수영장 이용하시는 투숙객분들이 많아서 놀랐어요. 편의점가깝고 공항도 **가까워서** 편리했습니다 ---------- ---- 룸 업그레이드해주셨고, 룸컨디션이 너무 깨끗하고 좋았습니다. 주변에 맛집들도 왠만큼 걸어다닐수 있었고, 공항도 **가깝고** 모든것이 좋았네요. 다만, 방음이 조금.. 아쉬웠네용;

1420번 호텔은 Countervectorizor와 같이 방음이 상위키워드에 위치하고 있고 가성비는 LDA모델에서 새롭게 등장하였으므로 가성비가 들어간 키워드를 살펴보면 가성비가 좋다는 리뷰가 많이 등장했다.

<표 8> LDA_1420번 호텔 리뷰

********** 가성비 **********
가성비는 좋은 것 같습니다. 근데 린스 없는 거랑 수건 쫌만 더 주셨음 좋겠어요,,,, ---------- ---- **가성비** 최고 입니다 ---------- ---- **가성비** 호텔! 위치대박 ---------- ---- 아빠와 함께한 여행이었는데, 침대도 2개 있어서 편하게 잘 잤어요:) 깨끗하다는 리뷰 하나 믿고 예약했는데 방, 침대 화장실 깔끔해서 만족했습니다. 가격도 **가성비** 좋습니다. 다만 창문이 두껍고 방음도 괜

찮은데 도로가 바로 옆에 있어 자동차나 오토바이 소음은 어쩔수 없이 들려요. 차로 3분 거리 이내에 검은모래해변이 있어서 좋았습니다. ---------- 주변 검은모래해변의 접근성도 좋고 **가성비** 좋아요~

48829호텔 또한 방음이 상위 키워드로 나왔으나, 괜찮다는 단어 또한 상위로 등장했다. “비교적 괜찮다”, “가성비는 있는 편”이라는 리뷰가 다소 있었다.

<표 9> LDA_48829번 호텔 리뷰

********** 괜찮다 ********** 화장실 물 수압도 너무 약하고 흠...바다가 보이는 방도 있겠지만 제가 묵었던 방은 잘봐야 바다가 보임 방은 많은데 주차칸 3칸..... 3~4만원이라면 **괜찮은** 방인데 그 이상은 좀 그런듯 앞으론 돈 조금 더 보태서 프랜차이즈호텔로 갈듯 ---------- 가성비 **괜찮았**어요~ ---------- 1. 근처에 올레시장이 있어서 저녁에 입이 심심하지 않음 2. 바로 앞에 산책할 수 있는 공원도 있어서 좋았음 3. 근처에 외돌개, 폭포 등.. 구경할 것도 많았음. 4. 올레7길 코스도 근처에 있어서 산책하기도 좋음. 5. 침구도 편안하고 커튼열면 밖에 풍경도 **괜찮았음.** 6. 걸어서 5분거리에 편의점도 있고 편했음. 7. 제주도는 역시 렌트해서 다니십기오 ㅎㅎ ----------

----

비대면이라서 전화연럭받았구요 그냥 **괜찮았어요**~ 3명 예약했는데 2명으로 오인하신거 빼고는 그냥 쏘쏘합니다. 조식이 없어서 그랬지만 그냥 쏘쏘합니다. 나중에는 조금더 알아보고 할께요~~^^

접근성, 위치 좋습니다. 다만 밤 늦게 9시 30분 이후 체크인을 하려하니, 직원이 안계시어 전화 통화해서 체크인하였습니다. 전화는 한번 만에 연결되었으나, 프런트가 밤 늦게 운용이 안되더군요. 체크인은 프런트 데스크 위에 호수가 적힌 카드들 중 호수를 찾아 들어갔습니다. 전반적으로 그럭저럭 **괜찮았으나**, TV가 있는 벽에 TV를 싸고 있는 좌측 유리 장식이 금이 크게 가있었습니다 (테이프로 고정되어 있었습니다). 욕조 벽에 금이 좀 가 있는 것 외엔 괜찮았습니다. 호텔 자체 시설은 크게 나무랄데가 없었는데, 제가 묵은 방의 문제라 생각됩니다.

각 호텔코드별로 uniqueness를 측정한 결과는 아래와 같다. 29562번 호텔은 공항이 가깝다는 특징을 보여주고 있으나, 이와 비슷한 특징을 가지고 있는 호텔이 있으며, 1420은 방음 이슈가 다른 호텔보다 있다고 판단할 수 있다. 마지막으로 48829호텔은 괜찮다는 단어가 등장했으나, 다른 호텔에서도 등장하여 호텔을 대표하는 단어로 부적절하다고 보인다.

<표 10> LDA Uniqueness

code	1_word	1_uni	2_word	2_uni	3_word	3_uni	4_word	4_uni	5_word	5_uni
29562	공항	0.82	가깝다	0.75	아쉽다	0.53	접근성	0.86	괜찮다	0.13
1420	방음	1.69	가성비	0.18	전체적	0.23	괜찮다	0.09	냄새	0.24
48829	전화	11.77	괜찮다	0.28	사람	9.81	가성비	0.35	사진	8.83

LDA는 문서에서 주제를 분류하는 모델이기 때문에 문맥을 고려하여 앞뒤 단어와의 연관성이 높은 단어를 상위단어로 지정하는 경우를 발견할

수 있었다. 48829 호텔의 경우 전화, 괜찮다가 같이 등장하는 리뷰가 다소 존재하여 괜찮다가 상위 단어로 올라온 것으로 보인다.

3. TF-IDF

countvectorizer와 LDA와는 다소 다른 키워드가 등장했으나, 다소 이해하기 어려운 단어들이 등장하여, 해당 키워드가 들어간 리뷰를 살펴보지 않으면 이해하기 어려운 단어들이 다소 등장했다.

<표 11> TF-IDF 결과

code	1_word	1_tfidf	2_word	2_tfidf	3_word	3_tfidf	4_word	4_tfidf	5_word	5_tfidf
29562	수영장	0.74	공항	0.37	수영모	0.15	쏘카	0.11	수모	0.07
1420	삼양	0.54	모래	0.21	해수욕장	0.18	해변	0.17	검다	0.13
48829	별관	0.34	외돌개	0.28	무인텔	0.22	본관	0.16	전화번호	0.15

29562 호텔에서는 다른 모델과 달리 수영모, 쏘카, 수모가 등장했으며, 수영모가 들어간 리뷰는 아래와 같다. 수영장에 들어가려면 수영모가 꼭 필요하다는 내용이 중점적이다. 호텔사이트에서 수영장이 있다는 점은 쉽게 알 수 있으나, 수영모가 필요하다는 점은 놓칠 수 있는데, 리뷰를 통해서 얻어낼 수 있는 특별한 정보로 보인다.

<표 12> TF-IDF_29562번 호텔 리뷰

********** 수영모 **********
1박2일로 짧게 하는 여행이라 숙박에 별 의미 안두기로 했는데 금액면 서비스 위치 조식 등등 다 좋은 듯 룸 타임 여러개로 금액별로 다르겠지만. 청소 상태 및 룸 컨디션은 좋았으나 뭔가 모를 쾌쾌한 담배냄새가 올라옴. 기분 나쁘다 하는건 아니라 치킨 사켜서 치킨 냄새로 탈취 함. 특히 숙박고객 수영장 이용이 좋음 수영 하실분은 수영복이

랑 **수영모** 챙겨가세요. 다음에 제주 이용시, 친구, 가족 등 모두 추천할 만 함.

위치가 너무 별로 방도 같은 급 효텔에 비해서 작아요 수영장도 자은데..**수영모**를 써야하는데 보통 가지고 다니지 않으니까 거기서 사야했어요 결국 추가요금이 발생하는셈이죠 미리 공지해 주지 않았으니까요 울며 겨자먹기로 사서 하거나 그냥 눈으로만보고 가거나...

다음엔 꼭 **수영모** 가져와서 수영할껍니다..!

수영장사용시 **수영모**를 써야되서 변거롭지만 깔끔했어요

고○항에서 매우 가까움! 시내 비지니스 호텔급! 실내수영장 **수영모** 필수(대여가능 3천원)

1420번 호텔도 이전모델들과 달리 새로운 키워드들이 등장했으며, 그 중 삼양이 들어간 키워드는 아래와 같다. 호텔주변에 삼양(검은모래)해수욕장이 있다는 점이 주요한 내용으로 보인다.

<표 13> TF-IDF_1420번 호텔 리뷰

********** 삼양 **********
삼양검은모래해수욕장 바로 위에 위치해 있어서 좋았어요

삼양해수욕장으로 가는길 바로 초입에 있어서 접근성 좋고, 호텔입구에 주차장도 있습니다. 객실과 화장실도 넓은편입니다. 좋은 가격에 하루 투숙했네요. ---------- ---- **삼양**해변 초입! 시설도 깨끗하고 방도 쾌적했음 ---------- ---- **삼양**해수욕장과 걸어서 10분정도 걸리는 위치구요, 뷰는 그닥인데 그냥 나름 깔끔하고 침대가 편한편이예요. 삼양 가야할일이 있어서 여기서 묵었고, 조식은 단체있을때만 있다고해서 못먹어봤어여~ 주차편리하고 다만 체크인시 너무 늦게오면 주차자리 없을 수 있는데 그래도 근처에 주차할데 꽤있구요, 바로 앞에 경찰서 맞은편에 교회 있어서 여자 혼자와도 뭐 안전한 느낌입니당ㅎㅎㅎ ---------- ---- 잘쉬다가요^^**삼양**좋네요

48829번 호텔 또한 이전과 다른 키워드들이 등장했으며, 그 중 별관을 포함한 리뷰는 아래와 같다. 호텔이 별관과 본관으로 나눠져 있고, 다소 구분하기 어렵다는 리뷰가 중점을 이루고 있다.

<표 14> TF-IDF_48829번 호텔 리뷰

********** 별관 **********
후기가 너무 안좋아서 기대 안했는데 냄새도 안나고 생각보다 나쁘지 않았습니다 다만 바닥이 장판이 아니고 신발이나 슬리퍼를 신고 생활해야해서 좀 불편했습니다 저희 가족은 **별관** 복층 이용했습니다 사진과는 마니 다른 뷰기는 했지만 거의 잠만 자고 나왔기때문에 큰 영향

은 없었고 올레시장이 가까워서 좋았고 가까운 곳에 선녀탕과 외돌개가 있어서 좋았습니다 실제 뷰 사진 올리니 참고하세요

체크인할 방이 본관인지 **별관**인지 설명되어있지 않아 헛걸음했고, 오션뷰했는데 시티뷰라 전화했더니 무인? 숙박이라 대응도 안되네요 무엇보다 방이 비좁아요

너무 별로였습니다 카운터에 사람도 없고 객실에 휴지도 사용한 거 하나밖에 없고 객실에는 곰팡이 냄새에 새벽에는 모기인지 벌레인지 모기인지 모르겠지만 기어다니는 것처럼 보이는것도 있고 날아다니는 것도 있고 아우 다시는 안가고 싶네요 별관은 청소좀 더 깔끔하기 하셔야 겠습니다~

호텔측과는 별개로 데일리호텔 이용하면서 너무 기분나쁜 체험을 했습니다. 방옵션에 외돌개 **별관** 써놓으면 다인가보네요. 외돌개 **별관** 뷰는 괜찮은데 싱글침대 두개가 더블마냥 딱 붙어있고, 샤워기에는 물때에 곰팡이, 세면대 밑 수납은 나무가 다 삭아서 문짝이 떨어져가고.. 직원이 프론트에 없고 하.. 그래서 데일리호텔에 전화해서 하루치 내고 퇴실하겠다 했더니 규정상 안된다. 계속 그말만 하더라구요 ^^ 정말 대단한 기업입니다. 그래놓고 한다는 말이 10% 적립금을 주겠다. 언제 어디로 놀러갈지도 모르고 특히 제주는 시간 내고 맞춰서 어렵게 와야하는데, 이따위로 서비스해놓고 가시 이용해달라며 적립금? 기도안차네요. 물건으로 따지면 불량 팔아놓고 환불불가 라며 안된다 하고.. 그럼 하다못해 다른데로 옮겨주기라도 하셔야하는데, 그마저도 제가 풍경호텔 본관에 전화해서 ok사인 받고 데일리호텔에 전화해서 요청했습니다.

고객알기를 정말 호구처럼 아는 기업 참 대단하네요 ^^ 다시는 데일리 호텔 안쓸겁니다. * 이것과는 별개로 호텔에서 바꿔주신 룸 컨디션은 만족하고 잘 쉬고왔습니다. 저 위에 평점은 외돌개 별관에 대한 평점입니다.

별관에서 잤는데, 안내사항이 없어서 본관으로 가는 바람에 2번 이동해서 불편하고, 본관에서 직원이 상주하고 있는게 아니라 체크인하려고 전화를 해도 연결이 안되서 너무 불편했어요. 따뜻한 물이 정말 안나와서 씻는데 불편했어요.

각 호텔코드별로 uniqueness를 측정한 결과는 아래와 같다. 29562번 호텔은 3번째 단어인 수영모의 uniqueness가 가장 높아 수영모가 필요하는 특성을 보여주 있다. 1420은 삼양, 모래, 검다, 해변이 높은데 이는 삼양검은모래해변이 각각 나뉘어서 이 호텔의 특징으로 볼 수 있다. 48829는 외돌개, 무인텔이 가장 uniqueness가 높은데 이로 인해 외돌개에 위치한 무인텔이 큰 특징으로 보인다.

<표 15> TF-IDF Uniqueness

code	1_word	1_uni	2_word	2_uni	3_word	3_uni	4_word	4_uni	5_word	5_uni
29562	수영장	0.88	공항	0.74	수영모	0.95	쏘카	0.29	수모	0.17
1420	삼양	1.93	모래	1.29	해수욕장	0.64	해변	0.85	검다	0.97
48829	별관	1.96	외돌개	2.94	무인텔	2.94	본관	1.47	전화번호	1.96

TF-IDF는 모델의 특성상 모든 리뷰에 등장한 단어는 값을 적게 하고 일부에만 등장한 단어에는 값을 높게 계산하기 때문에 해당 호텔에서만 등장하는 단어를 찾는 데는 용이한 것으로 보인다. 하지만 오탈자나 잘못 토크나이즈된 단어들 또한 높은 점수를 부여하므로 전처리, 후처리를 크게 요구하는 편이다.

4. TextRank

TextRank의 결과는 아래와 같다. Countvecrorizor와 크게 유사하며, LDA와는 다소 차이를 보이고 있다.

<표 16> TextRank 결과

code	1_word	1_tr	2_word	2_tr	3_word	3_tr	4_word	4_tr	5_word	5_tr
29562	수영장	17.60	공항	11.49	가깝다	5.35	접근성	3.53	방음	3.52
1420	방음	2.42	냄새	1.81	가깝다	1.59	주차	1.40	해변	1.30
48829	사람	2.16	청소	1.49	별관	1.48	전화	1.41	근처	1.32

29562번 호텔의 경우 countvectorizer와 동일한 결과를 보여주고 있다. 공항, 가깝다, 접근성은 서로 연관성이 있으므로 '방음'과 관련된 리뷰를 살펴보면 아래와 같다. 방음의 이슈가 있는 것으로 보인다.

<표 17> TaxtRank_29562번 호텔 리뷰

********** 방음 **********

나머지는 다 만족스러웠는데 **방음**이 너무 안좋아요 말하는소리도 들리고 드르륵 소리도 다 들리더라구요 심지어 새벽에 옆방에서 은밀한 그… 그소리까지 다 들려서 자다가 깼습니다… 다 좋았는데 소리에 예민하시면 비추천이요

방음이 조금 별로였고 뷰가 좋지는 않았지만 가성비 전체적으로 만족스러운 호텔입니다~

방이나 침구류 등 청결상태는 좋았습니다. 다만 옆방과의 **방음**이 너무

안좋았습니다. 창문쪽에 옆방과의 틈이 작게 있는데 그 사이로 옆방에서 말하는 소리가 또렷하게 들려서 (무슨 말을 하는지 알 수 있을 정도) 방을 바꿔주셔서 그나마 잘 쉴수 있었습니다. 저층의 경우 창을 닫아도 자동차 소음이 비교적 잘들립니다. 주차 공간은 낮에는 넉넉한 편이지만 외출 후 밤에들어갔더니 자리가 없어서 라인 없는곳에 주차했습니다. 위치는 공항과 매우 가까워 좋습니다. ---------- ---- 룸 업그레이드해주셨고, 룸컨디션이 너무 깨끗하고 좋았습니다. 주변에 맛집들도 왠만큼 걸어다닐수 있었고, 공항도 가깝고 모든것이 좋았네요. 다만, **방음**이 조금.. 아쉬웠네용; ---------- ---- 침구 깨끗하고 룸컨디션은좋았어요 꼭대기층에 있는 수영장도 좋았구요 그런데 잠을자려고누우니 옆방의대화소리가 같은방에 있는것처럼 잘들렸습니다 이렇게 **방음**이 안되도되나싶을정도였네요 꼭 개선되었으면좋겠습니다 ---------- ----

1420번 호텔의 키워드 중에서 '주차'는 TextRank모델에서 처음 등장했다. 주차장이 여부에 관한 리뷰가 주를 이루며, 주차장 여부는 쉽게 확인할 수 있는 정보이나, 넓이에 관한 정보는 리뷰에서만 얻을 수 있을 것으로 보인다.

<표 18> TaxtRank_1420번 호텔 리뷰

********** 주차 ********** 업무 및 주말부부라 매달 2-3회 이용중인데, 직원분들이 친절하셔서 좋

습니다. 다만 객실 청소면이 다소 미흡한 부분이있습니다. 지저분한 정도는 아니나 먼지뭉텅이가 묻어날때, 보일때 가 종종 있습니다. 동,서 주변 관광지 를 이동하기에 무리없고 차가 없으신분들도 숙소 5분거리 내 버스정류장이 있어 관광하기에도 좋습니다. (숙소 주차장 이 가득찼을땐 인근 공영주차장도 있으니 크게 **주차**문제의 대한 부담은 없을 것 같습니다.)

딱 가성비좋네요.또한 **주차**장이 넓어서 편리했어요

삼양해수욕장으로 가는길 바로 초입에 있어서 접근성 좋고, 호텔입구에 **주차**장도 있습니다. 객실과 화장실도 넓은편입니다. 좋은 가격에 하루 투숙했네요.

프런트 직원은 불친절 끝판왕에 조명은 다켜도 어두컴컴하고 샤워할때 샤워부스문도 없어 바닥에 물 홍건하고 샤워기도 수압이 별로 **주차**공간은 협소하고 마음에 드는게 하나도 없습니다! 다신 이용안하려고 합니다.

깨끗하고 좋습니다. **주차**장도 넓고

48829번 호텔은 countvectorizer에서 등장한 청소가 상위키워드로 나왔다. 청소가 들어간 키워드는 아래와 같다. 청소상태가 불결하다는 이슈가 주를 이루고 있다. 청결도 관련 이슈는 호텔이용에 민감한 문제이나, 호텔 측에서는 제공하지 않는 리뷰만 얻을 수 있는 정보인 점에서 잘 뽑힌 키

워드로 생각된다.

<표 19> TaxtRank_1420번 호텔 리뷰

********** 청소 **********
3박 연박했어요. 매일 **청소**해달라 했는데 첫날 안해주셔서 두번째날은 전화 따로 해서 부탁했어요. 그런데도 수건이랑 침구교체, 휴지통비우기가 끝이었고요, 화장실 머리카락도 그대로였어요. 다른 방은 모르겠는데 저희방에 주기적으로 담배냄새가 자꾸 들어와 확인해보니 환풍기가 뜯어져 있더라고요. 전반적으로 청소상태나 시설 관리 상태가 나빴습니다. 이건 조금만 신경쓰면 더 쾌적하게 할수 있는 부분인데 안되어 있어서 이용객 입장으로 불쾌했어요. 콘센트도 3개인데 침대를 붙여놓은 벽쪽으로 애매한 위치에 달려있어 너무너무 불편했어요. 안되겠어서 냉장고 있는 뒤편 벽쪽에 콘센트 있는지 보느라 수납장을 살짝 앞으로 당겼다가 경악했네요. 쓰고난 ㅋㄷ이 바닥에 떨어져 있더라고요. 청소좀 제대로 하셨음 하네요. 진짜 너무 더러웠어요.

2박 묵어서 중간일에 **청소**를 해주시지 마셔요 했는데 **청소**를 해주셔서 난감한 점이 아쉬었지만 저렴한 가격에 잘 쉬었습니다 문앞 비상등이 흰빛으로 계속 켜져있어서 숙면에 안타까웠지만 룸 업그레이도 해주셔서 좋았습니다 인터넷 티비도 좋았고요 저렴한 삼푸와 샤워코롱은 아쉬웠지만 욕조가 있는 룸이라 좋았습니다 404호 변기 앉는 곳이 덜렁거리더군요 비데는 안 바라지만 덜렁거림을 잡아야겠습니다 샤워기도 자꾸 왼쪽으로 치우치는 것도요 그래도 아침에 창문 열고 왼쪽 보면 보이는 한라산이 절경이었습니다!!

청소는 되어있으나 여기저기 지저분함. 건물입구에 쓰레기를 모아놨다

던가 엘레베이터에 손자국이 가득하다던가 .. 침구가 깨끗하긴한데 먼지가 많이 앉았음. 이름은 호텔인데 샴푸랑 바디워시, 비누, 수건만 있었음. 사진이랑 동일한데 너~~~무 잘찍어놔서 인스타셀기꾼을 실제로 본것 같았네요.배신감 ㅠ

방 배정받을때 프런트에 사람이 없어요 방배정 전화로 받고 스스로 알아서 키챙겨서 들어가야합니다.. **청소** 그런건 안되어있고요 그냥 알아서 다 해야합니다 숙소가 없어 그냥 하루 보내려 갔다가 고생한 기억이..

주차장 작아서 당황했지만 길가주차 가능햐서 다행였어요 그런데 청결은 좀 신경 쓰셔야 할듯요 닦지않은 전기포트, 욕실에 구석구석 물때, 3박내내 **청소**상태까지 큰 기대도 안했지만 기본은 해 주시길 바랍니다, 수건도 낡아서인지 너무 거칠었습니다 ㅠㅠ

TextRank의 Uniqueness결과는 countvectorizer와 유사하다. 모델결과가 유사했기 때문에 유사한 결과를 보이며, TextRank에서 uniqueness값이 가장 높은 단어가 최종적으로 호텔별 특징을 나타낸다고 할 수 있는데, 29562번 호텔의 특징으로 수영장, 1420번 호텔은 해변, 48829호텔에서는 전화의 해당키워드가 호텔의 중심키워드로 추출되었다.

<표 20> TextRank Uniqueness

code	1_word	1_uni	2_word	2_uni	3_word	3_uni	4_word	4_uni	5_word	5_uni
29562	수영장	0.71	공항	0.6	가깝다	0.31	접근성	0.43	방음	0.44
1420	방음	0.68	냄새	0.15	가깝다	0.12	주차	0.23	해변	1.69
48829	사람	9.81	청소	0.46	별관	5.88	전화	11.77	근처	0.74

Uniqueness값 또한 countvectorizer와 유사한 결과를 보이고 있다. 1420번 호텔에서 등장한 “냄새”에 대한 uniqueness값은 0.15로 동일하며, 48829번 호텔에서 등장한 “사람”에 대한 수치도 9.81로 동일하다. 이로 인해 TextRank가 countvectorizer와 같이 단어의 빈도수에 따라 큰 영향을 받는 모델임을 추측할 수 있다.

5. 종합

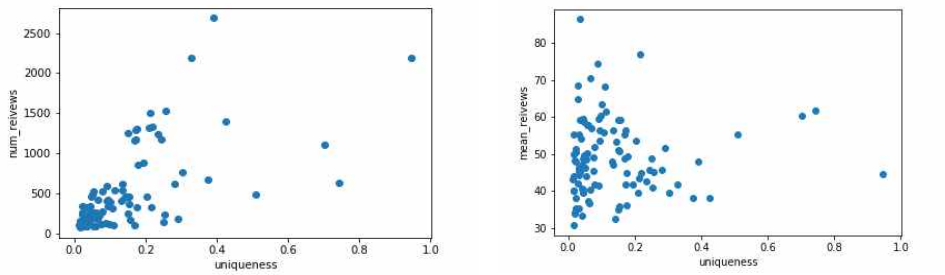
각 모델의 성능을 비교하기 위해 각 호텔별 uniqueness값을 구하고 이 값의 평균으로 각 모델의 성능을 비교한다. 각 호텔별 uniqueness는 상위 5개의 키워드의 uniqueness값에 가중치를 곱하여 산출한다. 가중치를 부여하여 같은 키워드가 등장했더라도 상위키워드에 위치할수록 높은 점수를 부여받도록 설정했다.

모델별 uniqueness값은 TF-IDF가 가장 높고 Textrank, LDA, countvectorizer순으로 나타났다. 특히 TF-IDF는 다른 모델 대비 높은 uniqueness값을 보이고 있어 다른 모델 대비 차별화된 키워드를 더 잘 찾는 것으로 확인되었다.

<표 21> 모델별 uniqueness

CountVectorizer	LDA	Textrank	TF-IDF
0.126	0.127	0.128	0.134

각 호텔별로 uniqueness을 산출했을 때 호텔의 uniqueness는 리뷰가 많은 호텔의 경우 uniqueness가 높은 경향을 보이고 있어 리뷰의 개수가 호텔의 차별화딘 키워드를 뽑는데 영향을 주는 것으로 나타났다. 반면 리뷰의 길이는 큰 영향을 주지 않았다. 리뷰를 길게 작성했더라도 특별한 정보가 없이 “침구 및 바닥 등 전체적으로 깨끗했어요”와 같은 호텔에서의 일반적인 정보가 나열되어 있는 경우 높은 uniqueness를 얻기 어려웠기 때문으로 확인된다.



<그림 2> 리뷰개수별, 리뷰 평균길이별 uniqueness

제 V 장 결론

1.분석 결과와 시사점

본 연구에서는 호텔리뷰에서 호텔별 키워드를 추출하여 호텔별 특징을 추출하는 모델을 만들고 각 모델을 비교해보았다. CountVectorizer, LDA, Textrank TF-IDF 중 TF-IDF모델이 대표 키워드를 추출하는데 가장 높은 성능을 보였다. TF-IDF를 제외한 다른 모델들은 유사한 키워드를 추출하고 uniqueness을 보여 성능의 큰 차이를 보이지 않았다. 하지만 TF-IDF에서 추출되지 않은 키워드 또한 의미를 가지고 있어 TF-IDF를 보조하는 역할로 사용할 수 있다.

호텔별 uniqueness를 추출할 때 리뷰의 평균 길이보다는 리뷰의 개수가 더 중요한 특성인 것으로 확인되었다. 일부 호텔예약 플랫폼에서는 100자 이상의 리뷰를 작성하기를 강제하고 있는데, 길이는 100자 이상이지만 평이한 리뷰인 경우 특별한 키워드를 추출해내기는 어렵다.

본 연구를 통해서 호텔관련업을 수행하고 있는 기업에서는 추출된 키워드를 통해 해당호텔의 장단점을 연구하여 장점을 강화하고 단점을 보완하는 전략을 세울 수 있을 것으로 생각된다. 또한 호텔을 방문하고자 하는 고객의 입장에서는 모든 리뷰를 읽어보기에는 피로감을 동반하기에, 핵심 키워드를 참고하여 숙박업체를 선정할 수 있을 것으로 생각한다.

2.연구 한계점 및 향후 연구 방향

본 연구의 한계점은 첫째, 지역을 제주도 한정하였다. 지역을 한정했기 때문에 성산일출봉과 같은 지역적 특색을 가지는 속성은 특별한 키워드로 뽑히기 어려웠을 것이다. 둘째, 102개의 호텔을 대상으로 분석했기 때문에 더 많은 호텔을 대상으로 분석을 진행했다면 다른 결과가 나왔을 수도 있었을 것이다. 셋째 전처리 및 후처리에 대한 로직이 모델 내에 반영되어 있는 것이 아니기 때문에 신조어나 오탈자에 민감하게 반응할 수 있기 때문에 추후 분석 결과가 상이할 수 있다. 넷째, 리뷰데이터의 품질과 양을 더 확보한다면 더 풍부한 결과를 얻을 수 있었을 것이다.

본 연구의 향후 연구과제는 첫째, 시계열 데이터로 구분하여 시간에 흐름에 따라 키워드가 변화하는 양상을 통해 해당 호텔이 문제점을 개선했는지 여부 등을 파악할 수 있을 것이다. 둘째, 데일리호텔 뿐만 아니라 야놀자, 여기어때 등 다양한 플랫폼에서 리뷰데이터를 확보하고 이를 취합하여 분석을 진행하면 더 많은 데이터 확보 및 플랫폼별 분석이 가능하다. 셋째, 호텔별 키워드에 집중하였기 때문에 감정 또는 평점분석을 연계한 분석을 진행하여 연구를 더 확장할 필요가 있을 것이다.

참고문헌

곽민정, 최지유, 박소현(2019), "호텔 서비스 속성별 고객만족도 분석을 위한 온라인 리뷰 감성분석" 관광경영연구, 90, 1-25

김도경, 김인신(2017), "텍스트 마이닝을 이용한 온라인 리뷰의 호텔 선택속성 분석" 관광학연구, 155, 109-127

박영욱, 정규엽(2021)," DMR(Dirichlet Multinomial Regression) 토픽모델링을 이용한 온라인 리뷰 빅데이터 기반 고객감성 분석에 관한 연구: 국내 5성급 호텔의 외국인 이용객 리뷰를 중심으로" 호텔경영학연구, 130, 1-20

이병철, 변효정(2014), "온라인 리뷰가 관광상품 구매행동에 미치는 영향 호텔리조트를 중심으로", 관광레저연구, 86, 59-79

임영희, 김홍범(2019), "호텔 온라인 리뷰 빅데이터를 활용한 감성분석에 관한 연구", 호텔경영학연구, 119, 105-123

최자영, 김현아, 김용범(2020), "온라인 리뷰가 매출에 미치는 영향력 분석: 텍스트 기반 감성지수를 중심으로", 유통연구, 25, 1-21

Rada Mihalcea, Paul Tarau(2004), “TextRank: Bringing Order into Texts”, Association for Computational Linguistics, 404 - 411

데일리호텔, 데일리호텔, https://www.dailyhotel.com/ 검색일 2022.01.27

안상준, 딥 러닝을 이용한 자연어 처리 입문, https://wikidocs.net/30708 검색일 2022.01.27

야놀자, 바른 후기 정책, https://www.yanolja.com/policy/review 검색일 2022.01.27

엠브레인, ‘호갱’이 되고 싶지 않은 소비자들, 습관처럼 ‘소비자 리뷰’ 확인해, https://www.trendmonitor.co.kr/tmweb/trend/allTrend/detail.do?bIdx=1599&code=0201&trendType=CKOREA 검색일 2022.01.27

위키백과, 토픽모델 https://ko.wikipedia.org/wiki/ 검색일 2022.01.27

lovit, "TextRank 를 이용한 키워드 추출과 핵심 문장 추출 (구현과 실험)", https://lovit.github.io/nlp/2019/04/30/textrank/ 검색일 2022.01.27

lovit "customized KoNLPy", https://github.com/lovit/customized_konlpy 검색일

2022.01.27
nate뉴스, 2021년 02월 17일자 "여기어때, 정보 품질 향상 위해 후기 정책 개편하자…고품질 리뷰 '쑥'"

Abstract

Extracting hotel keywords using hotel review data

Kwon Jeong Hyeon
Seoul School of Integrated Sciences and Technologies
Advisor: Chang Joong ho, Ph.D.

Most consumers refer to reviews when purchasing products. Suppliers want to get good quality reviews because consumers check reviews and after purchase products. Reviews are high importance on both suppliers and consumers, but it is difficult to check all reviews. Therefore, in this study, keywords are extracted with Countvector, Textrank, and TF-IDF models, including LDA, a topic modeling technique that can extract keywords for each hotel based on hotel review data, and the performance of each model is compared. In the case of previous studies, using emotional analysis or network analysis for throughout the hotel industry, but this study differs in that keywords characteristic of each hotel are extracted. The data used for the analysis were collected about 90,000 data from 186 hotels at the Daily Hotel and then after pre-processing and post-processing data, finally 60,000 data from 102 hotels. The TF-IDF model showed the highest perform-

ance by comparing the uniqueness by model. The characteristics that influence the high uniqueness of each hotel had a greater influence on the number of reviews than the length of each review. Through the results of this study, the characteristics of the hotel can be known through keywords without checking the overall review of each hotel, thereby promoting the convenience of both suppliers and consumers.

Key words: Hotel review, Keyword extract, TF-IDF, LDA, Textrank

Student Number: 2025418001