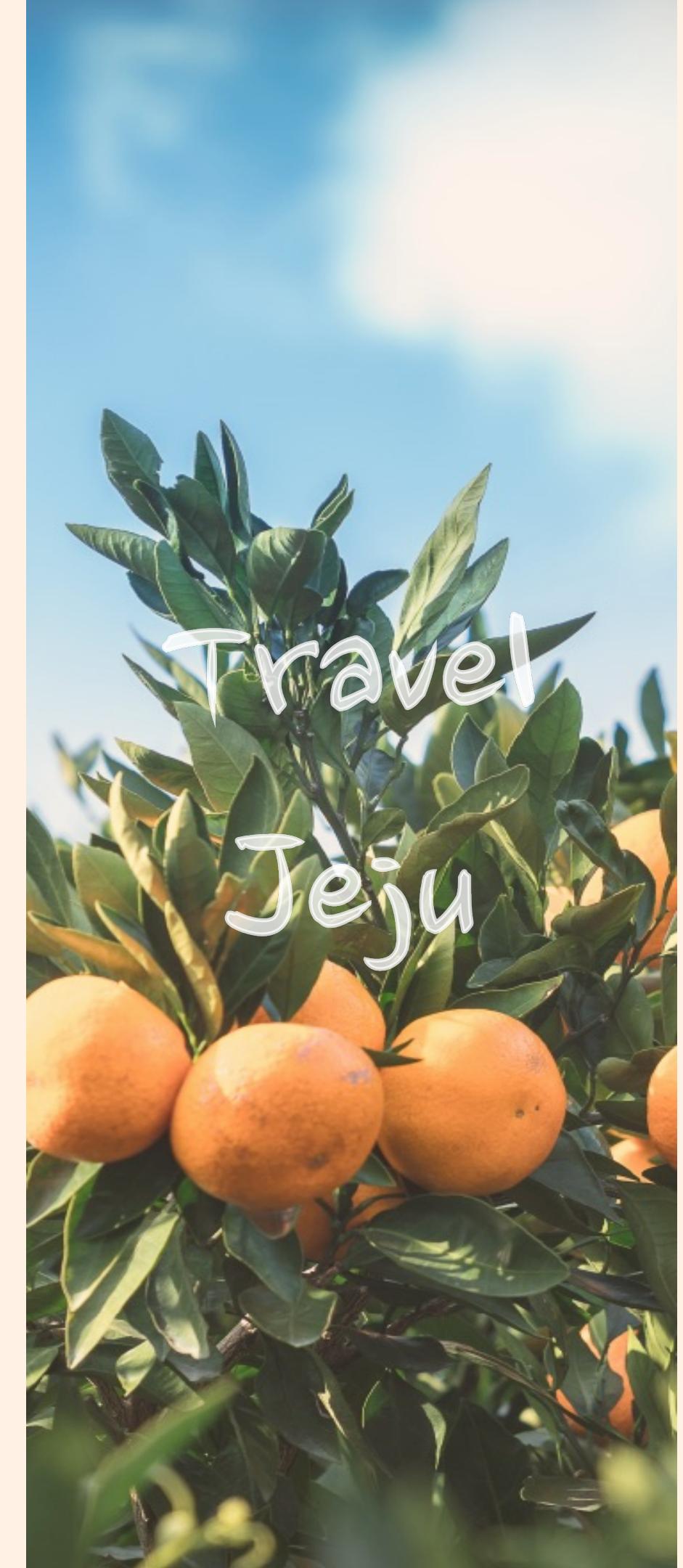


호텔 예약 사이트 리뷰 분석을 통한 고객 맞춤형 호텔 추천 시스템 - 제주 지역



a7ve - 지다영, 유경민, 이중호, 정새롬



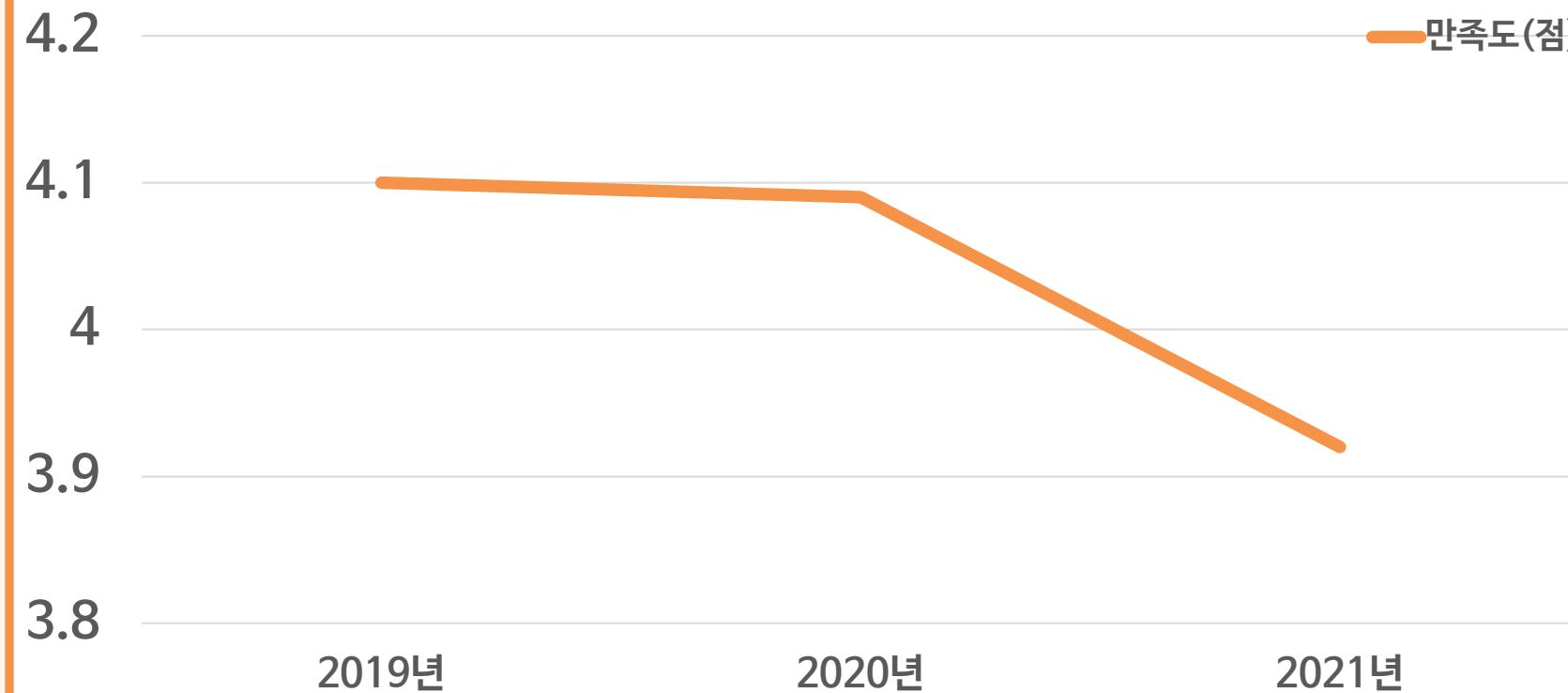
목차

01. 주제 선정 배경
02. 분석 개요
03. 프로젝트 과정
04. 최종 시스템 알고리즘
05. 추천 시스템 활용 예시
06. 분석의의 및 기대효과
07. 역할 분담

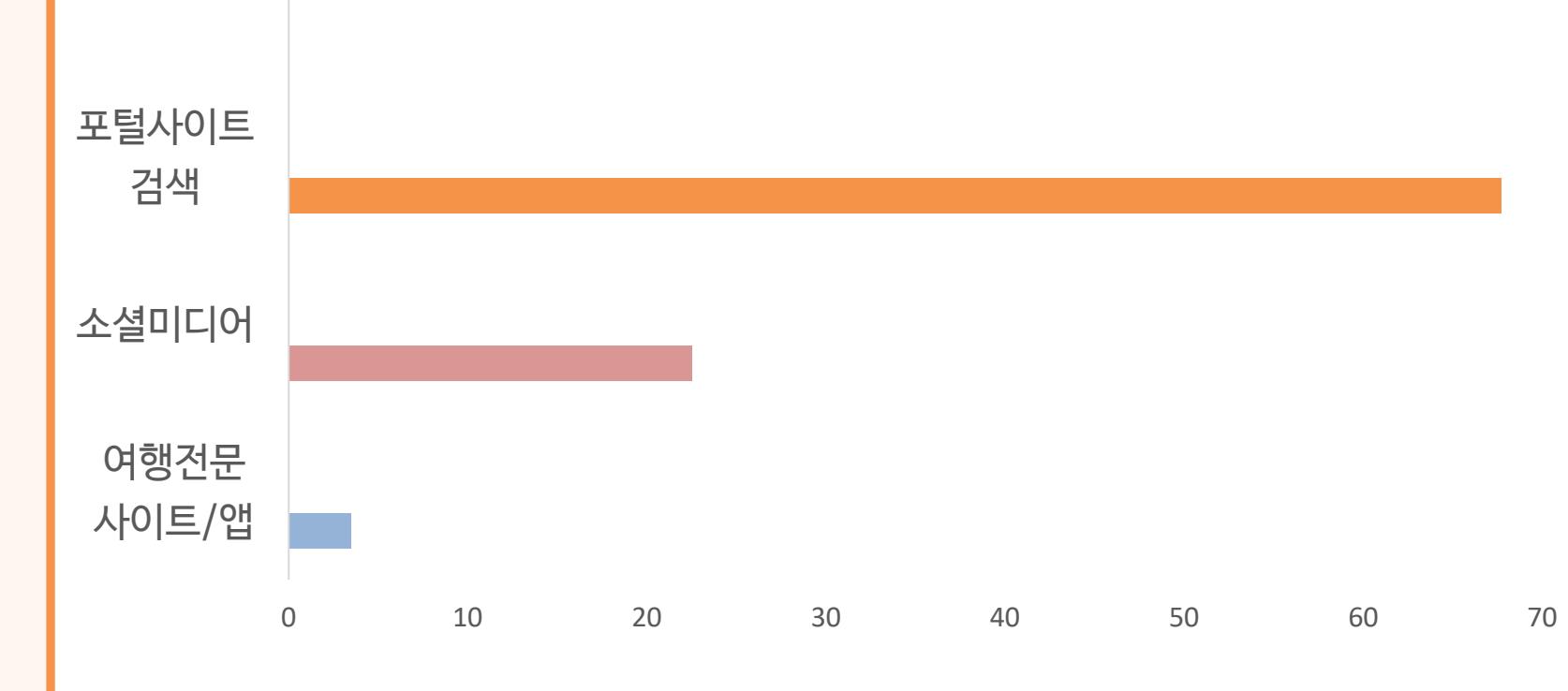
‘코시국’엔 외국 대신 제주…체류 기간 늘었지만 만족도는 떨어져

한겨례, 허호준 기자, 2021.04.09.

연도별 제주 여행에 대한 전반적 만족도

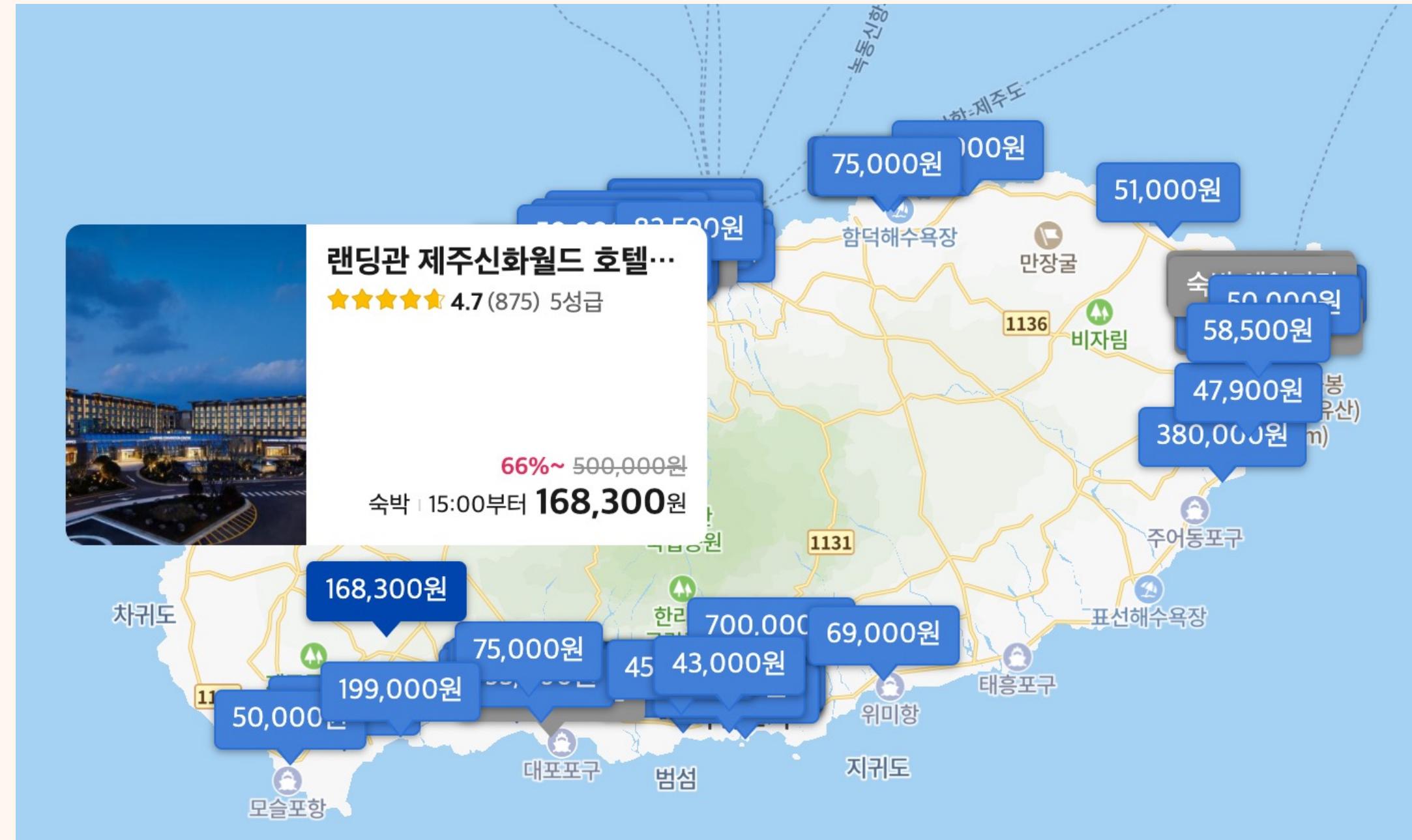


인터넷 정보습득경로



*출처 : 2020년 제주특별자치도 방문관광객 실태조사 분석보고서 | 제주관광공사

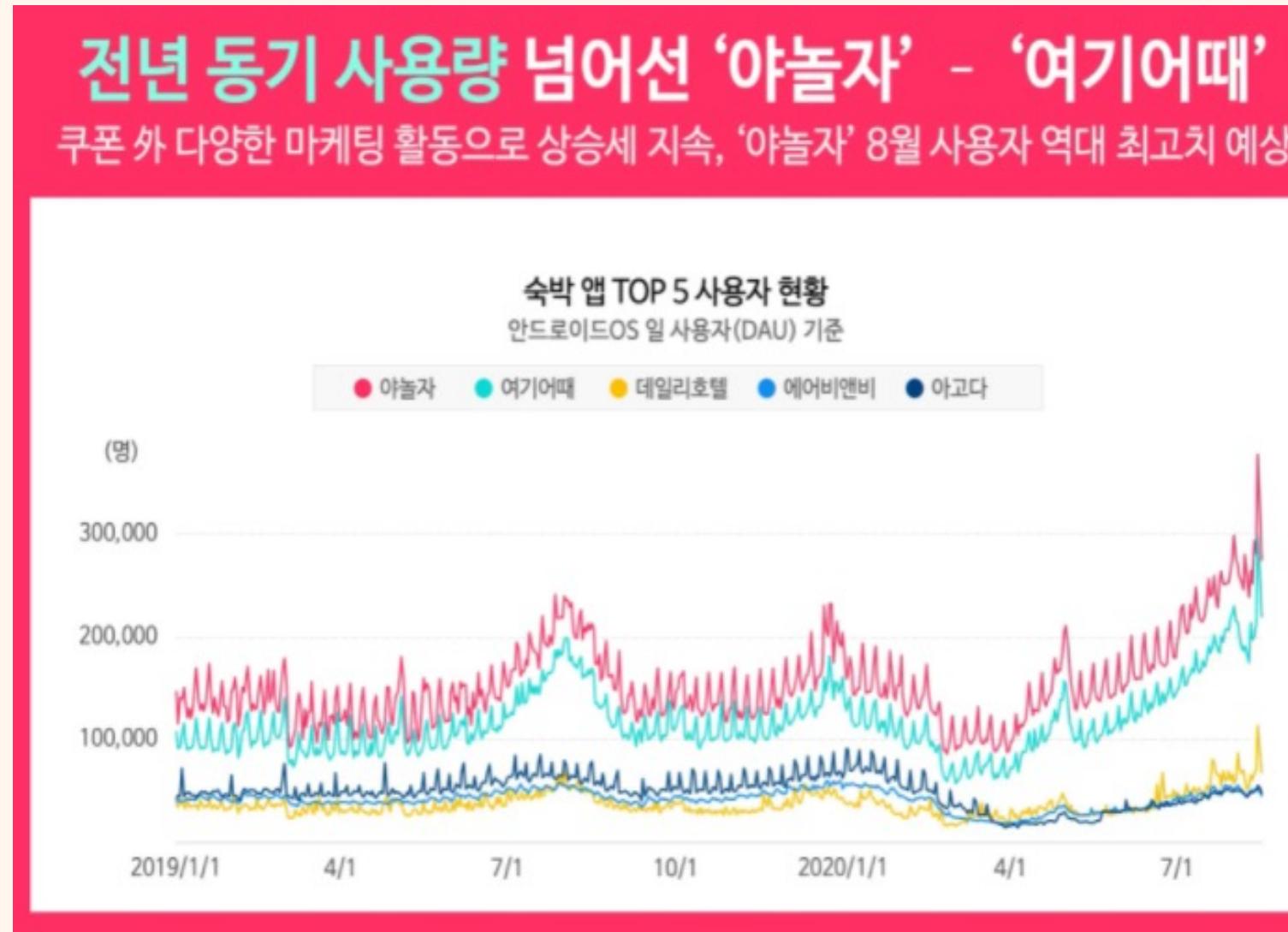
- 제주여행의 만족도는 점점 낮아지고 있는 추세 (2019년 대비 9.1% 하락)
- 여행정보 습득경로 중 한국의 인터넷 사이트/앱 64.4%
- 인터넷 정보습득경로에서 포털사이트 검색 67.7%, 소셜미디어 22.2%, 여행전문사이트 또는 앱 3.5%



*출처 : 야놀자



“나는 호텔을 선택할 때 청결이 가장 중요해.
그러나 호텔과 리뷰가 너무 많아 제대로 알 수가 없네.
내가 원하는 호텔을 편하게 찾을 수 있을까?”



*출처 : 아이지에이웍스 국내외 숙박 앱 사용자 현황 분석

호텔 성급 / 그 외 유형

5성급	<input checked="" type="checkbox"/>	4성급	<input checked="" type="checkbox"/>	3성급	<input checked="" type="checkbox"/>	2성급	<input checked="" type="checkbox"/>
1성급	<input checked="" type="checkbox"/>	리조트	<input checked="" type="checkbox"/>	가족호텔	<input checked="" type="checkbox"/>	비즈니스	<input checked="" type="checkbox"/>
레지던스	<input checked="" type="checkbox"/>	콘도	<input checked="" type="checkbox"/>				

0개 선택

테마

스파/월풀/욕조	주차가능	야외수영장	바다전망	커플PC	풀빌라	계곡인접	바베큐
애견동반	개별바베큐	수영장	조식운영(뷔페)	객실내PC	키즈	해수욕장인근	무료영화(OTT)

1 / 6

가격

1만원 ~ 50만원 이상

*출처 : 야놀자

- 여행예약사이트 이용률 1위 플랫폼 '야놀자' 선정
- 야놀자 사이트 내 호텔에서 제공하는 정보에 의한 필터 존재하지만, 개인화된 추천 시스템의 부재
- 제주 여행 만족도를 높이기 위해 호텔 리뷰 데이터를 기반으로 고객맞춤형 호텔 추천 시스템 개발

02 분석 개요

분석 목표

- 사용자 리뷰를 기반으로 호텔을 평가하여
호텔 이용자별 선호에 맞춘 호텔 추천

분석 채널

yanolja

분석 도구

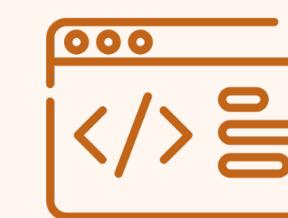


분석 설명

- 기간 : 2021.08.01. ~ 2021.09.02.
- 크롤링한 리뷰 : 12,920개 (40개 호텔)
- 분석에 사용한 리뷰 : 6,200개 (31개 호텔)

03.

프로젝트 과정



0	1	0	0
0	0	0	0
1	0	1	1

데이터 선택

데이터 분석

데이터 모델링

데이터 벡터화

호텔 필터링

형태소 분석

가설 설정

웹 크롤링

키워드 추출

데이터 라벨링

TF-IDF

데이터 전처리

유용한 리뷰 필터링

감성분석

03 (1) 데이터 선택

1. 웹 크롤링

- 데이터 수 : 총 40개 호텔, 리뷰 12,920개
- 기간 : 2019.06. ~ 2021.08.07. (약 2년)
- 기준 : 특수문자 제외, 띄어쓰기 포함 15글자 이상 500글자 이하

hotel_code	hotel_name	hotel_location
3009313	아트스테이 서귀포 하버	서귀포 이중섭문화거리•474.9km
3008436	휘슬락 호텔	제주국제공항 차량 10분•445.3km
3010282	호텔위드 제주	제주국제공항 차량 10분•449.2km
3001285	봄그리고가을 호텔&리조트	성산일출봉 3km•447.6km
3009057	라마다 제주시티 호텔	제주국제공항 4.86km•447.8km



라마다 제주시티 호텔

★ 4.5 /5 후기 2,006개 >

최근 작성순 ▼



바른후기

레귤 | [특가야놀자☆] 스… - 숙박

2021. 05. 31

편안하게 품 쉬다갑니다!! 주차가 아쉽지만 저렴하고 좋았어

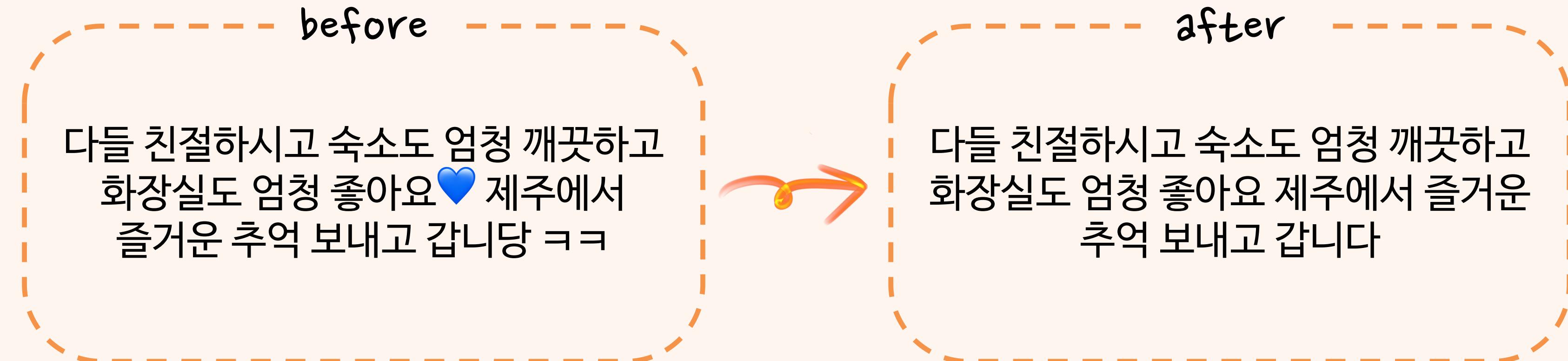
리뷰 개수가 200개 이상인 호텔을 기준으로 필터링 후,
html 코드를 이용하여 호텔의 정보(url, 이름, 위치) 수집

Selenium을 이용하여 필터링 적용 후,
호텔의 리뷰(호텔명, 리뷰, 별점, 날짜) 수집 자동화

03 (1) 데이터 선택

2. 데이터 전처리

1. Sampling : 크롤링 후, 호텔별(31개) 리뷰 200개 랜덤 추출 → 분석에 사용할 data : $31 \times 200 = 6,200$ 개
2. 자음, 모음 제외 (ㄱ ㄴ ㄷ ㄹ ...)
3. 이모티콘 제외 (😊❤️😊❤️ ...)
4. 맞춤법 검사 (네이버 맞춤법 검사 라이브러리 hanspell 사용)
5. 형태소 분석 시 발생하는 문제를 위한 추가적인 전처리 : 가성비, 뷰 → 구성비, 경치



03 (2) 데이터 분석

1. 형태소 분석

데이터 토큰화

텍스트 데이터를 하나의 특정 기본 단위로 자르는 것

숙소도 엄청 깨끗하고 화장실도 엄청 좋아요

KoNLPy의 Mecab 형태소 분석기 사용

('숙소', 'NNG'), ('도', 'JX'), ('엄청', 'MAG'), ('깨끗', 'XR'),
('하', 'XSA'), ('고', 'EC'), ('화장실', 'NNG'), ('도', 'JX'),
('엄청', 'MAG'), ('좋', 'VA'), ('아요', 'EF')

일반명사(NNG), 형용사(VA), 어근(XR), 일반부사(MAG)

03 (2) 데이터 분석

1. 형태소 분석

1. 전처리

- 불용어 제거 ('도', '하', '고' ...)
- 한 글자 단어 제거
- 형태소 중 일반명사(NNG), 형용사(VA), 어근(XR), 일반부사(MAG) 추출

2. 호텔별 다빈도 단어 50개 추출

3. 다빈도 단어 시각화 : Bubble Chart

제주 센트럴 시티 호텔



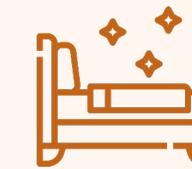
03 (2) 데이터 분석

2. 키워드 추출

- 리뷰 기반 7개의 상위 키워드 선정



청결



시설



서비스



접근성



가격



위치



방음

3. 유용한 리뷰 필터링

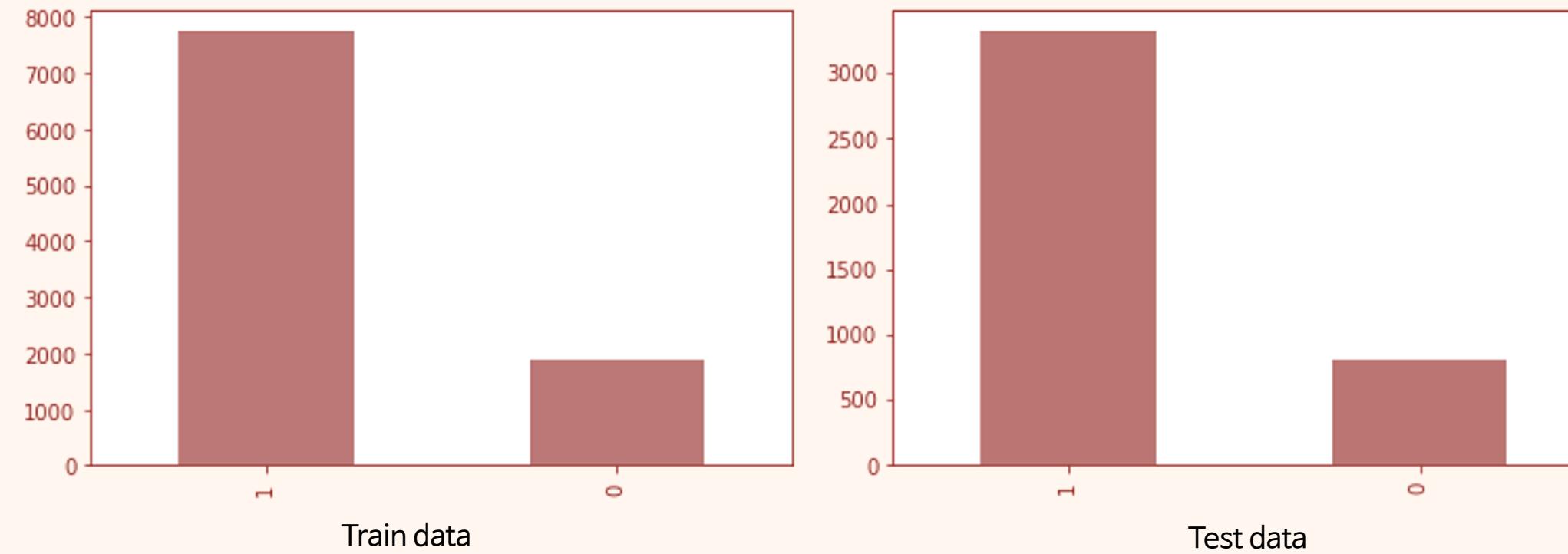
- 키워드 여부와 리뷰 길이에 따라 리뷰에 유용한 정보가 있다고 판단하여 최종 추천된 호텔의 유용한 리뷰를 제공하여 호텔을 파악하는데 도움
- 필터링 조건
 - 1) 하위 키워드 2개 이상
 - 2) 하위 키워드 1개 + 리뷰 길이 42자(중간값) 이상

Review	helpful
방이 좁았지만 위치가 좋았어요	0
주차가 생각보다 어려워요 조식은 저렴한 가격에 괜찮았습니다 프런트에 계시던 직원분들은 고객 응대에 조금 미숙한 거 같았습니다	1
냄새가 너무 나서 머리가 아팠어요	0
생각보다 너무 좋았습니다 깔끔하고 방도 업그레이드해주셔서 좋았어요 다음에는 날짜 맞춰서 루프탑 이용도 해보고 싶습니다	1

03 (3) 데이터 모델링

1. 가설 설정

- 리뷰 별점 4점 이상 = 1 (긍정) , 3점 이하 = 0 (부정)으로 라벨링 시 데이터 불균형 발생 → 다른 라벨링 기준 모색



약 13,000개 데이터를 직접 읽어 본 결과, 별점 1점과 5점의 리뷰는 각각 부정 • 긍정이 대체로 확실하나
별점 2~4점의 리뷰에는 긍정과 부정 내용이 혼재되어 있음

따라서, 별점 2~4점이 실제 리뷰 내용을 정확하게 반영하지 못하므로 실제 리뷰 기반의 긍 • 부정 라벨링 재진행

03 (3) 데이터 모델링

2. 데이터 라벨링

- 1) Mecab을 이용해 종결 어미, 일반 부사, 합성어(동사+어미)를 기준으로 토큰화
- 2) 라벨링 진행
 - 별점 1, 5점 리뷰 (총 11,564 건) : 별점 1점= 0 (부정), 별점 5점 = 1 (긍정)
 - 별점 2~4점 리뷰 (총 6,033 건) : 지도학습을 위해 직접 라벨링 (1 = 긍정, 0 = 부정, 2 = 의미없는 내용)

3. 감성 분석

LSTM

Long Short-Term Memory models

: RNN 모델 중 하나로 장/단기 기억을 가능하게 설계한 신경망의 구조로, 주로 자연어 처리에 사용

- 감성 분석에 사용한 데이터 (총 18,365 건) : 직접 라벨링한 별점 2~4점 리뷰 6,950건 + 별점 1, 5점 리뷰 11,415건
- 별점 기반 감성 분석 정확도 약 81% → 실제 리뷰 기반 감성 분석 정확도 약 91%

star	review	감정
1	모텔보다도 못한 방음 그리고 너무 더러웠어요	부정
5	성산일출봉 보며 물놀이 짱! 가성비가 너무 좋아서 다음에 또 갈 것 같아요	긍정
3	후기처럼 샤워커튼 곰팡이 때문에 찢찝했네요 / 그래도 수압은 최고네요 침구도 괜찮아요	부정 / 긍정

03 (4) 데이터 벡터화

1. TF-IDF 벡터화

TF-IDF

문서 빈도 (Document Frequency) + 역문서 빈도 (Inverse Document Frequency)
: 어떤 단어가 흔하지 않으면서도 특정 텍스트에서는 자주 사용된 정도를 나타낸 지표



긍정인 리뷰 데이터를 바탕으로 호텔별 리뷰를 하나의 문서(총 31개)로 합친 후,
호텔별 7개 키워드의 TF-IDF 지표 확인

호텔명	청결	서비스	가격	위치	시설	접근성	방음
더 베스트 제주 성산	0.362461	0.304163	0.111526	0.380204	0.382739	0.159686	0.058298
더 포 그레이스 리조트	0.212888	0.266110	0.130636	0.198373	0.641083	0.193535	0.048384
더큐브 리조트 제주	0.343227	0.301199	0.063042	0.336222	0.340892	0.189125	0.086390

04.

최종 시스템 알고리즘

7개의 키워드 중
1~3순위 선택



선택된 키워드에 따라
가중치 조정 후
추천 호텔 3개 선정

STEP 2

3개의 추천 호텔에 대한 시각화
(Map, Radar, WordCloud)
+
유용한 리뷰 제공

STEP 3

05.

추천 시스템 활용 예시

STEP 1



제주 여행이 처음인

20대 사용자

청결

서비스

가격

위치

시설

접근성

방음

7개의 키워드 중
1~3순위 선택

STEP 2

가중치 조정된
TF-IDF

라마다 제주시티 호텔	1.597985
원스토리 호텔 서귀포	1.515364
호텔 에어시티 제주	1.452326

1순위 - 청결
2순위 - 가격
3순위 - 방음

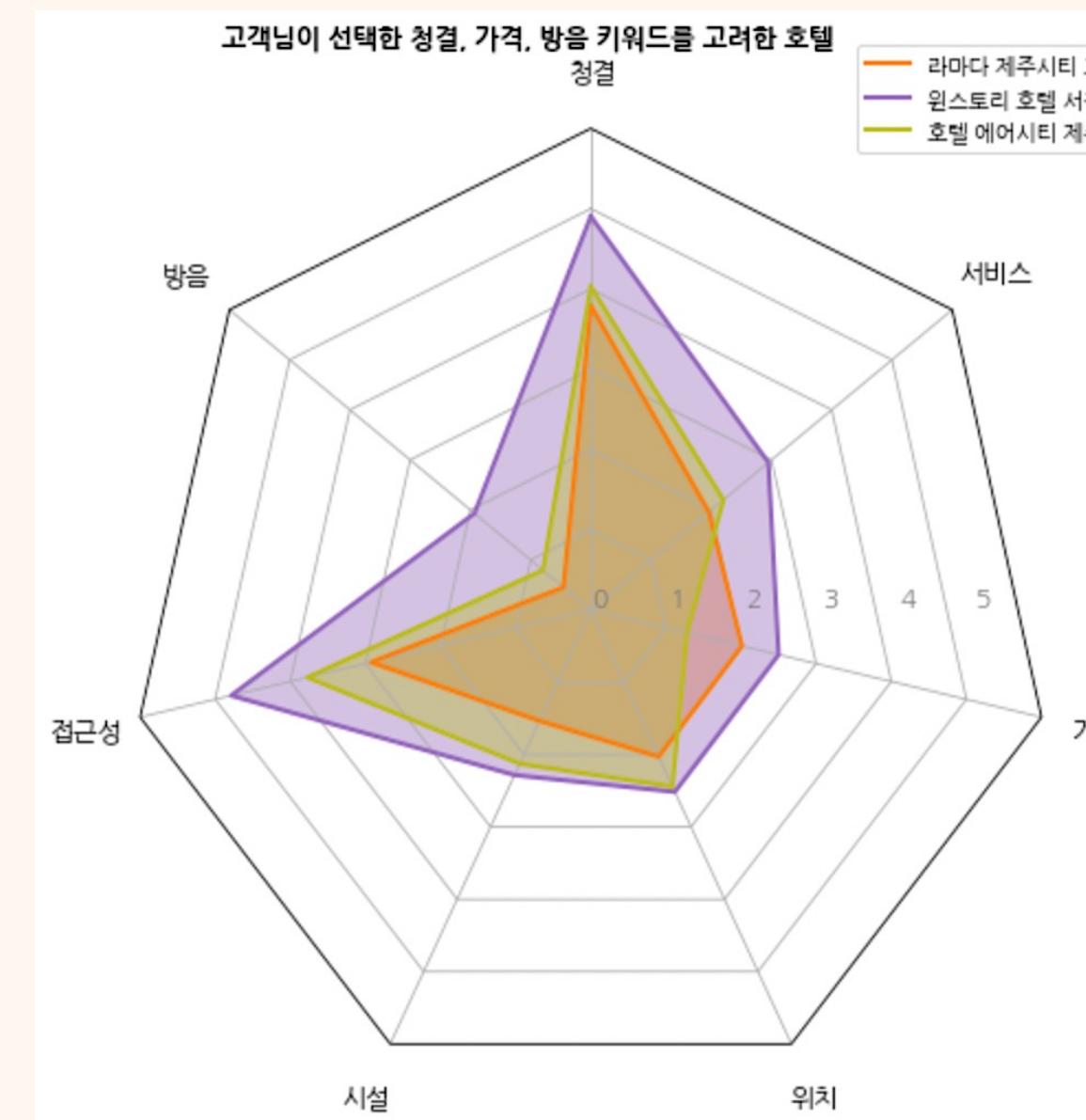
가중치 조정한 TF-IDF 기준 상위 3개 호텔 선정

05.

추천 시스템 활용 예시

STEP 3

"고객님에게 추천하는 호텔은 1순위 라마다 제주시티 호텔,
2순위 원스토리 호텔 서귀포, 3순위 호텔 에어시티 제주입니다."



05.

추천 시스템 활용 예시

STEP 3

라마다 제주시티 호텔 – Helpful Review

성인 4인가족이 머물 디럭스 침대 2개인 방을 찾기가 힘든데 운좋게 발견했어요. 공항에서 가깝고 근처에 식당도 가깝고 좋았어요.

주차장이 협소하지만 그래도 미리 전화하면 시간맞춰서 준비 도와주셨어요~ 가성비대비 좋아요.

밤늦게 도착했지만 공항에서 가까워서 좋았고 발렛파킹이 참 편했어요! 밑에 편의점도 크고 깔끔해서 좋았습니다.



라마다 제주시티 호텔에 대한 긍정 키워드의 WordCloud



라마다 제주시티 호텔에 대한 부정 키워드 WordCloud

분석의의

1. 자신이 원하는 키워드에 따라 개인화된 호텔 추천으로 제주도 여행 시 만족도 향상
2. 리뷰를 모두 읽어 호텔을 파악하거나 별점으로 호텔을 정하기 어려웠던 문제를 실제 사용자의 리뷰 분석으로 정확한 만족도 파악
3. 고객이 선택한 키워드에 따라 상위 호텔 3개의 시각화된 정보를 바로 출력하여 호텔 찾는 시간 감소

기대효과

1. 학습시킨 감성분석 모델로 새로운 리뷰가 추가됐을 때 자동화하여 실제 리뷰 기반으로 리뷰 감성분석 가능
2. 국내 유명 관광지역 숙박시설에 분석 알고리즘을 적용하여 국내여행 만족도 향상 및 관광 산업 활성화에 기여
3. 숙박시설 외 고객별 관심분야(예 : 맛집, 문화공연 등)에 적용하여 객관적인 양질의 정보 제공

07.

역할 분담



유경민

호텔 리뷰 크롤링 초안
형태소 분석 초안
train을 위한 데이터 긍/부정 라벨링
LSTM 모델 초안
데이터 시각화 초안(Wordcloud)
TF-IDF 벡터화 초안



이중호

train을 위한 데이터 긍/부정 라벨링
데이터 시각화 초안
(Bubble Chart, Folium)



지다영

프로젝트 계획 및 진행 관리
호텔 정보 수집 자동화
웹 크롤링
데이터 전처리
형태소 분석
데이터 시각화
(WordCloud, Bubble Chart)
형태소 분석 결과에 따른 키워드 추출
데이터 라벨링 (유용한 리뷰)
LSTM 모델링(data 구성)
TF-IDF 벡터화
최종 시스템 알고리즘 구현
문서화



정새롬

데이터 전처리
형태소 분석 + 명사 추출
데이터 시각화
(WordCloud, Folium, Radar Chart)
train을 위한 데이터 긍/부정 라벨링
LSTM 모델링 및 튜닝
가설 재설정 및 프로젝트 방향 재조정
최종 시스템 알고리즘 구현

THANK YOU

