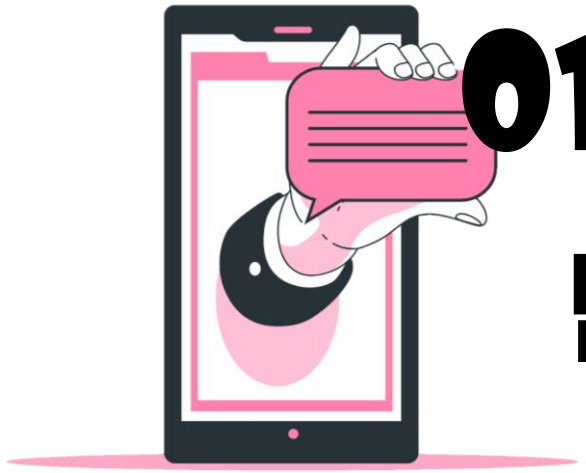
리뷰분석을 통한 개인 맞춤형 숙소 추천 서비스

검색엔진 & 추천시스템

방과방가

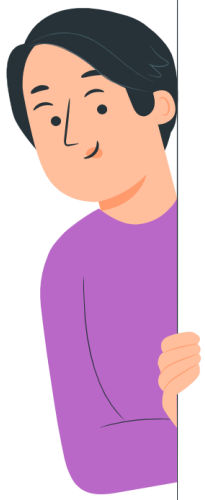
고준호 이선영
김다나 이지수
김경연

목차



01

팀 소개



방과방가

Room&Hi

: 내가 마음에 드는 방과 만날 수 있다.

팀소개

이선영

데이터 수집
데이터 전처리
프론트엔드 및 백엔드
추천 시스템

고준호(조장)

데이터 수집
데이터 전처리
프론트엔드 및 백엔드
추천 시스템

김다나

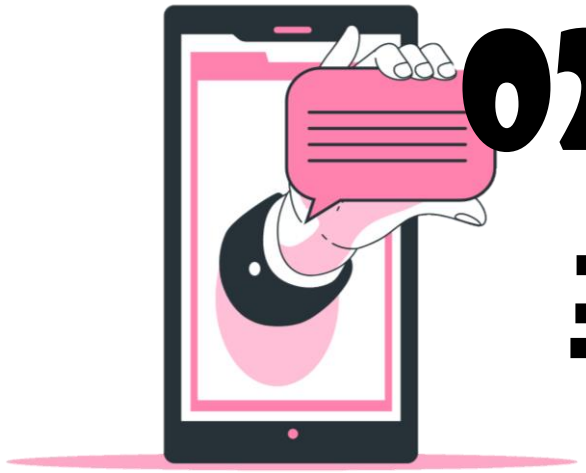
데이터 수집
데이터 전처리
프론트엔드 및 백엔드
엘라스틱서치

이지수

데이터 수집
데이터 전처리
프론트엔드 및 백엔드
엘라스틱서치

김경연

AWS

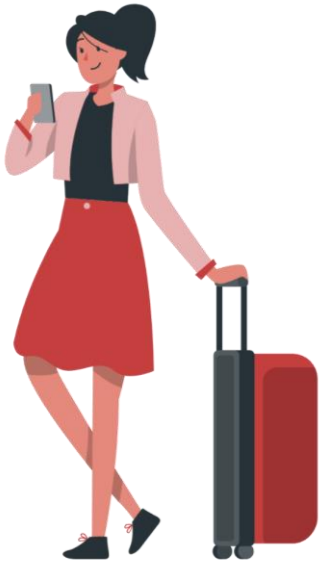


02

프로젝트 소개

배경

숙소 정보, 리뷰

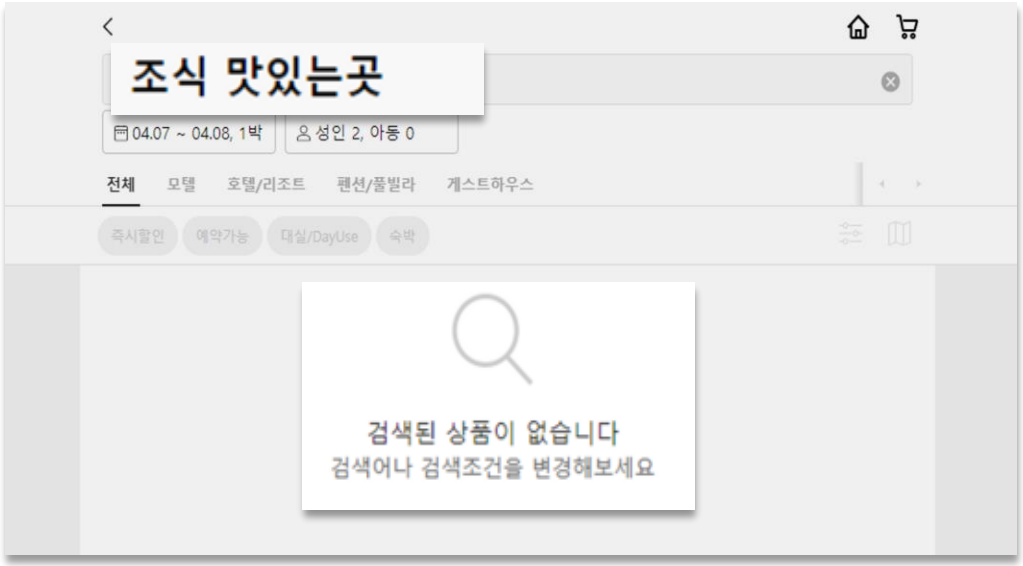


시맨틱 서치

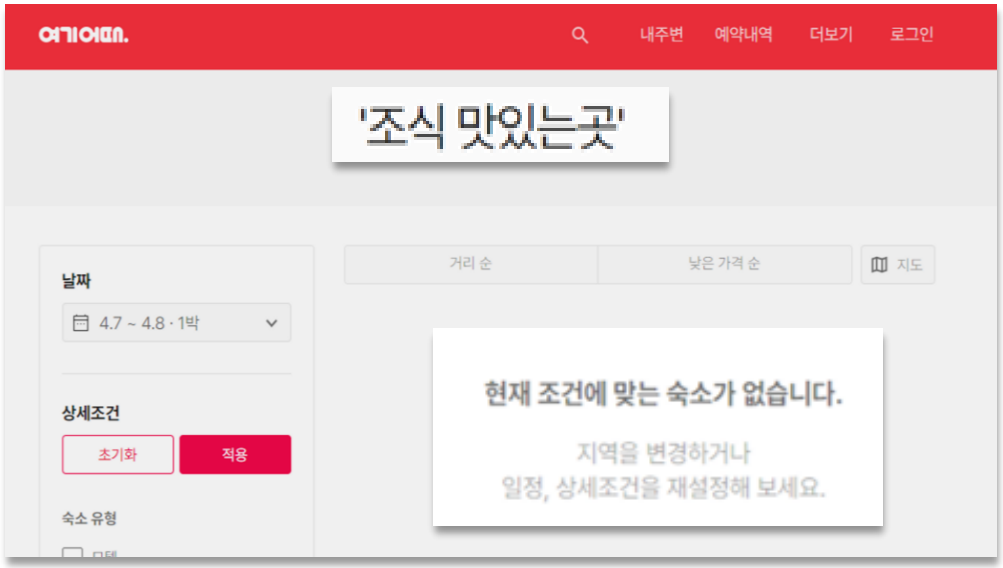
시맨틱(Semantic)서치

사용자의 검색의도를 파악하고, 문서에 기술된 어휘의 의미와 문맥을 분석하여, 사용자가 원하는 검색 결과를 제시하는 것

배경



국내 숙박 플랫폼 1위 y nolja



국내 숙박 플랫폼 2위 여기어때.

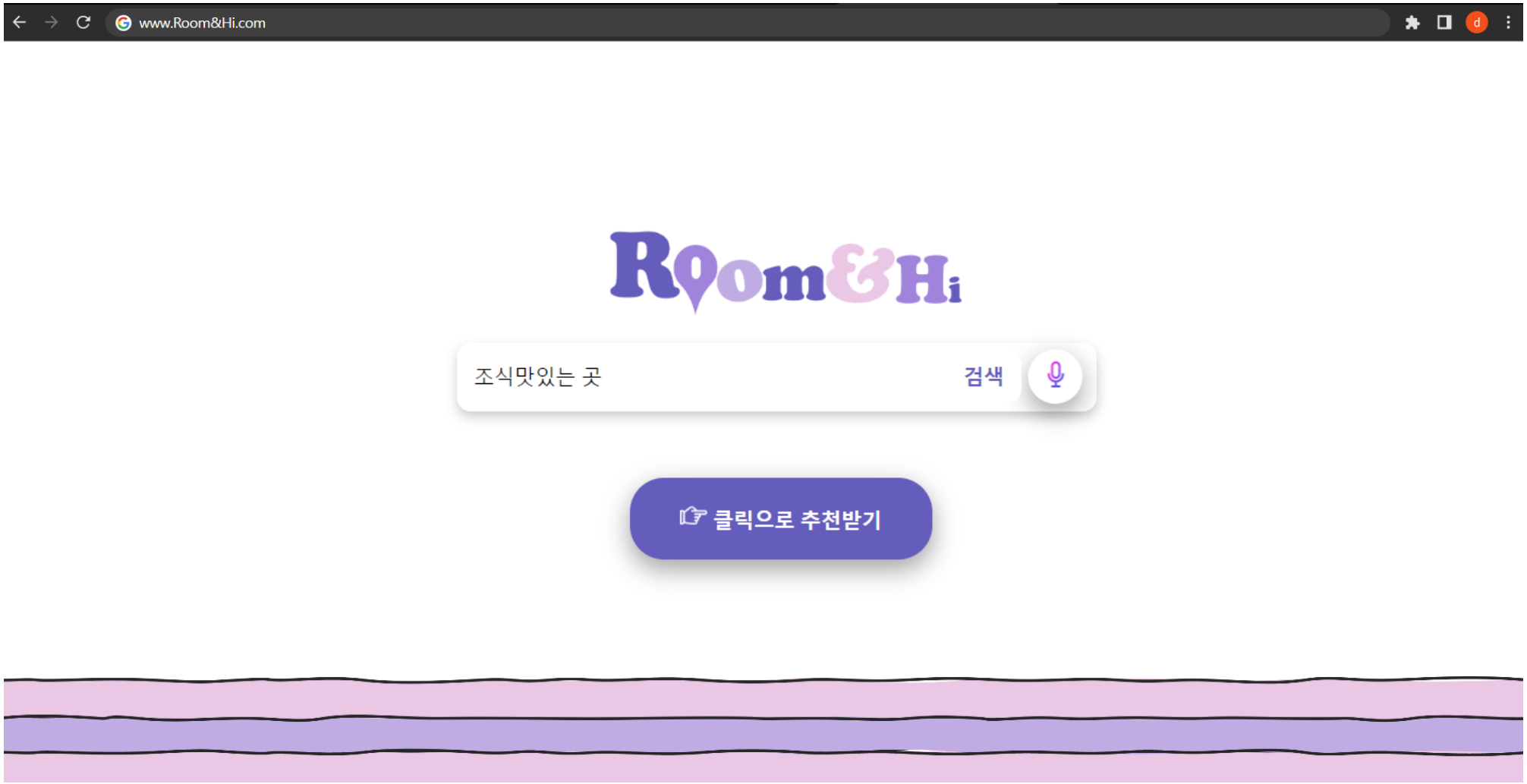
배경

국내 숙박 플랫폼 1위 y

내 숙박 플랫폼 2위 여기어때.

조식 제공 여부 확인 ⇨ 조식 관련 리뷰 확인하기

배경

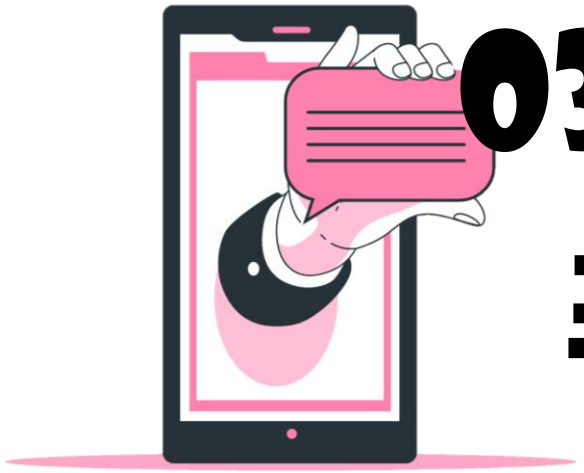


최종 목표

1. 소비자의 검색피로도 최소화
2. 맞춤형 숙소 추천 서비스

Project Timeline

Tasks	Feburary	March	April
1. 프로젝트 기획 및 사전 조사	2.15 - 2.21		
2. 프로젝트 설계	2.20 - 2.21		
3. 데이터 수집 및 전처리	2.21 - 3.7		
4. 추천 시스템 구현	2.28 - 3.8		
5. 웹 구현 및 AWS	3.6 - 3.16		
6. TEST 1차 - 보완		3.16 - 3. 22	
7. TEST 2차 - 보완		3.23 - 3. 31	
8. 최종 TEST 및 배포			4.1 - 4.6



03

프로젝트 설계

시스템 아키텍처

Input

검색창 또는 음성인식 사용

"영등포" 에서 "무료조식에다가 가성비 좋은 호텔 알려줘"

stt webkitSpeechRecognition

사용자

카테고리 추천 선택

EC2 instance contents

Flask
web development, one drop at a time

Model

지역 필터링 + 문장 벡터화 (SBERT)

ElasticSearch
Review data DB

지역 필터링 + 선택지에 따른 가중치 계산

Score.csv
카테고리별 숙소 점수화

피어슨 상관계수 유사 호텔 추천

시스템 아키텍처

Output

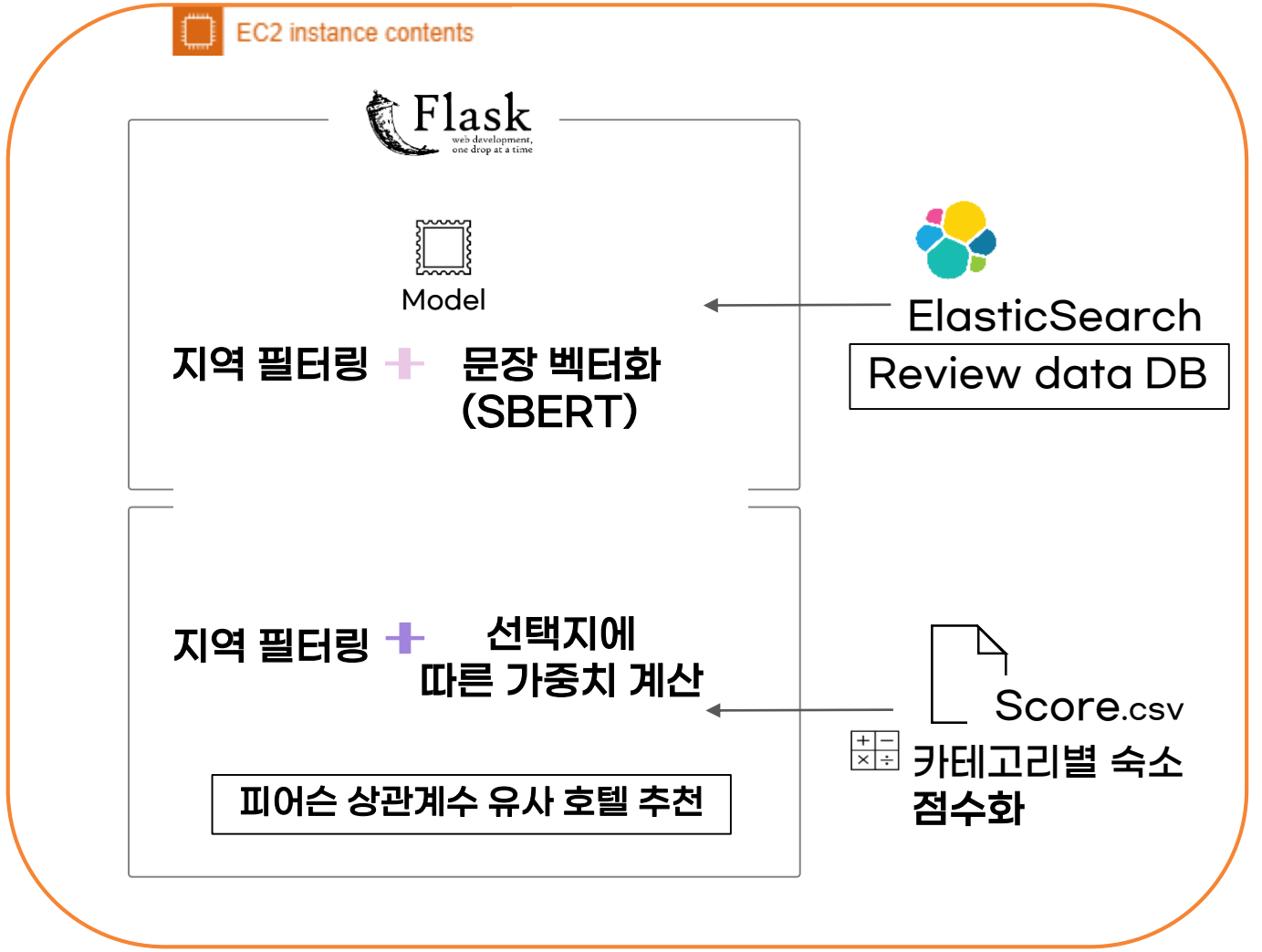
유사리뷰 기준 호텔 결과 도출

score	gu	review
450.68558	영등포구	깨끗하고 구성비 좋은 호텔 무료로 제공되는 아침식사도 알차고 맛있었어요

사용자

추천 결과 도출

No.	호텔	등급	보러가기	더보기
1	토요코인서울영등포	3	B. × ×	더 보기
2	영등포GMS호텔	3	× ya	더 보기
3	더스테이트선유	3	B. ya	더 보기
4	호텔부티크9	3	× ya ×	더 보기
5	영등포부띠크HotelSB	3	B. ×	더 보기



엘라스틱서치 사용이유

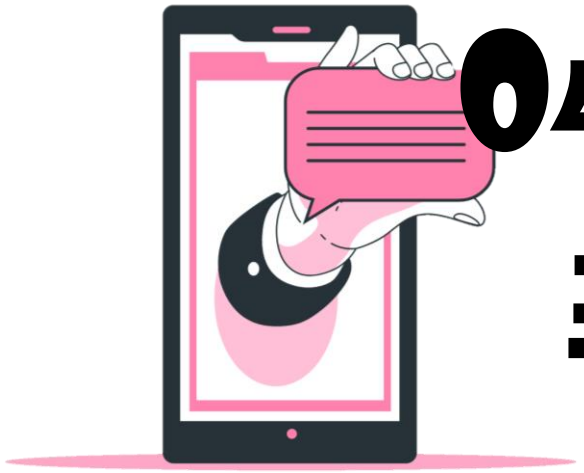
Vs

정형 데이터 처리에는 효율적

비정형 데이터인 리뷰를 처리하는데 있어서
속도, 유연성측면에서 효율적인 검색이

엘라스틱서치란?

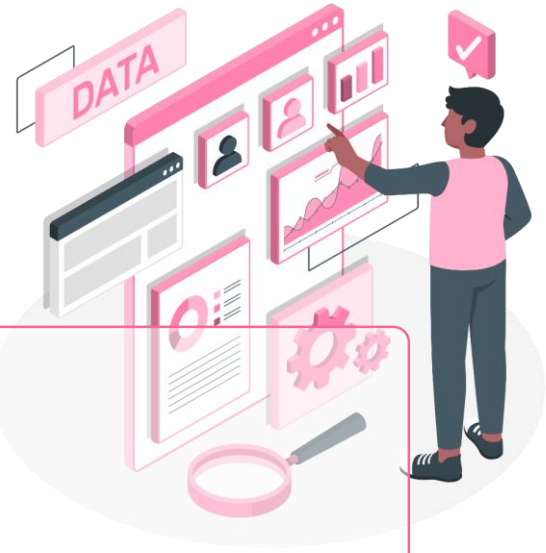
Apache Lucene기반의 Java 오픈소스 분산 검색 엔진으로 방대한 양의 데이터를
신속, 실시간으로 저장, 검색, 분석할 수 있는 검색엔진



04

프로젝트 진행

데이터



데이터 수집

서울 386개 숙박 시설

yanolja 310개

여기어때 257개

Booking.com 157개

- 수집 데이터 -
날짜 . 리뷰 . 별점
총 397,152개

데이터 전처리

1. 특수문자, 개행문자 필터링
2. 날짜 형식 맞추기
   : '16분전', '4일 전' , '3개월 전' → yyyy-mm-dd
3. 불용어 제거
   :한국어 불용어 사전 + 서울,호텔,리뷰 …
4. 문장 분리 (KSS)
5. 맞춤법 검사 (Hanspell)

리뷰 감성분류

니스모래시계04 | 2022.10.15

객실명 이그제큐티브 스카이뷰 킹

체크인 40분 이상 대기했음....

토닥토닥꼬리풀꽃0096 | 2022.10.07

객실명 이그제큐티브 스카이뷰 킹

너무 좋아요 동대문 뷰랑 근처 산책도요

!? 평점과 텍스트 리뷰 간 괴리 존재

-> 평점 기준으로 리뷰를 긍부정으로 구분하는 것은 정확도가 떨어짐

17000개의 학습데이터 직접 라벨링 진행

→ 성능 개선을 위해 데이터 비율 조정 (2:2:1) , 14000개 학습

* 평점과 리뷰간의 괴리 발생으로 인한 분류 정확성 저하 방지

리뷰 분류 기준

0	부정
1	긍정
2	긍부정 복합, 의미 없음

감성분류 모델

5개의 사전학습 모델 테스트 결과

모델 명	Accuracy
KoBERT	0.9122
KLUE-BERT-base	0.8912
KLUE-RoBERTa-large	0.9344
kykim/funnel-kor-base	0.9404
KoELECTRA-Base-v3	0.9443

	precision	recall	f1-score	support
0	0.95	0.96	0.96	1043
1	0.96	0.96	0.96	1062
2	0.90	0.88	0.89	624
accuracy			0.94	2729
macro avg	0.94	0.94	0.94	2729
weighted avg	0.94	0.94	0.94	2729

Classification report

라벨링 결과

ht_id	date	review	label
188	2020.02.27	미니바는 술 제외한 음료는 다 무료였고 어메니티 보디로션 향이 너무 좋네요	1
353	2021.11.21	위치 좋고 직원의 친절함은 정말 환상이었습니다	1
154	2020.02.25	침대가 4개 있었는데 위치 선정이 너무 좋았고 저렴한 가격 대비 너무 좋은 공간이었어요	1
302	2023.02.26	접근성이 좋고 주변 편의시설이 많아서 좋아요	1
225	2021.05.27	직원들이 상냥합니다	1

ht_id	date	review	label
190	2022.02.23	청소 및 물품 비치가 살짝 아쉬웠다	0
111	2020.06.04	실내 수영장까지의 이동경로가 불편	0
188	2022.01.15	샤워실 면 한쪽이 유리 문이라 좀 부담스러웠음	0
80	2022.12.27	천장 히터만 틀어지고 창문 쪽에선 찬바람이 너무 많이 들어와서 놀러 와서 감기 걸렸어요	0
246	2022.09.27	다만 저한테는 샤워기 키가 크신 분들한테는 불편하실 수 있습니다	0

ht_id	date	review	label
363	2022.02.27	다만 강남이라 주차는 이해해야 하죠	2
180	2022.08.27	중문 없는 게 이렇게 큰가 싶을 정도로 복도 소음도 크지만 거 빼면 다 좋아요	2
26	2023.01.03	역에서도 가깝고 조명이 조금 어둡긴 하지만 이용하는데 불편함 없어요	2
126	2022.03.01	시설이 오래되었지만 직원 서비스로 다 커버가 되네요	2
237	2022.07.04	위치 접근성 최곤데 침대 뜯어진 자국 있고 그런 것만 빼면 좋았어요	2

감성분류 모델

5개의 사전학습 모델 테스트 결과

모델 명	Accuracy
KoBERT	0.9122
KLUE-BERT-base	0.8912
KLUE-RoBERTa-large	0.9344
kykim/funnel-kor-base	0.9404
KoELECTRA-Base-v3	0.9443

	precision	recall	f1-score	support
0	0.95	0.96	0.96	1043
1	0.96	0.96	0.96	1062
2	0.90	0.88	0.89	624
accuracy			0.94	2729
macro avg	0.94	0.94	0.94	2729
weighted avg	0.94	0.94	0.94	2729

Classification report

라벨링 결과

ht_id	date	review	label
188	2020.02.27	미니바는 술 제외한 음료는 다 무료였고 어메니티 보디로션 향이 너무 좋네요	1
353	2021.11.21	위치 좋고 직원의 친절함은 정말 환상이었습니다	1
154	2020.02.25	침대가 4개 있었는데 위치 선정이 너무 좋았고 저렴한 가격 대비 너무 좋은 공간이었어요	1
302	2023.02.26	접근성이 좋고 주변 편의시설이 많아서 좋아요	1
225	2021.05.27	직원들이 상냥합니다	1

ht_id	date	review	label
190	2022.02.23	청소 및 물품 비치가 살짝 아쉬웠다	0
111	2020.06.04	실내 수영장까지의 이동경로가 불편	0
188	2022.01.15	샤워실 면 한쪽이 유리 문이라 좀 부담스러웠음	0
80	2022.12.27	천장 히터만 틀어지고 창문 쪽에선 찬바람이 너무 많이 들어와서 놀러 와서 감기 걸렸어요	0
246	2022.09.27	다만 저한테는 샤워기 키가 크신 분들한테는 불편하실 수 있습니다	0

긍정리뷰(라벨1)과 부정리뷰(라벨0) 사용

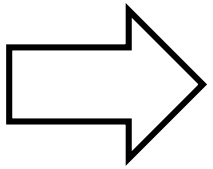
총 사용 리뷰 수 : 382,441개

카테고리 사전 생성

최지유·박소현·김상호·이규범·곽민정(2018).
“온라인 리뷰 마이닝을 통한 국내호텔 서비스품질 주요속성 분석”.『관광경영연구』.

“솔직히 기대 안했는데 조식 짱 맛있었어요”

“아침식사 먹으러 여기 와요 ”



카테고리
“조식”

MeCab 명사 빈도수 추출 → 총 36개의 카테고리

카테고리 사전 생성

최지유·박소현·김상호·이규범·곽민정(2018).
“온라인 리뷰 마이닝을 통한 국내호텔 서비스품질 주요속성 분석”.『관광경영연구』.

카테고리	키워드(하위단어)
가격	구성비, 가성비,가격
직원	친절, 직원, 응대, 대응, 사장
프론트	프런트, 체크인, 체크아웃
통신	인터넷, 와이파이
분위기	분위기, 조명
시설/인테리어	인테리어, 디자인, 외관, 엘리베이터, 에레, 신추, 건물 , 시설
로비	로비, 입구
라운지	클럽라운지, 옥상, 루프탑
주차장	주차장, 주차, 발레, 차량
산책	산책, 공원
부대시설	편의시설, 부대시설
수영장/사우나/스파	수영장, 스파, 사우나, 수영

카테고리	키워드(하위단어)
휘트니스	피트니스
비즈니스	비즈니스, 출장
위치	위치, 접근성, 거리, 이동, 도보, 도심, 식당
교통	지하철, 자히철역, 공항, 버스, 대중교통, 교통
관광	쇼핑, 백화점, 관광, 관광지, 볼거리
가구	소파, 책상, 가구, 옷장, 탁자,전등
침대/가구	침대, 침구류, 베드, 트윈, 매트리스, 푹신, 시트, 침구, 침실, 커버
커튼/카페트	카펫, 블라인트, 커튼 , 카페트
가전/전자제품	에어컨, 티브이, 히터, 컴퓨터, 가습기, 청정기, 정수기, 전자레인지, 세탁기, 스피커, 가전, 전자제품
화장실	욕조, 욕실, 비데, 배수구, 목욕, 반신욕, 샤워, 화장질
비품	어메니티, 수건, 칫솔, 용품, 타올, 일회용품, 물품, 비누,린스, 휴지, 면도기, 제품, 세면도구, 비품
물	수압, 온도, 온수, 물

카테고리	키워드(하위단어)
객실	객실, 공간, 천장, 테라스, 조리, 취사
방음	소리, 방음, 옆방
온도	냉방, 중앙난방, 보일러, 외풍
바닥/벽	바닥, 벽지, 벽
창문	창문
룸컨디션	컨디션, 노후, 연식, 낙후, 모기, 룸컨디션
냄새	냄새, 담배, 하수구
전망	경치, 풍경, 야경, 뷰, 전망
편의용품	충전기, 콘센트, 슬리퍼, 생수, 옷걸이 ,편의용품, 드라이기
청결	청결, 먼지, 머리카락, 얼룩, 곰팡이, 정리, 자국, 물때, 쓰레기, 벌레,청소 , 깔끔, 깨끗
룸서비스	룸서비스
조식	조식, 아침식사, 뷔페

추천 기준 - 만족지수

$$\text{만족지수} = \frac{\text{긍정비율}}{\text{긍정비율} + \text{부정비율}} \times 100$$

$$\text{긍정비율} = \frac{\text{해당 카테고리 긍정 리뷰 수}}{\text{해당 카테고리 언급 리뷰 수}} \times 100$$

*곽민정, 최지유, 박소현(2019)
“호텔 서비스 속성별 고객만족도 분석을 위한 온라인 리뷰 감성분석”

Hotel SeSAC의 **조식 만족지수** 구하기

$$\text{조식 긍정비율} = \frac{\text{‘조식’ 긍정(라벨1) 리뷰 수}}{\text{‘조식’ 언급 리뷰 수}} \times 100$$

$$\text{조식 부정비율} = \frac{\text{‘조식’ 부정(라벨0) 리뷰 수}}{\text{‘조식’ 언급 리뷰 수}} \times 100$$

$$\text{조식 만족지수} = \frac{\text{‘조식’ 긍정비율}}{\text{‘조식’ 긍정비율} + \text{‘조식’ 부정비율}} \times 100$$

추천 기준 - 만족지수

	ht_id	1	2	3	4	5	6	7	8	9	...
0	1	96.55	93.75	71.43	0.0	100	63.64	0.0	0.0	16.67	...
1	2	90.38	90.14	100	0.0	66.67	68.0	0.0	100	0.0	...
2	3	25.0	88.89	66.67	0.0	100	70.0	0.0	0.0	33.33	...
3	4	94.23	97.0	88.89	0.0	66.67	79.41	100	0.0	57.14	...
4	5	92.0	92.75	76.92	0.0	100	83.33	33.33	0.0	28.57	...
...	...	...	...	...	...	...	...	...	...	...	...

만족지수 = 긍정

긍정비율 = 해당 / 해당

구하기

00

00

x 100

조식 긍정비율 + 조식 부정비율

*곽민정, 최지유, 박소현(2019)
"호텔 서비스 속성별 고객만족도 분석을 위한 온라인 리뷰 감성분석"

엘라스틱 서치

리뷰데이터가 들어갈 필드와 타입 Mapping

```
es.indices.create(
    index='re_list',
    body={
        "mappings": {

                "properties": {
                    "ht_id": {
                        "type": "keyword"
                    },
                    "date": {
                        "type": "date",
                        "format": "yyyy.mm.dd"
                    },
                    "review": {
                        "type": "text"
                    },
                    "label": {
                        "type": "integer"
                    }
                }
            }
    }
)
```

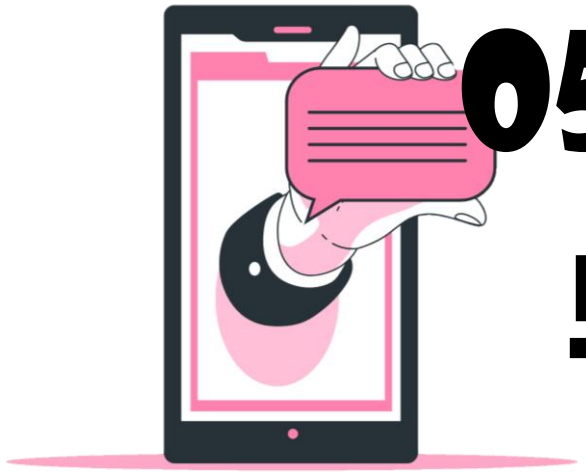
- "type": "keyword" → 고유값으로 검색 가능
- "type": "date" → 날짜 범위 지정하여 검색 가능
- "type": "text" → 일부단어, 유사도 검색 가능
- "type": "integer" → 숫자값으로 검색 가능

TF-IDF 유사도 검색 가능

아기랑 놀기 좋은 곳 [검색]

유사도 점수 / **"좋은 곳" 단어 일치**

```
12.538096 / 혼자 쓰기 딱 좋은 곳이에요
12.470367 / 위치도 좋고 무엇보다 아기와 함께하기 좋았어요
12.202542 / 위치도 매우 찾기 좋은 곳입니다
12.128616 / 새로 오픈한 곳이라 깔끔 간단한 취사가 가능 히노끼탕에서 놀기 좋음
12.0970955 / 광장시장 최고 낮에는 아기랑 수영장에서 즐거운 시간 보냈네요
11.885953 / 호캉스하기 딱 좋은 곳입니다
11.885953 / 깔끔하게  하루 머물기 좋은 곳입니다
11.8196335 / 가까운 곳에 카페랑 서브웨이 있어서 너무 좋았어요
11.814242 / 5명 호텔 숙박 되는 곳 찾기 쉽지 않은데 잘 놀고 갑니다
11.596999 / 대학로 근처라 놀기도 좋고요
```



05

보완

문제 제기

엘라스틱서치 검색 정확도 문제

TF-IDF 검색에서는 중요 단어, 키워드가 정확하게 추출 되지 않음

아기랑 놀기 좋은 곳 검색

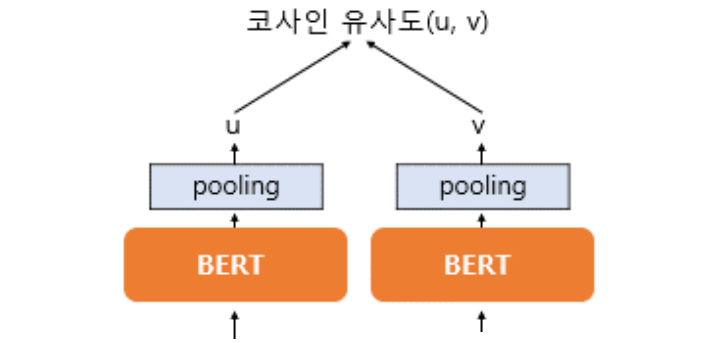
12.538096 / 혼자 쓰기 딱 좋은 곳이에요
12.470367 / 위치도 좋고 무엇보다 아기와 함께하기 좋았어요
12.202542 / 위치도 매우 찾기 좋은 곳입니다
12.128616 / 새로 오픈한 곳이라 깔끔 간단한 취사가 가능 히노끼탕에서 놀기 좋음
12.0970955 / 광장시장 최고 낮에는 아기랑 수영장에서 즐거운 시간 보냈네요
11.885953 / 호캉스하기 딱 좋은 곳입니다
11.885953 / 깔끔하게 하루 머물기 좋은 곳입니다
11.8196335 / 가까운 곳에 카페랑 서브웨이 있어서 너무 좋았어요
11.814242 / 5명 호텔 숙박 되는 곳 찾기 쉽지 않은데 잘 놀고 갑니다
11.596999 / 대학로 근처라 놀기도 좋고요

문제 해결

SBERT(Sentence BERT) 모델을 통해 문맥에 따라 유사도를 검색하는 문장임베딩 수행

➡ **Fast Elasticsearch Vector Scoring plugin 사용**

SBERT(Sentence BERT)란?



두 문장에서 문장벡터의 평균을 통해 유사한 문장임을 확인

아기랑 놀기 좋은 곳 | 검색

1.767511 / 저녁에 산책도 좋았고 아기랑 놀기 괜찮았어요

1.7231867 / 아기들 놀 수 있는 수영장 이 하나 있습니다

1.7157646 / 아기랑 같이 산책했는데 좋아하더라고요

1.7097986 / 위치도 좋고 무엇보다 아기와 함께하기 좋았어요

1.7081176 / 위치도 찾기 좋고 거리도 적당하고 애들 놀기에도 좋고 괜찮았어요

1.7040801 / 위치가 일단 아주 좋고 근처 놀 수 있는 곳들이 많다

1.6928089 / 아기 반려견과 같이 갈 수 있는 점도 좋았고 식사 수영장까지 포함돼 있어 좋았습니다

1.6909626 / 놀이공원이 같이 있어 좋아요

1.6898826 / 욕조가 크고 넓어서 아기가 놀기 좋았어요

문제 제기

리뷰 데이터의 불일치

데이터에서 언급되는 지역 명칭과 실제 호텔의 지역이 일치 하지 않는 문제

영등포에서 친구와 함께 놀기 좋은 호텔 검색

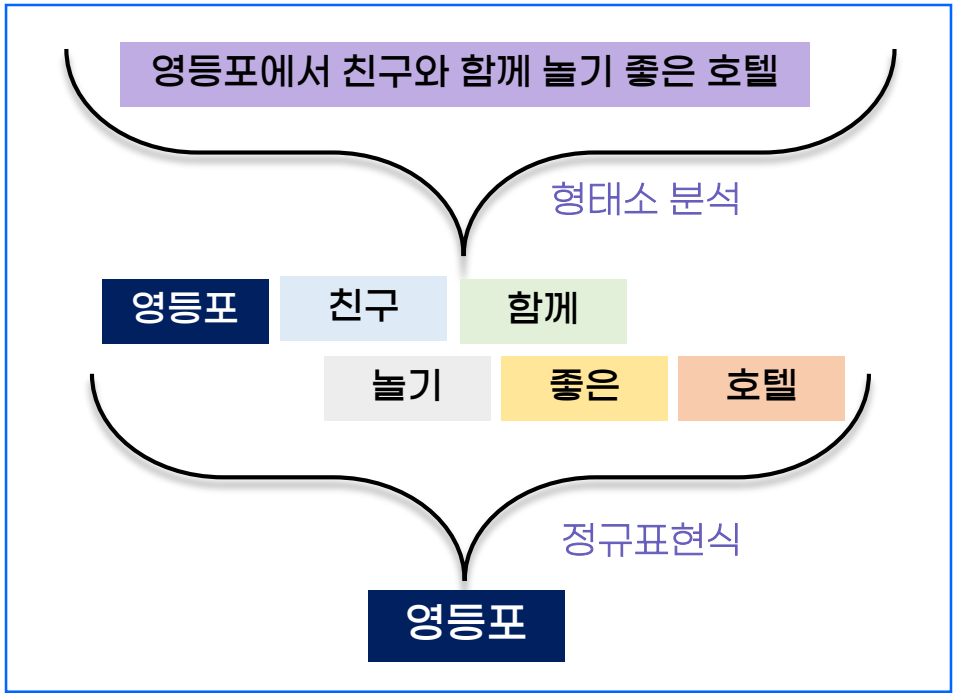
22.673677 / 중구 / 그리고 이 가격에 5성급 호텔에 친구 3명이서 함께 놀기에 딱 좋은 크기 구성이었어요
20.726984 / 중구 / 친구와 함께 놀려고 방문했는데 시설도 깔끔하고 복도에 정수기가 있어서 편리했어요
20.349552 / 중구 / 친구와 함께 구성비 서울 호캉스하기 너무 좋았어요
19.715473 / 강남구 / 고 시국에 친구와 구성비 좋게 놀기 좋은 곳 깨끗하고 담백하고 위치도 놀기 딱 좋음
18.857918 / 중구 / 친절한 프런트와 편안한 객실 가족 아이와 함께 하기 좋은 룸 컨디션 수영장 샤워실 너무너무 잘 놀고 갑니다
18.782799 / 중구 / 부산에서 친구 남편과 함께 놀러 왔는데 거부감 없이 깔끔하게 사용했습니다
18.674974 / 용산구 / 친구와 함께 갔는데 침구도 넓고 좋았습니다
18.636559 / 중구 / 친구와 함께 트윈 베드로 예약했어요
18.509735 / 영등포구 / 키즈 라운지 및 왜건 무료 이용 등 아기와 함께하기 좋은 호텔이었습니다
17.150501 / 구로구 / 영등포보다 덜 시끄럽고 좋네요

문제 해결

지역, 지하철역, 호텔 성급 등

숙소의 특성을 1차적으로 필터링하여 검색 정확도 향상

➡ 형태소분석, 정규표현식을 통해 해당 단어가 있을 경우
필터링하여 검색 조건에 가중치 부여



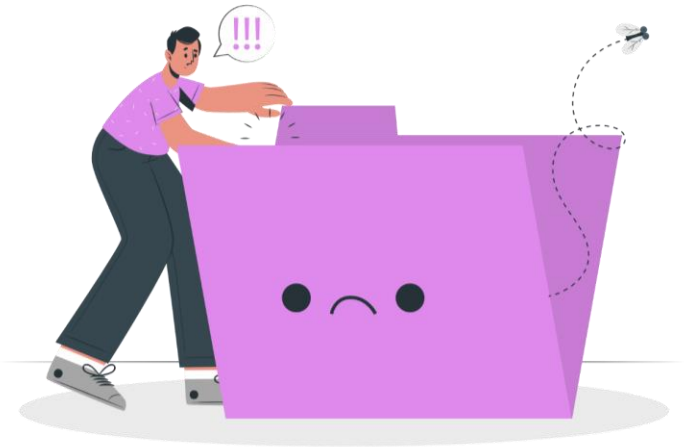
영등포에서 친구와 함께 놀기 좋은 호텔 검색

440.61902	영등포구	/ 여의도 한강공원 놀러 갔다가 친구들이랑 가기 딱 좋아요
430.20334	영등포구	/ 여자친구 생일이라 겸사겸사 여자 친구랑 놀러 갔는데 방도 크고 깔끔하고 너무 좋았네요
417.06793	영등포구	/ 깨끗 근데 탐스에서 놀다가 호텔 찾기가 너무 힘들었어요
415.2179	영등포구	/ 근처에 먹거리 타운도 가까워서 놀다가 술 마시고 자고 가기 좋아요
414.7667	영등포구	/ 다음에 친구들과 여행 시에도 저렴한 가격으로 숙박하기에 좋을 것 같습니다
414.6259	영등포구	/ 연인과 친구와 가족과 와도 좋은 위치 시설입니다
414.53528	영등포구	/ 역에서 가깝고 주변에 편의시설 등 많아서 이동하기 놀기 다 좋았어요
414.0736	영등포구	/ 타임스퀘어가 같은 건물에 있어서 놀기 너무 편하고 접근성이 좋아요
412.72748	영등포구	/ 쇼핑과 호텔 이용하기 너무 좋습니다
411.3476	영등포구	/ 객실 상태는 청결하고요 주변에 맛 집도 많고 쉬고 놀고 하기에는 좋아요

문제 제기

36개의 카테고리에 대한 의문

30만개의 리뷰 데이터를 확보했음에도 포함되는 리뷰 수 부족 문제



문제 해결

36개의 카테고리를 다시 통합하여 **10개**로 축소
-> 카테고리별 리뷰 수 확보

수정 전 카테고리별 평균 리뷰 수 : 14,428 개

⬇ 리뷰 수 3배 증가

수정 후 카테고리별 평균 리뷰 수 : 37,706 개

문제 제기

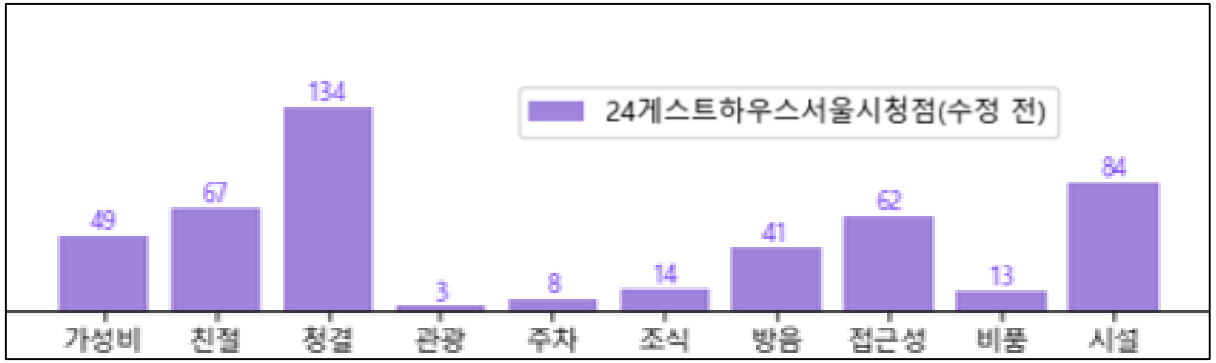
만족지수의 신뢰성 문제

각 호텔의 만족지수는 해당 호텔의 리뷰 수만을 기준으로 측정

리뷰 수가 적을 수록
상위 호텔에 랭크 될 가능성이 높음!

‘조식’언급 리뷰 수 14개

‘조식 긍정’ 언급 리뷰 수 12개



기본 조건에서 추천 받은 1위 숙소의 카테고리별 리뷰 수

문제 해결

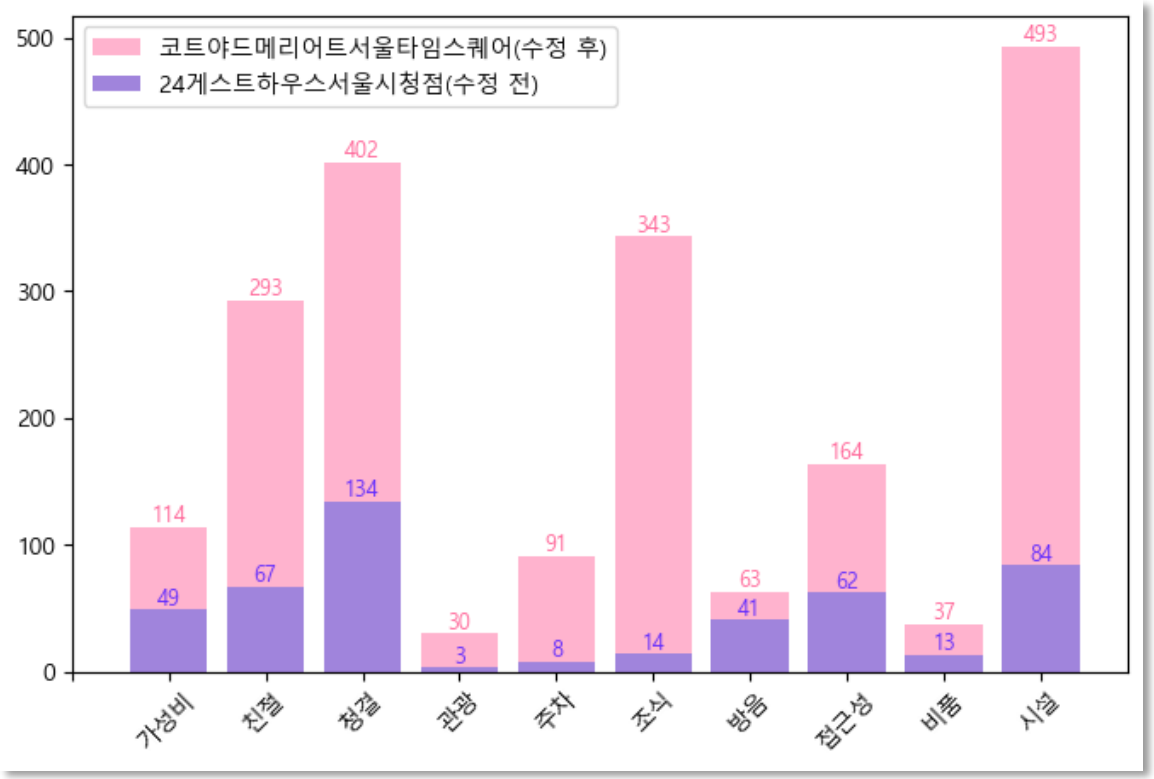
만족지수에 전체 호텔의 평균 카테고리 언급 리뷰 수로 가중치 부여

$$\text{호텔별 카테고리 점수} = \text{만족지수} \times \frac{\text{카테고리 언급 리뷰 수}}{\dfrac{\text{카테고리별 총 언급 수}}{\text{카테고리 언급 호텔 수}}}$$

Hotel SeSAC의 **조식 점수 다시** 구하기

$$\text{조식 만족지수} \times \frac{\text{SeSAC의 '조식' 언급 리뷰 수}}{\dfrac{\text{전체 호텔의 '조식' 언급 리뷰 수}}{\text{'조식'을 언급한 호텔 수}}}$$

동일한 조건에서 1순위 숙소의 리뷰 수 차이



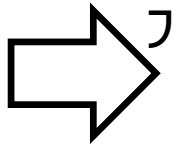
문제 제기

사용자가 지정한 카테고리 점수만을 고려하여 상위 숙소 추천
-> 사용자가 선택하지 않는 카테고리의 점수도 고려할 필요성

"가성비랑 조식이 중요해 "

	가성비	친절	청결	관광	주차	조식	방음	접근성	비품	시설
ht_id	x1.	x0	x0	x0	x0.	x1	x0	x0	x0	x0
1	66.8978	53.3862	38.9684	24.1121	3.7218	19.6568	14.2828	53.0949	9.5212	13.2588
2	79.7346	56.5584	27.5881	48.2243	0.0000	217.1516	12.1874	80.5917	7.6536	15.3840
3	1.8041	14.7865	8.4127	NaN	2.7914	16.3807	4.9268	6.7765	7.2508	8.0356
4	97.2546	85.0711	44.4523	16.0748	14.3557	2.8081	80.0375	56.5606	33.1656	32.0250
5	40.4117	61.9010	43.1160	16.0748	11.1655	12.6365	12.9954	20.2833	14.2818	16.5115
...	...	...	...	...	...	...	...	...	...	...

호텔별 카테고리 점수 테이블



가성비와 조식 카테고리 점수만 반영하여 계산
※그외 카테고리에서도 좋은 평가를 받은
숙소를 추천해주는 것이 중요!!

문제 해결

카테고리 가중치의 기본값을 1(보통)으로 설정
-> 사용자가 더 우수한 숙소를 추천받을 수 있도록 함

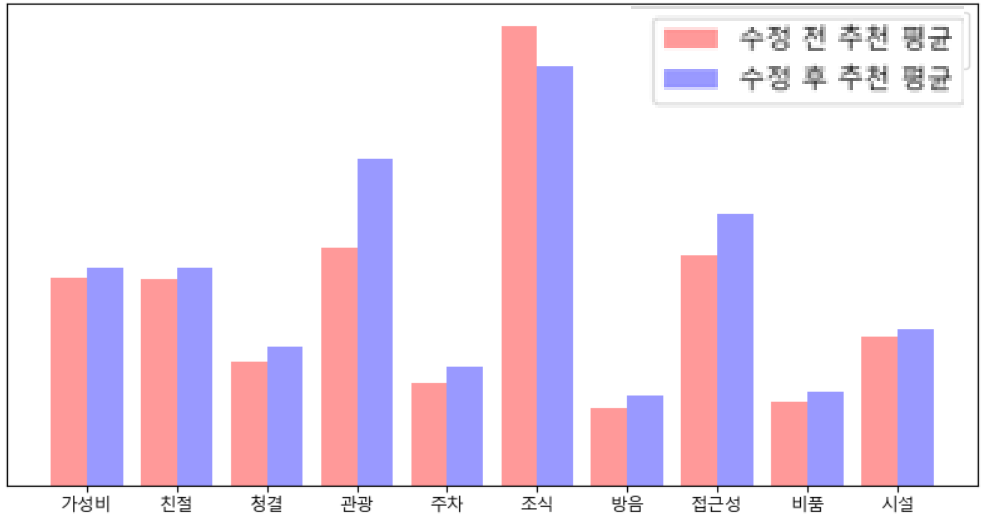
가중치 표

상관없음	0
덜 중요	0.5
보통(기본값)	1
중요	1.5
매우중요	2

가성비와 조식 외에도 가중치 부여 가능

	가성비	친절	청결	관광	주차	조식	방음	접근성	비품	시설
ht_id	x 2.	x 1.	x 1.	x 0.5	x 0.	x 2.	x 1.	x 1.	x 1.	x 1.5
1	66.8978	53.3862	38.9684	24.1121	3.7218	19.6568	14.2828	53.0949	9.5212	13.2588
2	79.7346	56.5584	27.5881	48.2243	0.0000	217.1516	12.1874	80.5917	7.6536	15.3840
3	1.8041	14.7865	8.4127	NaN	2.7914	16.3807	4.9268	6.7765	7.2508	8.0356
4	97.2546	85.0711	44.4523	16.0748	14.3557	2.8081	80.0375	56.5606	33.1656	32.0250
5	40.4117	61.9010	43.1160	16.0748	11.1655	12.6365	12.9954	20.2833	14.2818	16.5115
...	...	...	...	...	...	...	...	...	...	...

호텔별 카테고리 점수 테이블

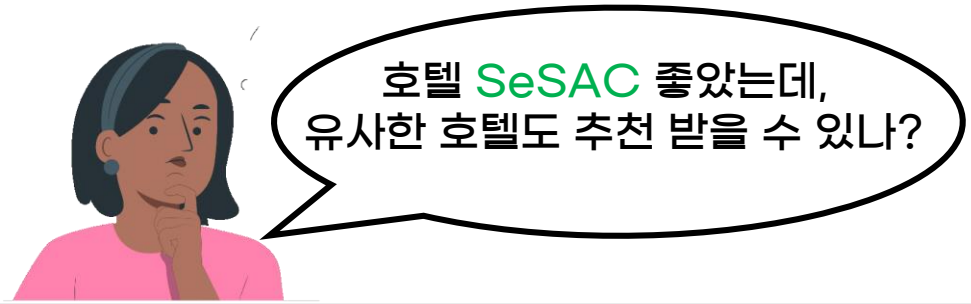


추천 호텔의 카테고리별 평균 점수 비교

	문제 제기 	문제 해결 
1	엘라스틱서치 검색 정확도 문제	문맥에 따라 유사도를 검색하는 문장임베딩 수행
2	리뷰 데이터의 불일치	숙소의 특성을 1차적으로 필터링하여 검색 정확도 향상
3	36개의 카테고리에 대한 의문	36개의 카테고리를 다시 통합하여 10개로 축소
4	만족지수의 신뢰성 문제	만족지수에 전체 호텔의 평균 카테고리 언급 리뷰 수 로 가중치 부여
5	사용자가 지정한 카테고리 점수만을 고려하여 상위 숙소 추천	카테고리 가중치의 기본값을 1(보통)으로 설정하여 사용자가 더 우수한 숙소를 추천받을 수 있도록 함

추가 기능

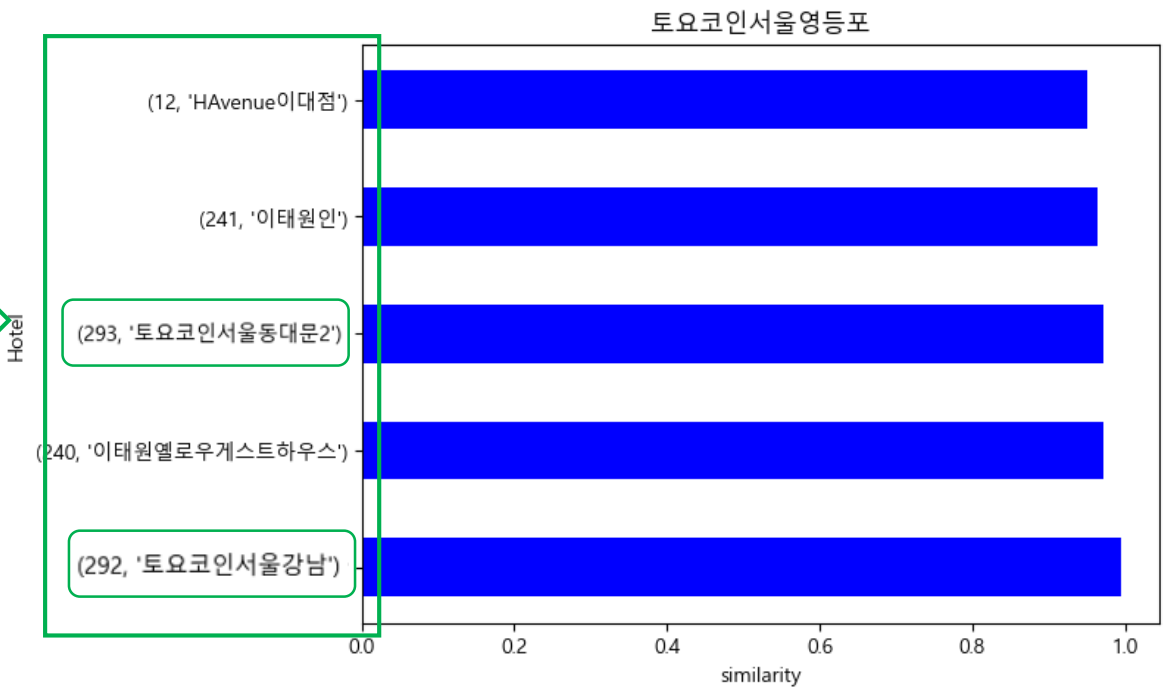
피어슨 상관계수를 사용하여
해당 호텔과 유사한 호텔을 추천



ht_id	가성비	친절	청결	주변시설	주차	조식	방음	위치	비품	시설
1	66.8978	53.3862	38.9684	24.1121	3.7218	19.6568	14.2828	53.0949	9.5212	13.2588
2	79.7346	56.5584	27.5881	48.2243	0.0000	217.1516	12.1874	80.5917	7.6536	15.3840
3	1.8041	14.7865	8.4127	NaN	2.7914	16.3807	4.9268	6.7765	7.2508	8.0356
4	97.2546	85.0711	44.4523	16.0748	14.3557	2.8081	80.0375	56.5606	33.1656	32.0250
5	40.4117	61.9010	43.1160	16.0748	11.1655	12.6365	12.9954	20.2833	14.2818	16.5115
...	...	...	...	...	...	...	...	...	...	...

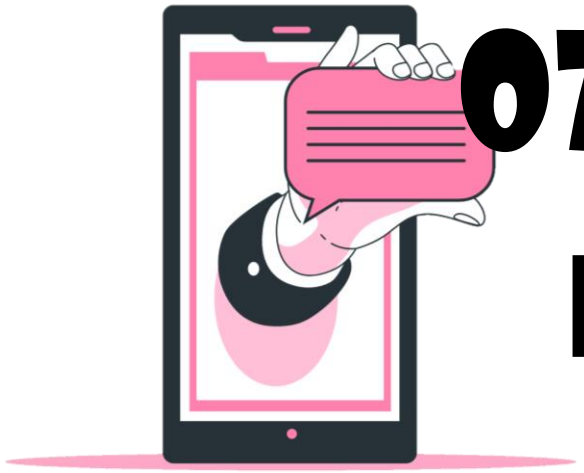
호텔별 카테고리 점수 테이블

상관분석(Correlation Analysis)



-> 동일한 체인의 호텔끼리 상관계수 높게 나옴

06 웹 구현



07

마무리

의의

문장임베딩을 통한 시멘틱 서치로 정확도 향상과 검색 피로도 최소화

비정형 데이터(리뷰)를 자연어처리 기술로 수치화하여 추천 알고리즘 구축

숙박리뷰 데이터 외에 타 리뷰 데이터에도 적용 가능한 서비스

한계점

중립데이터(Label 2)를 활용하지 못함

개인의 선호도를 긍정과 부정 이분법으로 한정

만족지수 가중치 반영에 있어서 임의성 부여

리뷰 크롤링으로 인해 상업화 불가 수익화 모델 생정하지 못함

향후계획

라벨2 데이터로 정보성 리뷰 선별 및 사용

주기적인 업데이트를 위한 스크래핑 및 전처리 자동화 시스템 구축

각 숙박 플랫폼과 제휴하여 법적 문제 없이 서비스를 구현할 수 있도록 함

국내 및 해외 주요 지역 서비스 확대

Q&A

방과방가

감사합니다

방과방가