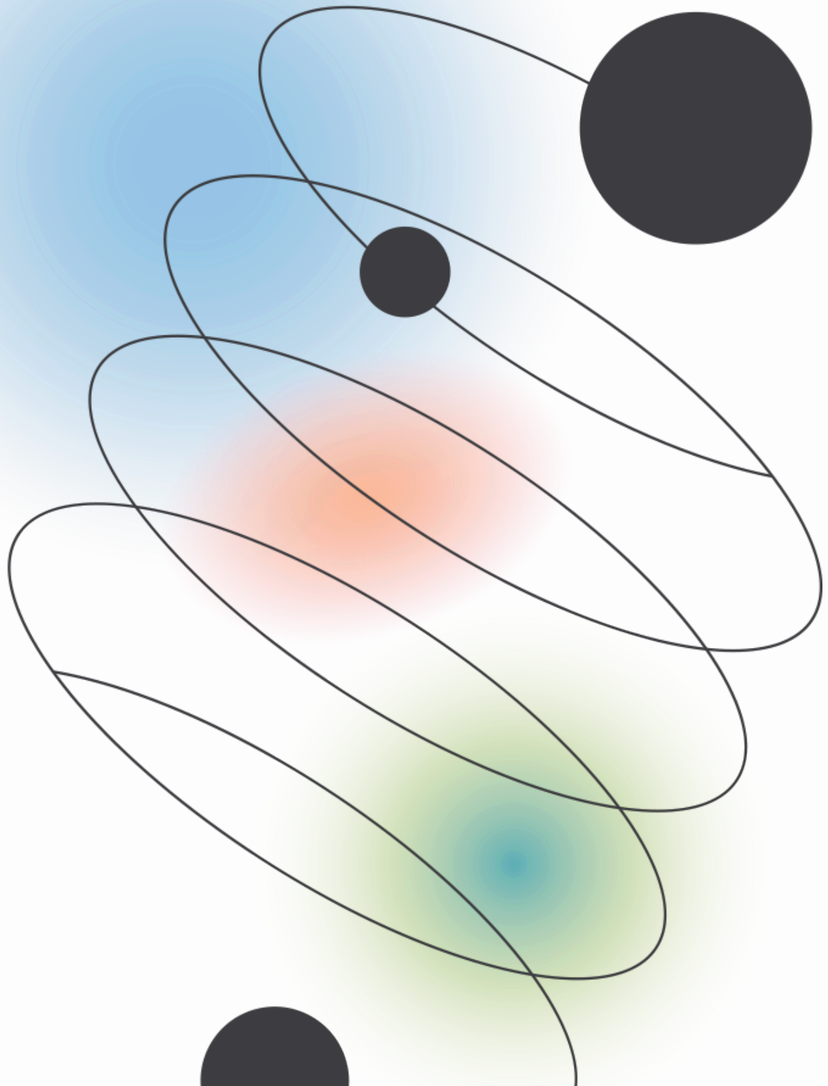
Agoda Hotel Review

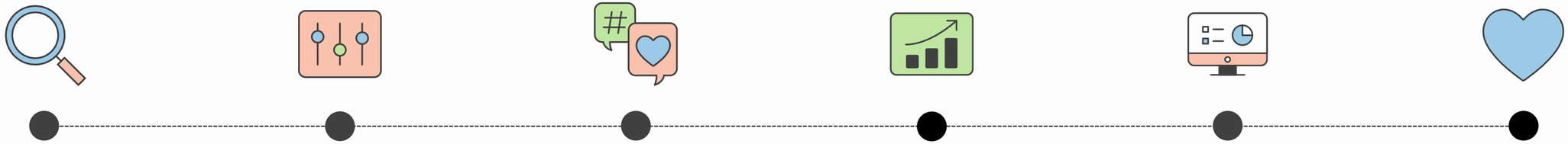


팀명: 웹크롤링 나빠 !

조영수 문예준

멋쟁이 부트캠프 9기

프로젝트 과정



01 주제 토론

호텔 리뷰 분석

02 데이터 수집

Selenium
웹크롤링

03 리뷰 토큰화

Gensim 불용어
nltk 토큰화

04 EDA

계절별
시각화
트렌드 분석

05 모델링

K-means, GMM,
DBSCAN 클러스터링
긍부정 예측모델

06 결론

의의 및 한계점

01
주제
선정 이유

01. 호텔을 이용하는 이유

-여가,출장,휴가 등 다양한 목적으로 이용

02. 호텔 플렛폼 선정

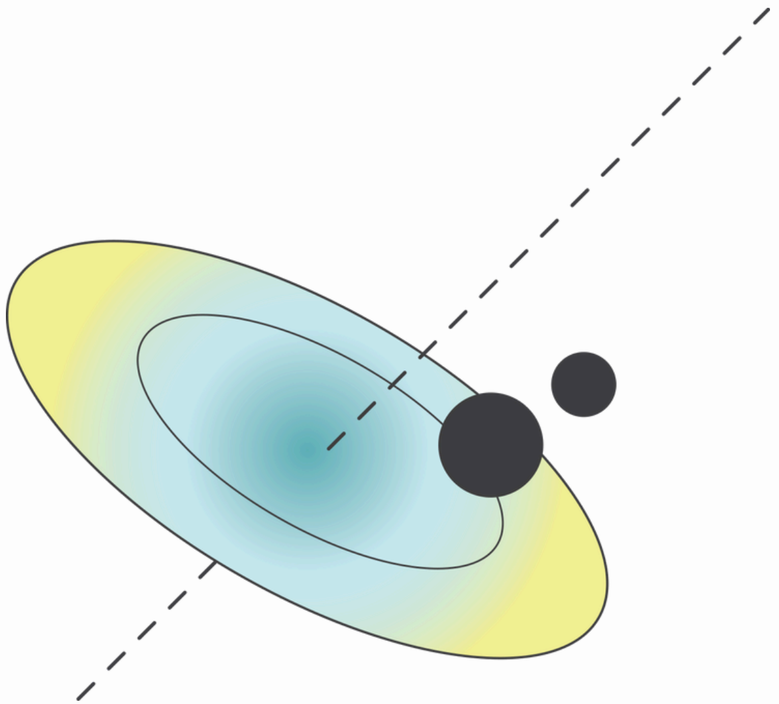
-아고다,여기어때,야놀자,호텔스닷컴 등 다양한 플렛폼이 존재
-가장 사용자가 많은 아고다에 다양한 리뷰들이 있기 때문에 선정

해외여행 숙소 검색 및 예약 채널

	검색 채널 (중복 응답)	예약 채널 (1~3순위)	예약 전환율 (평균 44.5%)
아고다	33.4	25.5	76.2%
호텔스닷컴	18.8	9.0	48.0%
호텔스 컴바인	18.2	8.5	47.0%
하나투어	13.8	8.4	60.6%
에어비앤비	13.2	7.7	58.4%
숙소에 직접 예약	7.7	5.8	75.6%
부킹닷컴	10.4	5.3	50.6%
모두투어	9.4	5.3	56.0%

오픈서베이 여행 트렌드 리포트2023

프로젝트 목표



01. 리뷰의 키워드 추출 후 전체적인 트렌드 분석

계절별(봄, 여름, 가을, 겨울),
Word Cloud, CountVectorize 를 통한 트렌드 분석

02. 리뷰 추천 모델 제안

사용자의 리뷰 내용을 바탕으로
호텔 리뷰의 긍부정 분류 예측 & 리뷰 점수에 대한 중요 요인 분석

클러스터링 모델들을 통해서 사용자 맞춤 리뷰 추천

02 데이터 수집

아고다 사이트에서 9 곳의 호텔 리뷰
약 1,900개의 리뷰 크롤링

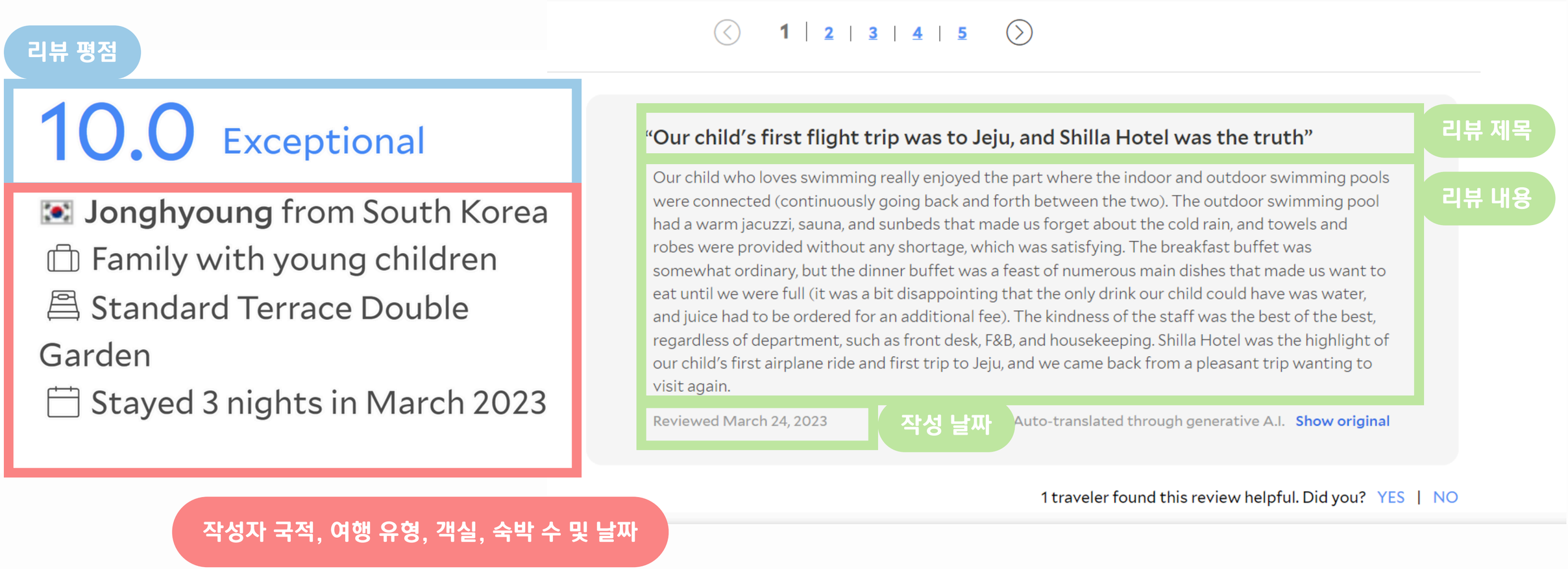
🍊 **제주**
신라호텔 + 호텔 2곳

🏘️ **서울**
신라호텔 + 호텔 2곳

🌉 **부산**
신라호텔 + 호텔 2곳

02 데이터 수집

Agoda 웹사이트 크롤링



02 데이터 수집

크롤링 결과

```
merged_df[merged_df.duplicated(subset='review_content')]
```

				children		
8	Worth every penny"₩nRoom is spacious, clean and neat. Every staff was so kind and willing to help that you don't really need to ask for a help(because they ask if you need help before you ask them for a help) Facilities are also clean and luxurious. If you are ready to pay for extra money for the higher quality service, this is the place.	Stayed 2 nights in September 2022	10.0	Couple	Worth every penny"	Reviewed September 14, 2022
9	Amazing grounds"₩nThe grounds and garden are simply beautiful. Really well kept and clean rooms.	Stayed 2 nights in April 2023	10.0	Solo traveler	Amazing grounds"	Reviewed April 14, 2023
...	...	...	...	...	...	...
3345	Nice hotel"₩nWe stayed @ Shilla Stay Haeundae for 4 nights - great location, very close to Haeundae beach & SeaLife Busan, very friendly staff, especially Front Desk and F&B Team. Highly recommend if you come & visit Busan.	Stayed 4 nights in December 2023	7.2	Business traveler	Uninspired"	Reviewed June 06, 2023
3346	Nice hotel"₩nWe stayed @ Shilla Stay Haeundae for 4 nights - great location, very close to Haeundae beach & SeaLife Busan, very friendly staff, especially Front Desk and F&B Team. Highly recommend if you come & visit Busan !!	Stayed 4 nights in December 2023	9.2	Couple	Exceptional"	Reviewed June 03, 2023

12,180 rows → 1,965rows

💡 크롤링 과정에서의 문제로
중복된 리뷰가 다수 존재

→ 최대한 다양한 호텔을
크롤링, 시간이 많이 소요

02 데이터 수집

파생 컬럼 생성, 추후 EDA 과정에서 활용

01

리뷰 길이

review_content
컬럼에서 len 함수 적용

02

숙박 날짜

date_text 컬럼에서
숙박한 연도-월 추출
-> 월은 Mapping

Stayed 2 nights
in April 2023

03

여행자 수

group 컬럼에서
couple: 2,
solo traveler: 1,
family: 4로 맵핑

Couple -> 2

04

숙박일수

date_text 컬럼에서
숙박한 일수 추출

Stayed 2 nights
in April 2023

05

작성자 국적

reviewer_country
컬럼에서 국적 추출

malco from Indonesia

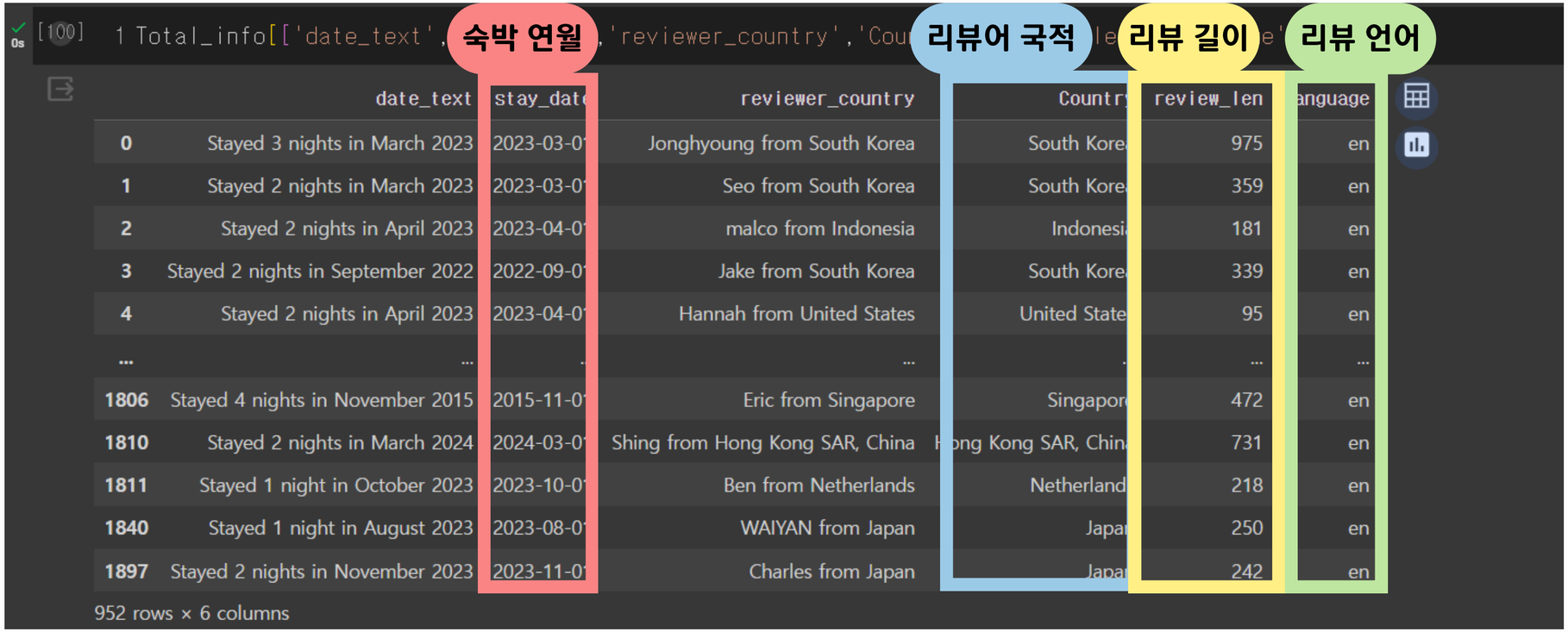
03 리뷰 토큰화

영어 추출, NLTK 토큰화, Gensim 불용어

03 리뷰 토큰화

전처리

- 전처리
  - 숙박 연월 형태 변환
  - 리뷰어 국적 추출
  - 리뷰 길이 추가
- 언어 통일
  - 영어 리뷰들만 추출



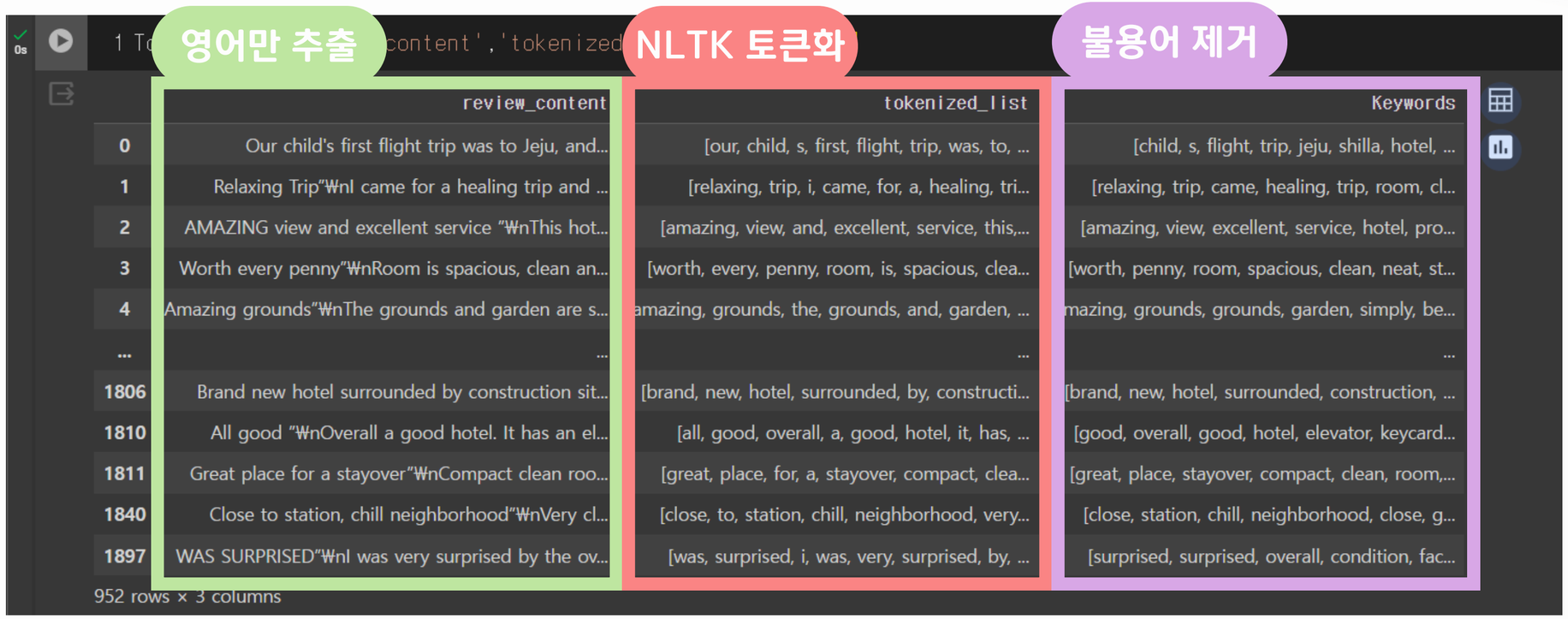
	date_text	stay_dat	reviewer_country	Country	review_len	anguage
0	Stayed 3 nights in March 2023	2023-03-0	Jonghyoung from South Korea	South Kore	975	en
1	Stayed 2 nights in March 2023	2023-03-0	Seo from South Korea	South Kore	359	en
2	Stayed 2 nights in April 2023	2023-04-0	malco from Indonesia	Indonesi	181	en
3	Stayed 2 nights in September 2022	2022-09-0	Jake from South Korea	South Kore	339	en
4	Stayed 2 nights in April 2023	2023-04-0	Hannah from United States	United State	95	en
...	...	.	...	.	...	...
1806	Stayed 4 nights in November 2015	2015-11-0	Eric from Singapore	Singapor	472	en
1810	Stayed 2 nights in March 2024	2024-03-0	Shing from Hong Kong SAR, China	ong Kong SAR, Chin	731	en
1811	Stayed 1 night in October 2023	2023-10-0	Ben from Netherlands	Netherland	218	en
1840	Stayed 1 night in August 2023	2023-08-0	WAIYAN from Japan	Japa	250	en
1897	Stayed 2 nights in November 2023	2023-11-0	Charles from Japan	Japa	242	en

952 rows × 6 columns

03 리뷰 토큰화

- 토큰화
  - NLTK를 사용하여 토큰화
- 불용어 제거
  - Gensim 불용어 사용

영어만 추출 NLTK 토큰화 불용어 제거



	review_content	tokenized_list	Keywords
0	Our child's first flight trip was to Jeju, and...	[our, child, s, first, flight, trip, was, to, ...	[child, s, flight, trip, jeju, shilla, hotel, ...
1	Relaxing Trip"₩nI came for a healing trip and ...	[relaxing, trip, i, came, for, a, healing, tri...	[relaxing, trip, came, healing, trip, room, cl...
2	AMAZING view and excellent service "₩nThis hot...	[amazing, view, and, excellent, service, this,...	[amazing, view, excellent, service, hotel, pro...
3	Worth every penny"₩nRoom is spacious, clean an...	[worth, every, penny, room, is, spacious, clea...	[worth, penny, room, spacious, clean, neat, st...
4	Amazing grounds"₩nThe grounds and garden are s...	amazing, grounds, the, grounds, and, garden, ...	mazing, grounds, grounds, garden, simply, be...
...	...	...	...
1806	Brand new hotel surrounded by construction sit...	[brand, new, hotel, surrounded, by, constructi...	[brand, new, hotel, surrounded, construction, ...
1810	All good "₩nOverall a good hotel. It has an el...	[all, good, overall, a, good, hotel, it, has, ...	[good, overall, good, hotel, elevator, keycard...
1811	Great place for a stayover"₩nCompact clean roo...	[great, place, for, a, stayover, compact, clea...	[great, place, stayover, compact, clean, room,...
1840	Close to station, chill neighborhood"₩nVery cl...	[close, to, station, chill, neighborhood, very...	[close, station, chill, neighborhood, close, g...
1897	WAS SURPRISED"₩nI was very surprised by the ov...	[was, surprised, i, was, very, surprised, by, ...	[surprised, surprised, overall, condition, fac...

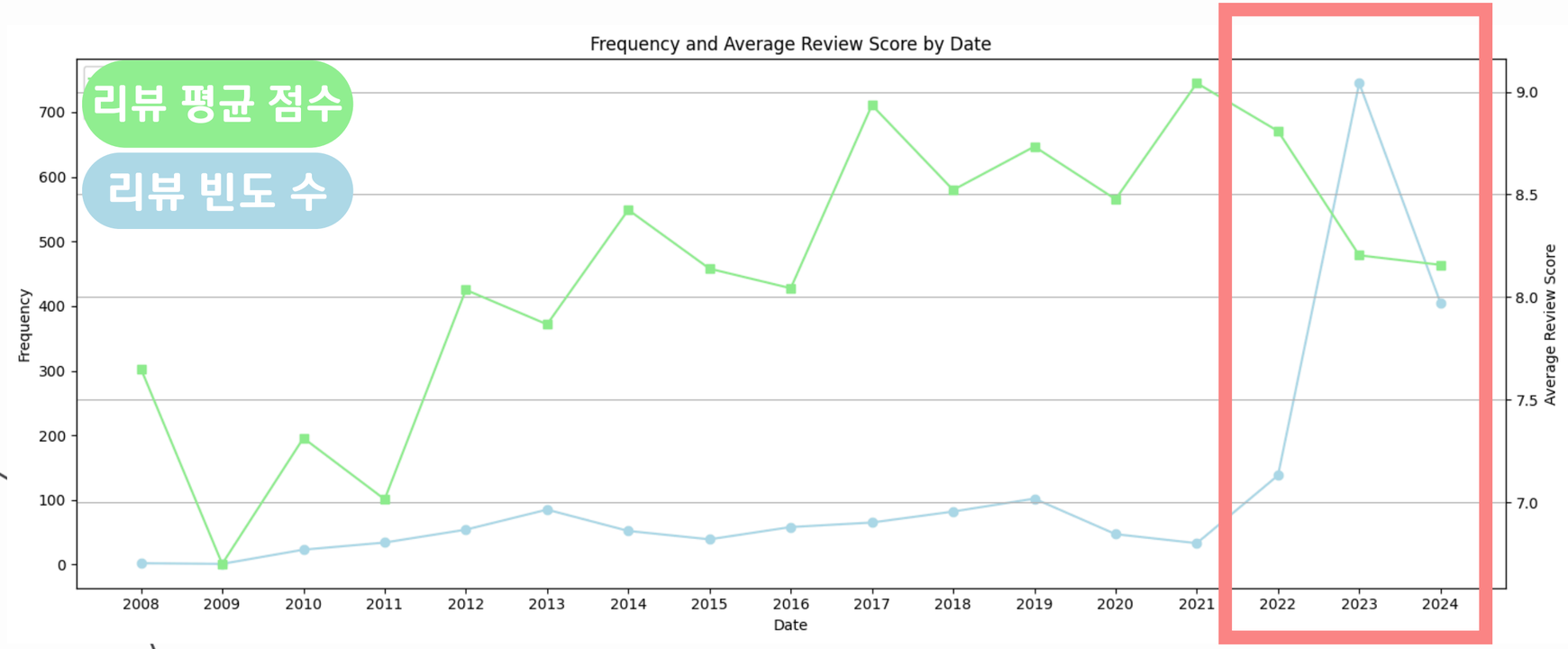
952 rows × 3 columns

04 EDA

호텔 데이터 트렌드 시각화
계절별 WordCloud

04 EDA

호텔 리뷰 트렌드 분석

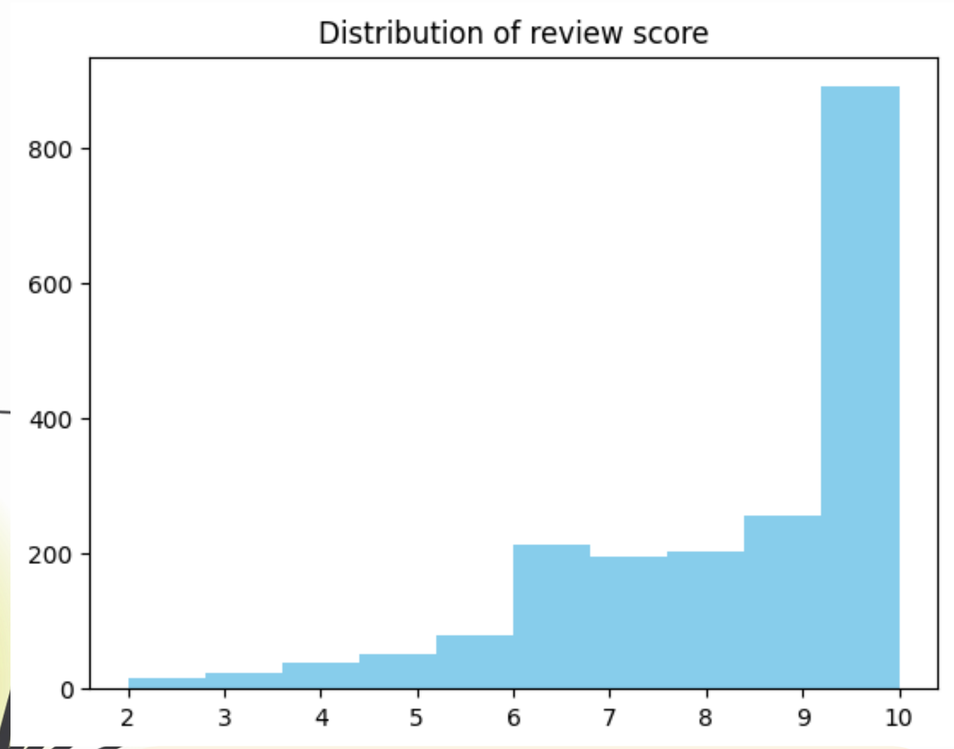


2023년도부터 아고다 사용 증가

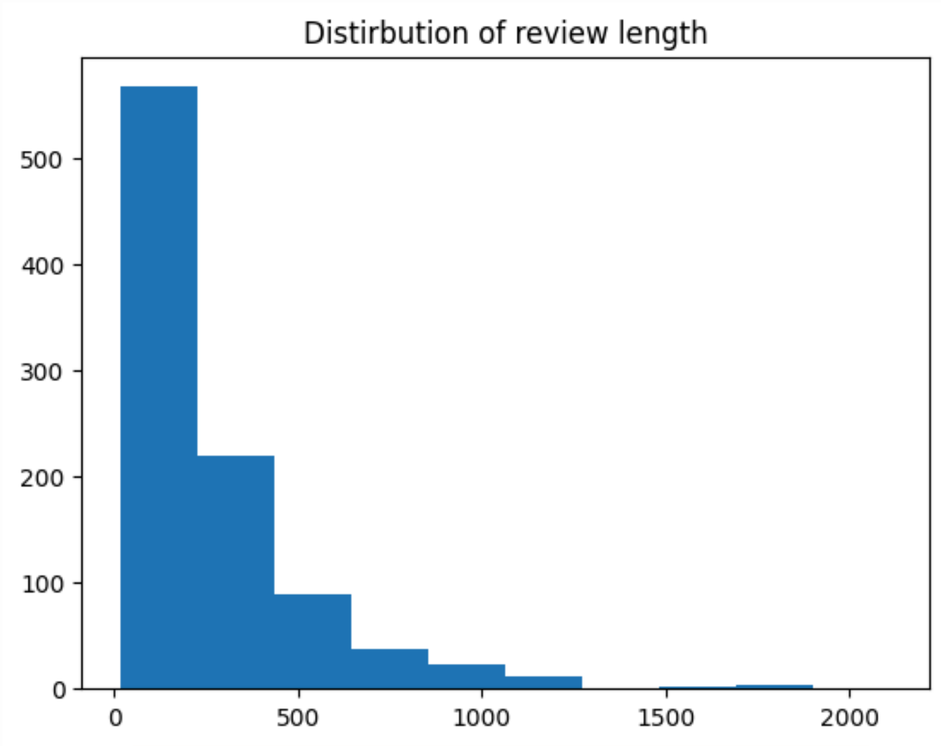
리뷰수 증가 📈 & 평균 리뷰 점수는 감소 📉

04 EDA

호텔 리뷰 트렌드 분석



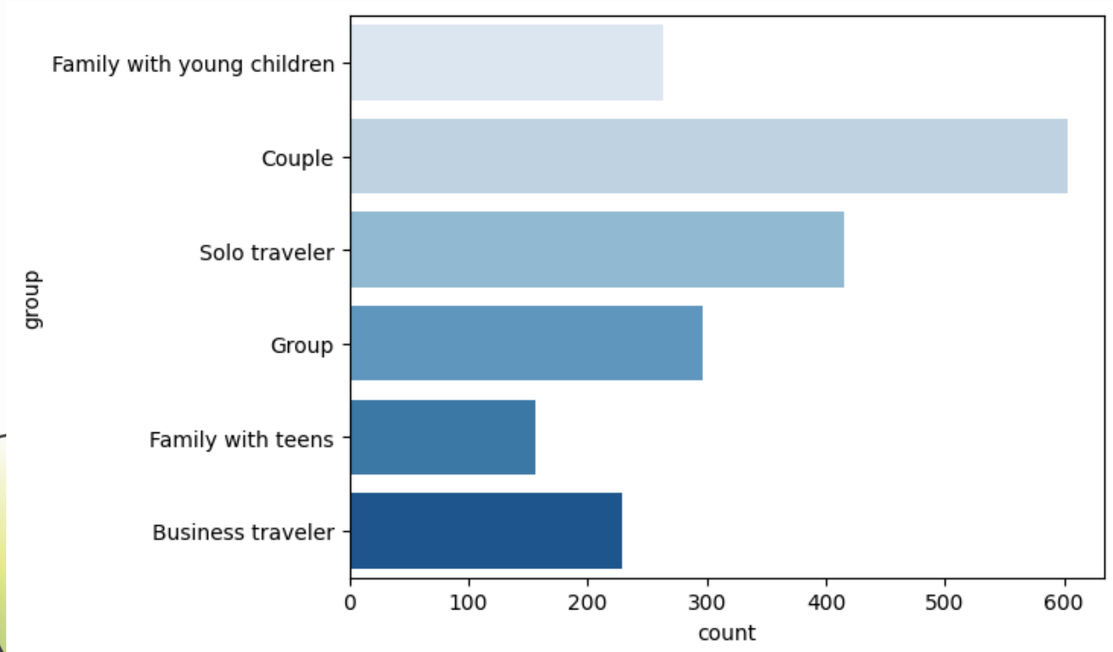
리뷰 점수 분포

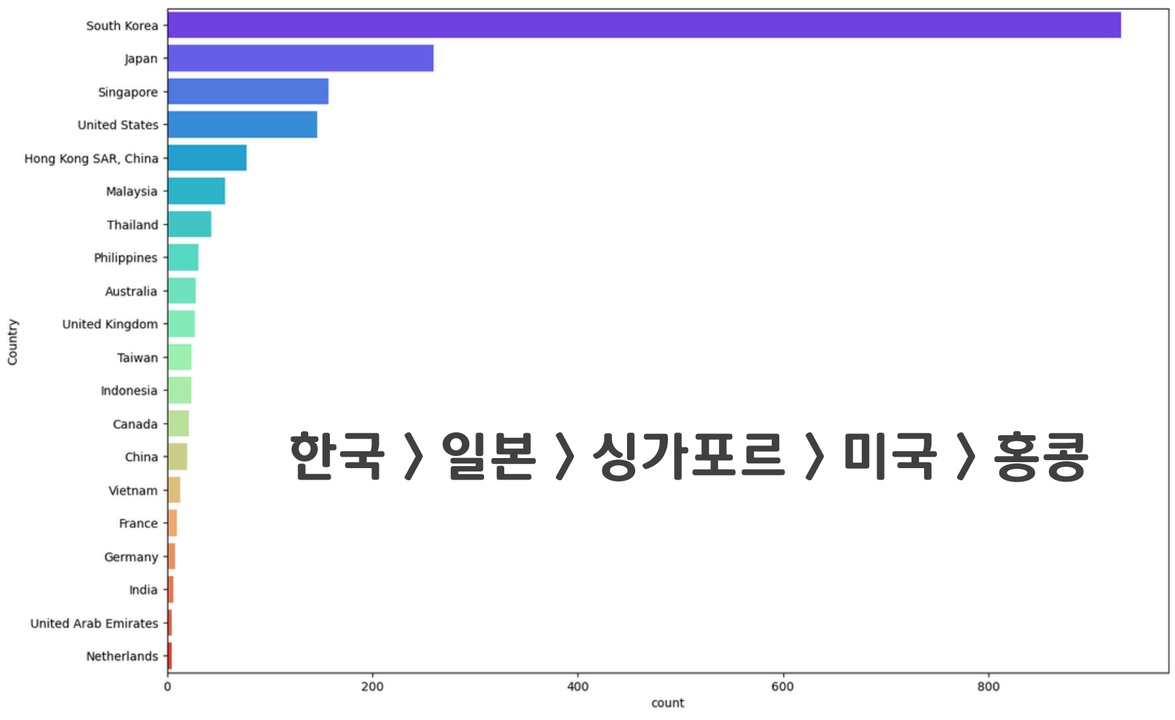


리뷰 길이 분포

04 EDA

호텔 리뷰 트렌드 분석





한국 〉 일본 〉 싱가포르 〉 미국 〉 홍콩

👫 여행 유형 빈도 👫

🇰🇷 리뷰자 국적 top 20 🏴

04 EDA

정규성 검정 : 샤피로-윌크 검정을 통해 정규성 검정

```python
import scipy.stats as stats
# 표본 크기가 2000 이하이므로 샤피로-윌크 검정 수행
for column in desired_columns:
    print(f"변수 '{column}'의 정규성 검정 결과:")

    # 샤피로-윌크 검정
    shapiro_stat, shapiro_p = stats.shapiro(df[column])
    print(f"샤피로윌크스 검정 - 통계량(statistic): {shapiro_stat:.4f}, p-value={shapiro_p:.4f}")
```

```
변수 'review_len'의 정규성 검정 결과:
샤피로윌크스 검정 - 통계량(statistic): 0.6881, p-value=0.0000
변수 'stay_day'의 정규성 검정 결과:
샤피로윌크스 검정 - 통계량(statistic): 0.5368, p-value=0.0000
변수 'people_num'의 정규성 검정 결과:
샤피로윌크스 검정 - 통계량(statistic): nan, p-value=1.0000
변수 'review_score'의 정규성 검정 결과:
샤피로윌크스 검정 - 통계량(statistic): 0.8622, p-value=0.0000
```

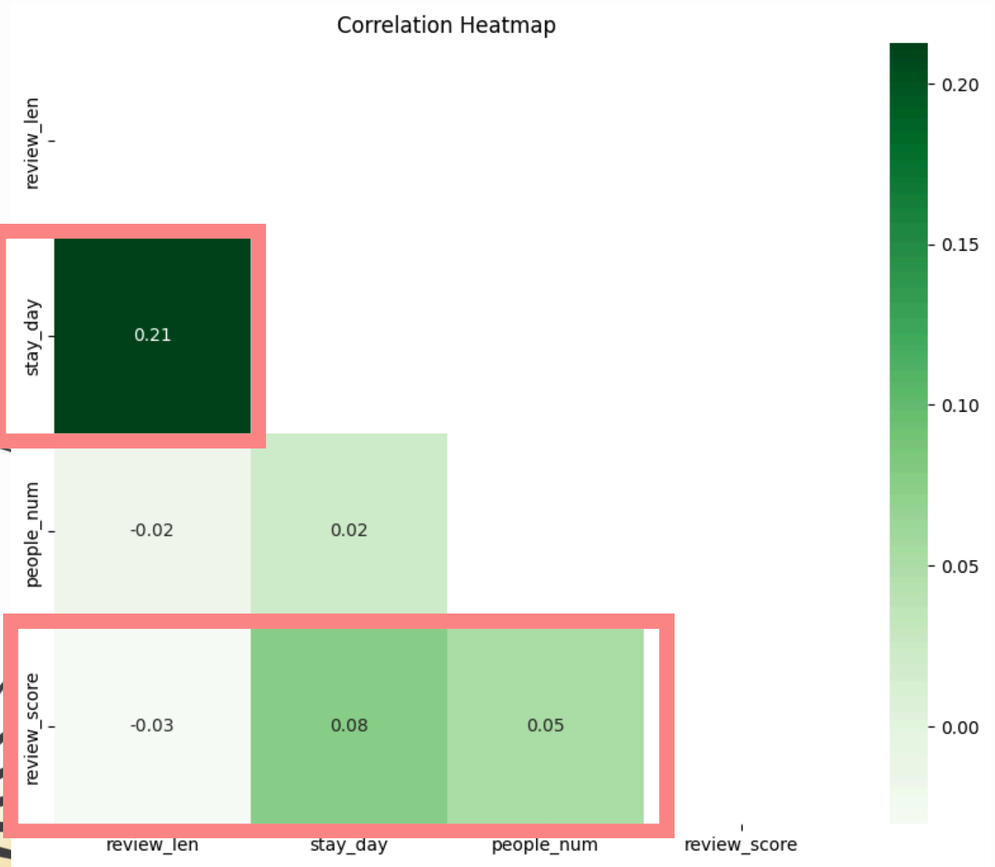
비모수 검정

추후 비모수 검정 방법을 실시하는 것이 적합하다 판단

표본의 크기가 2000 이하이기 때문에
샤피로-윌크 검정을 실시
정규성 검정 결과, 모든 변수가 정규성을
만족하지않기 때문에 비모수 검정을
실시해야 하는 것으로 판단

04 EDA

상관계수 Heatmap(비모수 검정인 spearman 방식)



stay_day(숙박일수) - review_len(리뷰길이)

0.21의 약한 상관관계이지만,
숙박일수와 리뷰길이와의 **양의 상관관계** 확인
ex) review_len = 1130, stay_day = 6

review_score(리뷰 평점)와 다른 컬럼

추후 **모델링**에서 활용할 컬럼을 찾기 위해
상관관계 분석을 진행하였으나,
리뷰 평점과의 **뚜렷한 상관관계를 갖는 변수 X**

04 wordcloud (제주 신라 호텔)

Spring

Summer

Winter

Fall

04 wordcloud (호텔 전체)

Spring

Summer

Winter

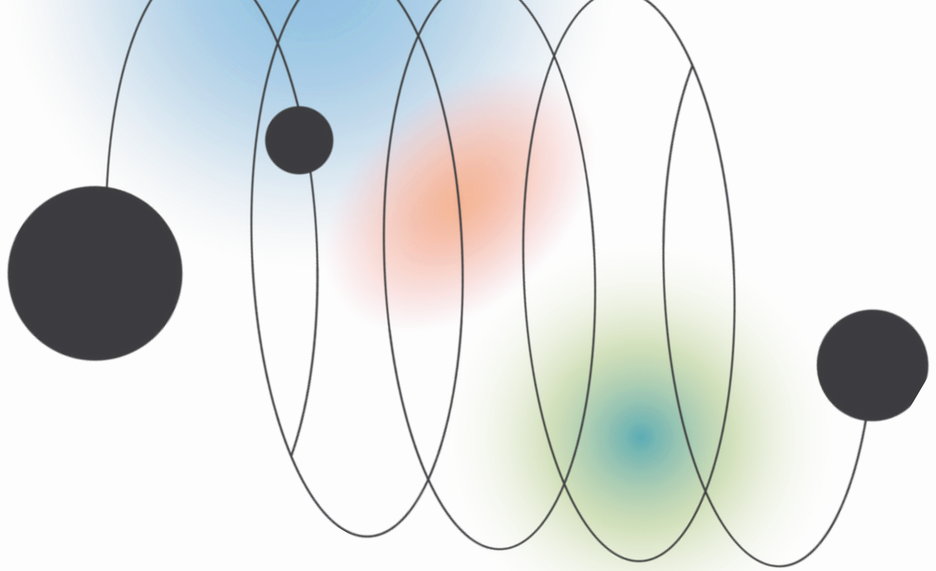
Fall

05 모델링

K-Means, GMM, DBSCAN 클러스터링
긍부정 예측 모델

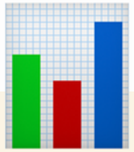
05 리뷰 분류

K-Means 클러스터링, GMM 클러스터링, DBSCAN



K-Means 클러스터링

유클리드 거리를 기반으로
가까운 클러스터에 할당, **wcss를 최소화**
최적화: elbow 메소드, silhouette 스코어



GMM 클러스터링

확률 기반 클러스터링 모델
데이터가 **여러 가우스 분포의 혼합**에서 생성된다고 가정, 최적화: BIC, AIC

DBSCAN

다양한 모양, 크기의 클러스터 + **노이즈**가 있는 데이터셋에 효과적, 특정 공간의 **밀도**로 클러스터링, 변수: n_samples, eqs

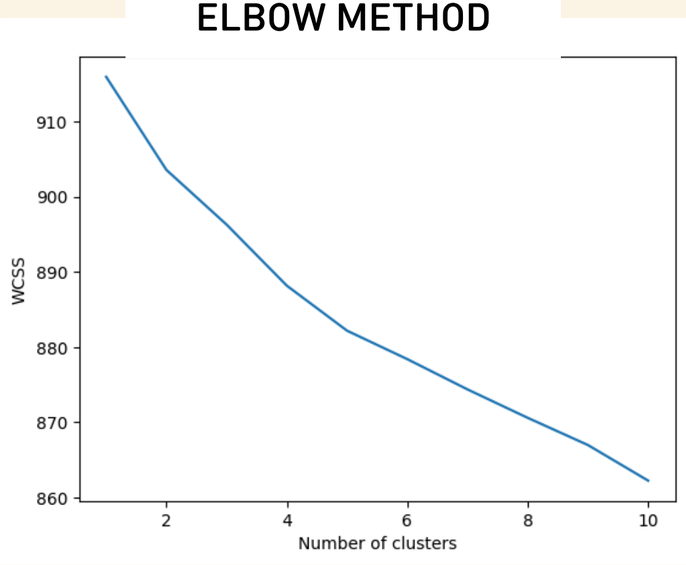
✅ 채택된 방식

리뷰 내용을 TF-IDF로 벡터화 한 후, 3가지 모델을 이용하여 리뷰 클러스터링

05 리뷰 분류

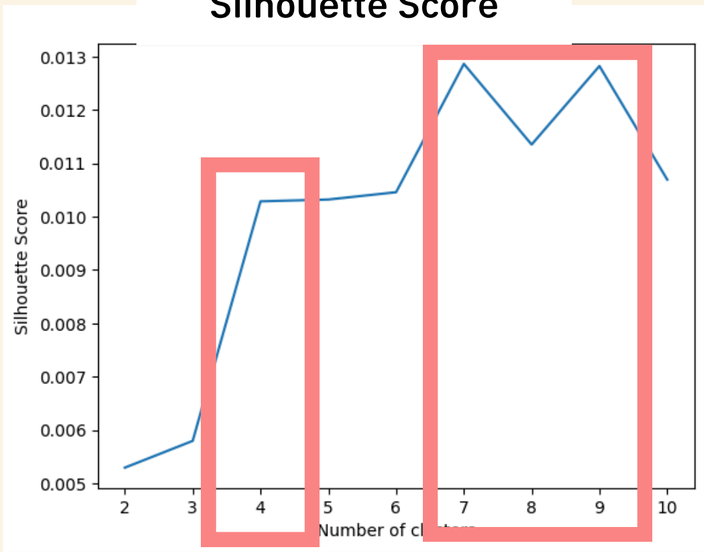
👫 K-Means 클러스터링 최적화

ELBOW METHOD



Elbow method를 통해 WCSS를 계산
WCSS가 크게 감소하는 뚜렷한 지점이 없음
4 또는 5 지점

Silhouette Score



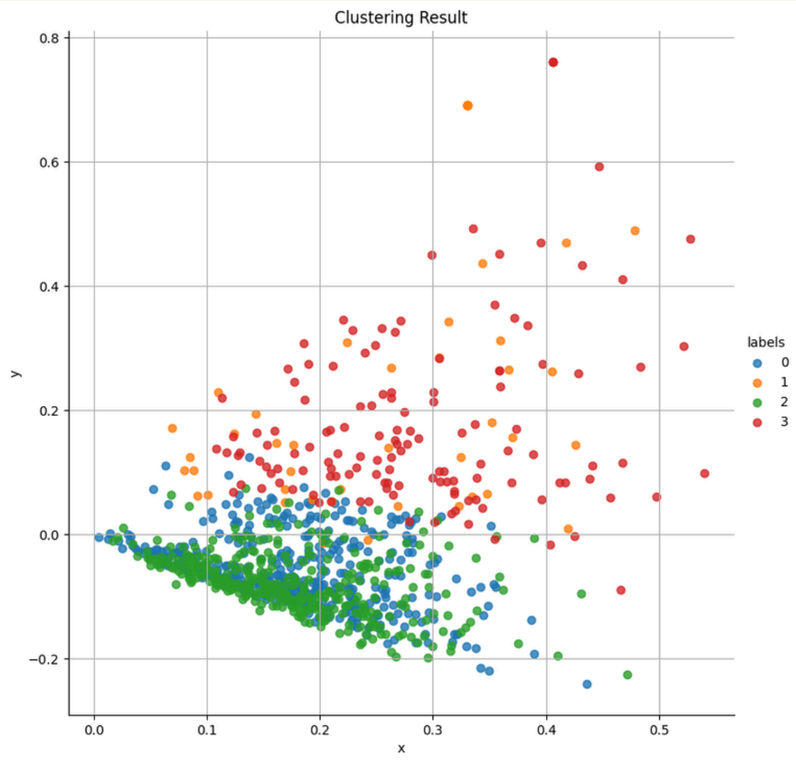
Silhouette Score를 확인한 결과,
7과 9가 가장 높았고
추가적인 두드러진 부분은 4

클러스터 4개

Elbow method에서는
두드러진 부분이 없었음
&
Silhouette Score은 7과 9가
높았으나, 데이터가 적기 때문에
많은 클러스터는 필요하지 않다고 판단

05 리뷰 분류

👫 K-Means 클러스터링, k = 4



4️⃣ k = 4일때 분포 시각화

```
kmeans = KMeans(n_clusters=4, init='k-means++', n_init=10, random_state=0)
kmeans.fit(X)
```

KMeans

KMeans(n_clusters=4, n_init=10, random_state=0)

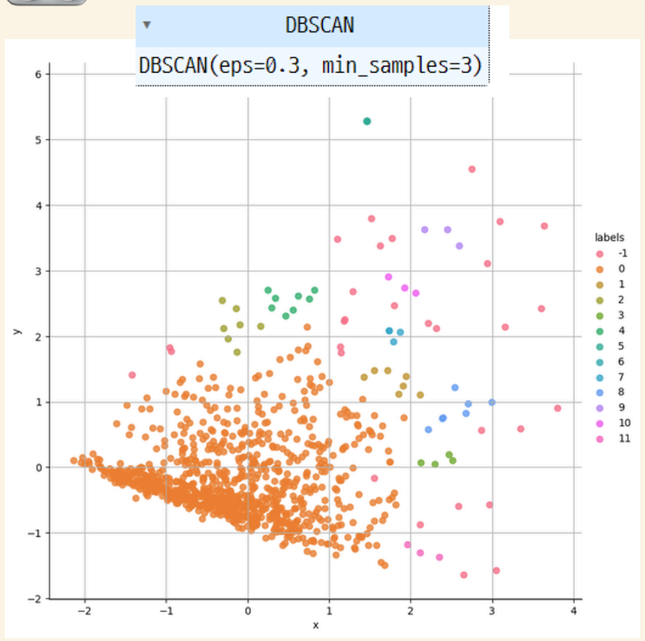
Label 0과 Label 2 / Label 1과 Label 3이 제대로 분류가 되지 않은 것을 판단

추가적으로 다른 모델 적용
GMM 클러스터링과
DBSCAN 클러스터링
적용

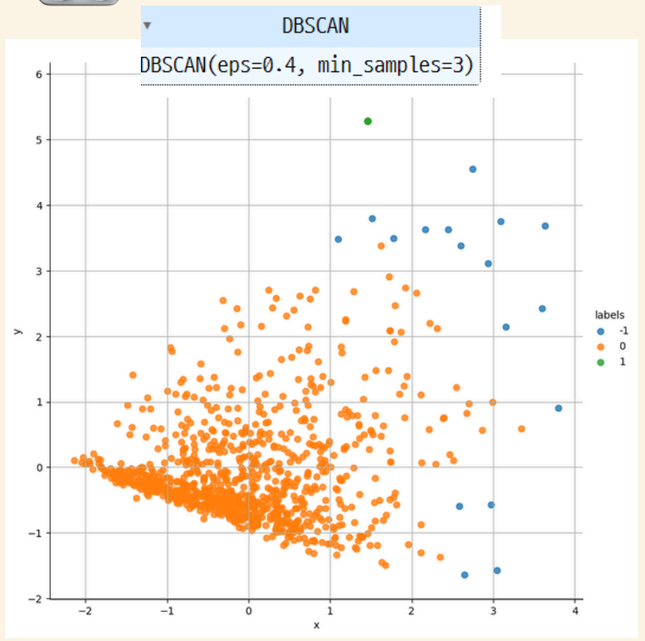
05 리뷰 분류

👥 DBSCAN 파라미터 탐색

🎛️ eps: 0.3, min_sample: 3



🎛️ eps: 0.4, min_sample: 3



EPS (eplison,최대 탐색거리)

TF-IDF 벡터화 + StandardScaler 표준화로
데이터 값들이 **다소 작음**
따라서 **EPS 값에 민감하게 반응**

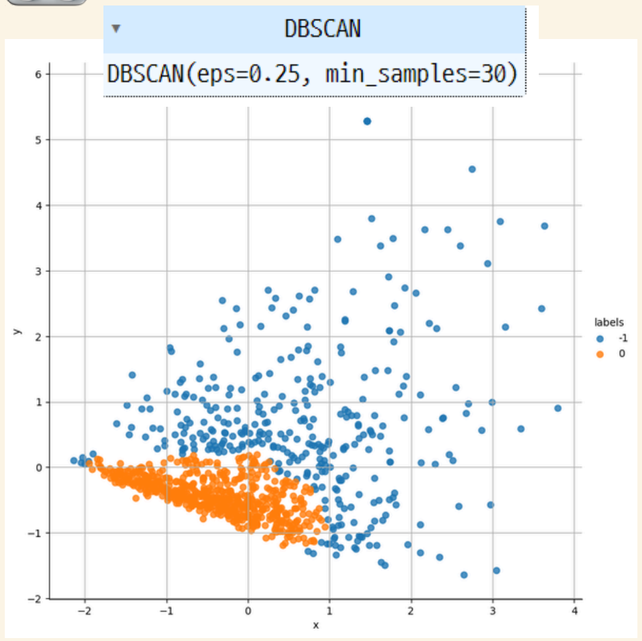
min_sample (최소 샘플개수)

min_sample이 3일때,
Label = 0 에 데이터가 몰림, 분류가 제대로 X
따라서 **min_sample 값을 키워야 함**

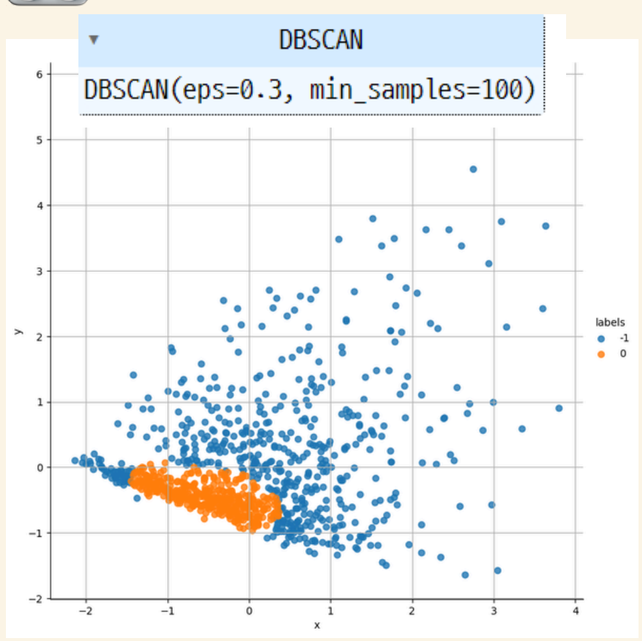
05 리뷰 분류

DBSCAN 파라미터 탐색

eps: 0.25, min_sample: 30



eps: 0.3, min_sample: 100



min_sample (최소 샘플개수)

min_sample의 값을 **키운 후**
Label = 0 (min_sample = 3)의 쏠림을 줄임
다른 모델과 달리 **패턴(곡선)으로 분류**하는 경향

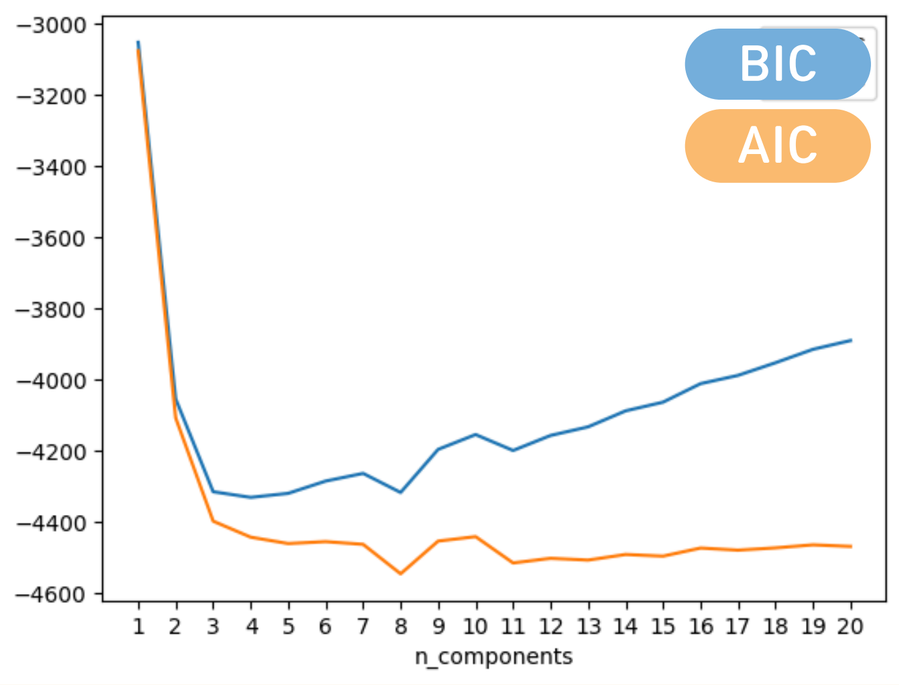
하지만 900개의 리뷰를
두개의 Cluster로
분류한 결과에 아쉬움이 존재

GMM 모델 사용 !!

05 리뷰 분류

GMM 클러스터링

n_components 변화에 따른 최적화



BIC **B**ayeisan **I**nformation **C**riterion

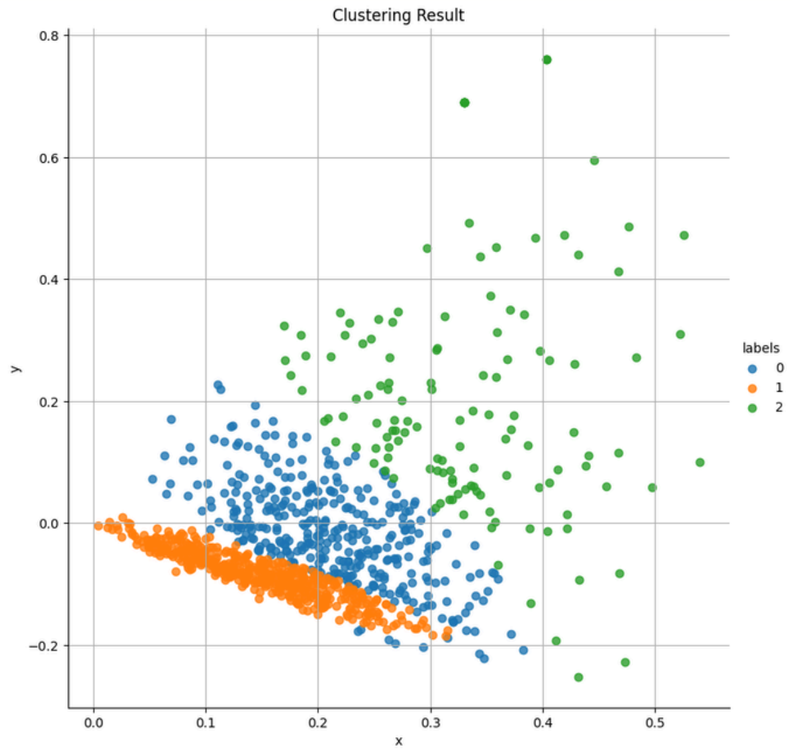
AIC **A**kaike **I**nformation **C**riterion

확률 모델 GMM -> **log-likelihood 기반** AIC & BIC 사용
값이 작을수록 모델 성능이 좋음
다를 경우, BIC 모델이 AIC 모델보다 간단한 경향

BIC가 가장 낮은 n_components가 3일때로
GMM 클러스터링 모델 최적화

05 리뷰 분류

GMM 클러스터링 결과



3 n_components = 3일때 분포 시각화

Cluster	review 수	평균 평점
label 0	348	8.588
label 1	474	8.547
label 2	132	8.70

총 3개의 Cluster
review 수는 label 1의 숫자가 가장 많았으며,
평균적으로 label 2의 **평점**이 높았지만, 큰 차이는 없음

05 리뷰 분류

GMM 클러스터링 결과

Label 0 워드 클라우드 (하위 80%)


```
cluster0.hotel.value_count
```

hotel	
busan_shilla	102
seoul_shilla	64
jeju_shilla	60
busan_paragon	49
seoul_riverside	30
jeju_bluespring	18
busan_marianne	16
jeju_cordelia	9

Name: count, dtype: int64

활동적인 단어들

- shops,
- outdoor,
- gym

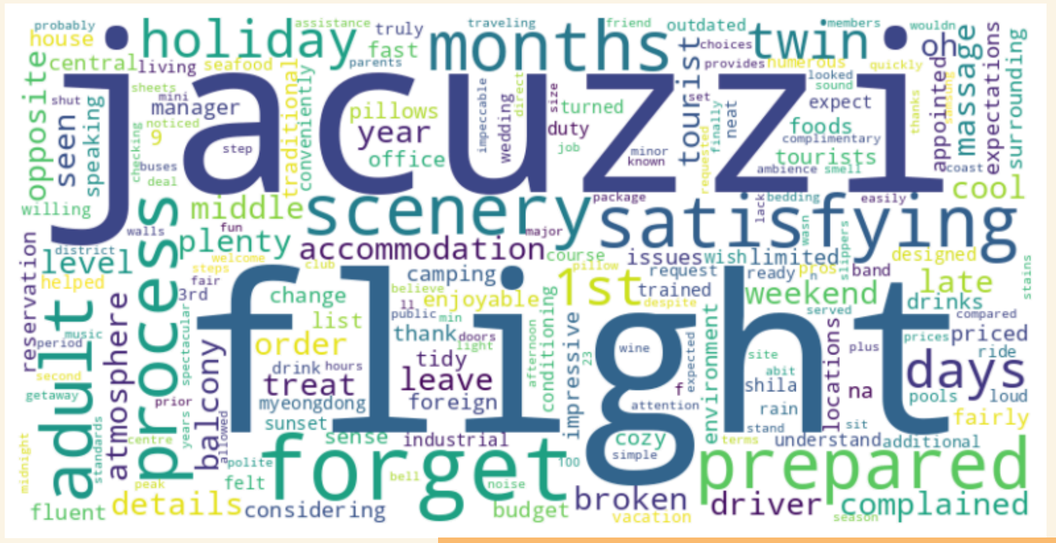
부산 호텔의 높은 비중

호텔 주변 Outdoor 활동이나 호텔 내에 있는
gym 에 관심이 많은 사람에게 적합

05 리뷰 분류

GMM 클러스터링 결과

Label 1 워드 클라우드 (하위 80%)




```
hotel
seoul_shilla        127
jeju_shilla         121
busan_shilla         90
busan_paragon        57
seoul_riverside      41
jeju_bluespring      15
busan_marianne       12
jeju_cordelia        11
Name: count, dtype: int6
```

자쿠지(욕조)
-> 서울 위주의 호텔,

flight(비행)
-> 제주 위주의 호텔

자쿠지와 같은 욕조나 스파에 대한 리뷰가
궁금한 사람에게 적합

05 리뷰 분류

GMM 클러스터링 결과

Label 2 워드 클라우드 (하위 80%)


```
hotel
busan_shilla        42
seoul_shilla        22
jeju_shilla         19
busan_paragon       16
jeju_bluespring     10
busan_marianne      10
seoul_riverside     10
jeju_cordelia        3
Name: count, dtype: int6
```

shilla
-> 신라호텔의 상호명

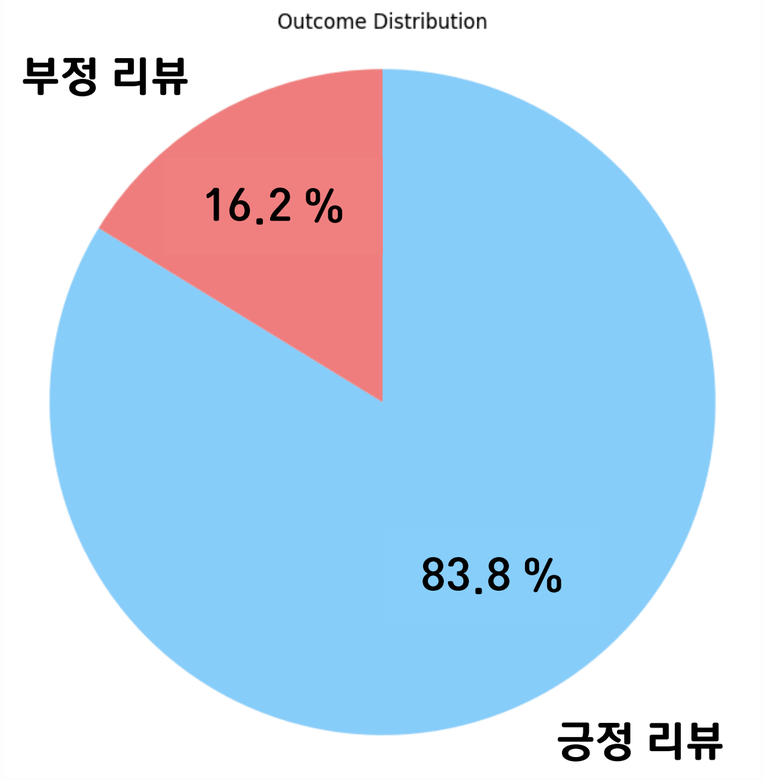
family
-> 가족 여행객

sea, busan
-> 부산 위주의 호텔

가족 구성원들끼리 가는 여행을 준비 중인 사람이나
바다 주변 부산 호텔을 찾아보는 사람에게 적합

05 긍부정 분석

부정(리뷰 평점: 0 ~7점) - 1 / 긍정(리뷰 평점: 7 ~ 10점) - 0
분류 모델인 LogisticRegression 사용



모델 사용 변수

Tf-idf 벡터화, stay_day(숙박 일수),
클러스터 label(GMM 모델), review_len(리뷰 길이)

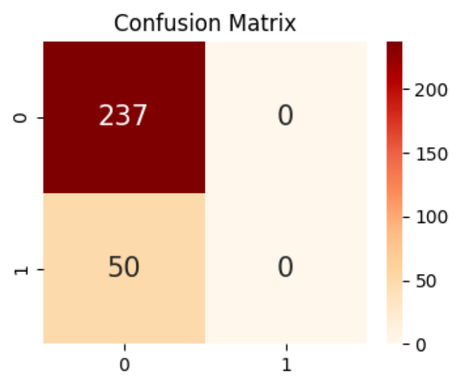
모델링

Logistic Regression 모델 사용

```
LogisticRegression
LogisticRegression(random_state=0)
```

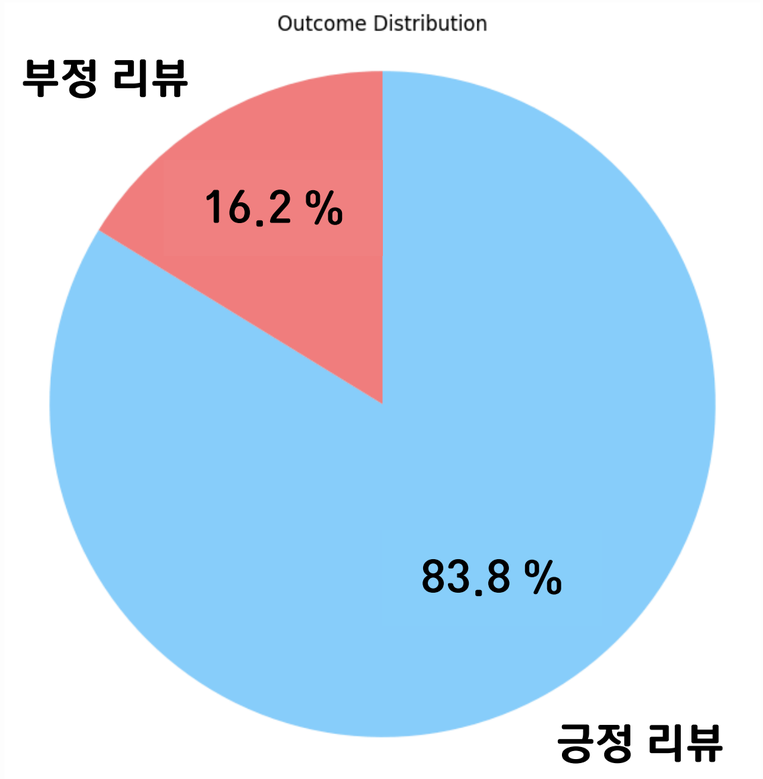
모델 성능

accuracy: 0.83
precision: 0.00
recall: 0.00
F1: 0.00



05 긍부정 분석

부정(리뷰 평점: 0 ~7점) - 1 / 긍정(리뷰 평점: 7 ~ 10점) - 0
분류 모델인 LogisticRegression 사용

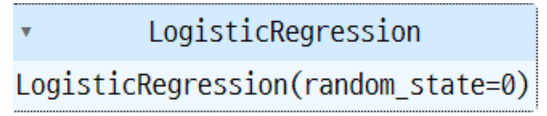


모델 사용 변수

Tf-idf 벡터화, stay_day(숙박 일수),
클러스터 label(GMM 모델), review_len(리뷰 길이)

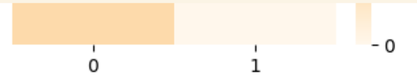
모델링

Logistic Regression 모델 사용



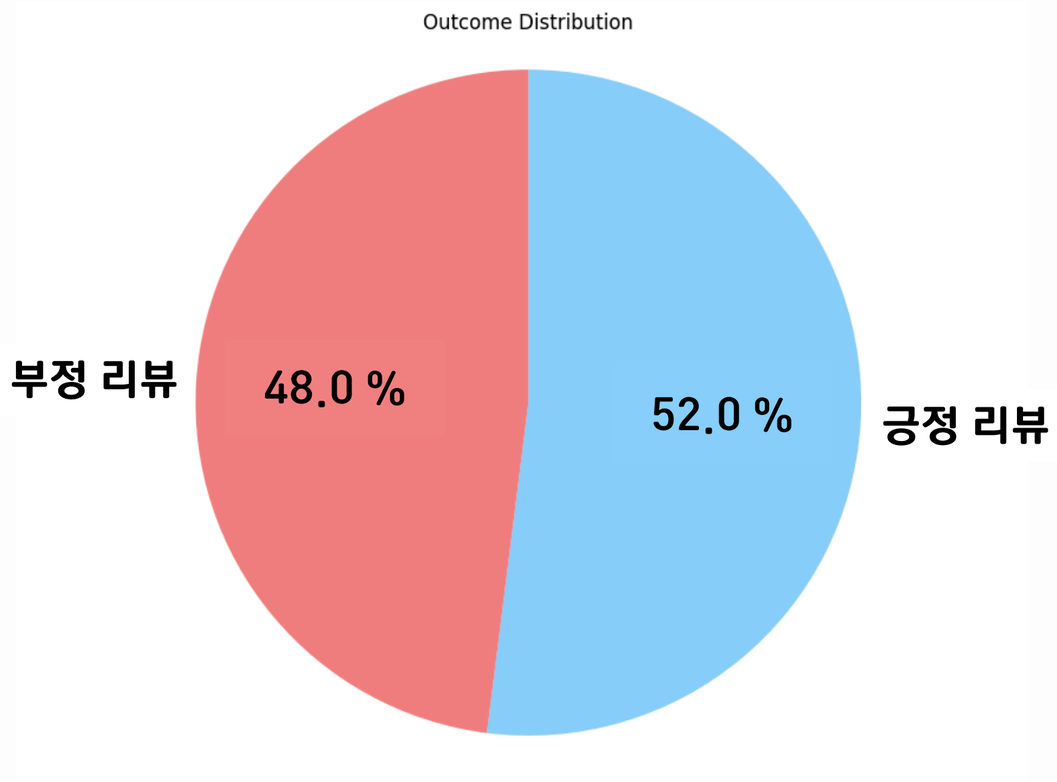
ac
pr
rec
F1: 0.00

예측 모델이 아예 분류를 하지 않았기 때문에
negative 리뷰의 수를 늘리는 방안을 택함



05 긍부정 분석

부정(리뷰 평점: 0 ~9점) - 1 / 긍정(리뷰 평점: 9 ~ 10점) - 0
분류 모델인 LogisticRegression 사용



모델 사용 변수

Tf-idf 벡터화(x, y), stay_day(숙박 일수)
여러 변수 조합 결과 성능이 가장 좋은 조합

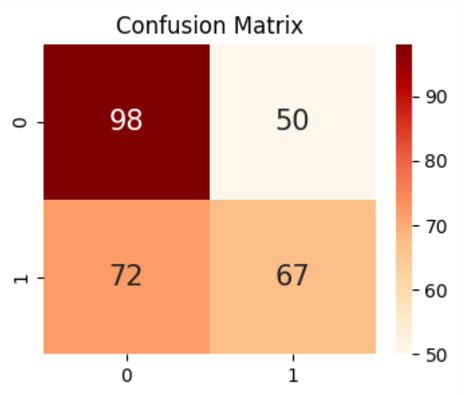
모델링

Logistic Regression 모델 사용

```
LogisticRegression
LogisticRegression(random_state=0)
```

모델 성능

accuracy: 0.57
precision: 0.57
recall: 0.48
F1: 0.52



05 긍부정 분석

부정(리뷰 평점: 0 ~9점) - 1 / 긍정(리뷰 평점: 9 ~ 10점) - 0
분류 모델인 LogisticRegression 사용

Feature 중요도

```
5] # Get feature importance
   feature_importance = np.abs(lr.coef_[0])

   # Normalize feature importance to sum up to 1
   feature_importance /= feature_importance.sum()

   # Print feature importance
   for i, importance in enumerate(feature_importance):
       print(f"Feature {i}: {importance}")
```

```
Feature 0: 0.5758554649183082
Feature 1: 0.26980214552338455
Feature 2: 0.15434238955830737
```

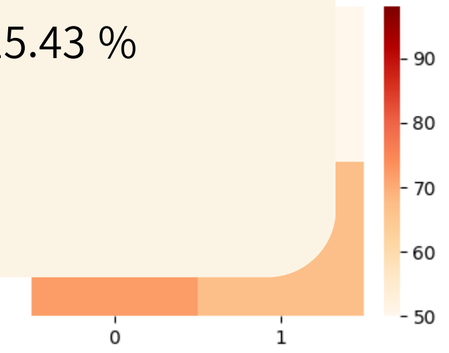
Tf-idf x 값 57.58 %

Tf-idf y 값 26.98 %

stay_day 변수 15.43 %

부정 리뷰

```
ession
ndom_state=0)
```

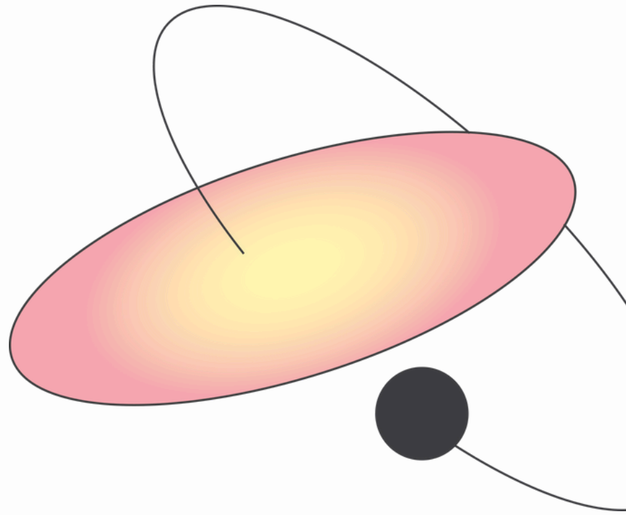


F1: 0.52

06 결론

의의 및 보완할 점

06 결론 및 의의



1. 데이터 수집
    a. Selenium을 사용하여 크롤링 진행
2. 데이터 전처리
    a. 영어와 숫자만 추출
    b. NLTK 토큰화
    c. Gensim 불용어 적용
3. EDA
    a. 계절별 wordcloud
    b. 호텔 리뷰 트렌드 분석
4. 모델링
    a. K-mean, GMM, DBSCAN 클러스터링
    b. 분류 모델을 통해서 사용자 적합 리뷰 추천
    c. 긍부정 분석 -> 리뷰 평점에 준 요인들 분석

06 보완할 점

1. 데이터 수 부족
    a. 총 900개의 호텔리뷰
        i. Overfitting 문제
        ii. 긍정 리뷰와 부정 리뷰 클래스 Imbalance
2. wordcloud
    a. 보다 의미 있는 계절별 키워드 추출 필요
3. 긍부정 분석에서 저조한 성능
    a. 더 다양한 데이터셋으로 성능 추후 개선