



제주도 호텔 리뷰 분석으로

당신에게 어울리는 특별한 하루를 선사합니다!



감성사전 기반 리뷰 분석을 통한 맞춤형 추천 시스템



박정환/한지은

목차

01

주제 선정 배경
데이터 분석개요

02

데이터 수집
데이터 정제
데이터 분석

03

데이터 모델링

04

추천 시스템
알고리즘

05

분석기대효과
개선방안

“지금 세계인은 한국 여행 중”

올해 1분기 여행 예약 현황을 분석한 결과,
지난해에 비해 호텔 예약은 월 최고 404%,

항공 예약은 월 최고 2862% 증가

출처: 트릿닷컴

떠나고 싶은 여행의 형태는?

호캉스 67.7%

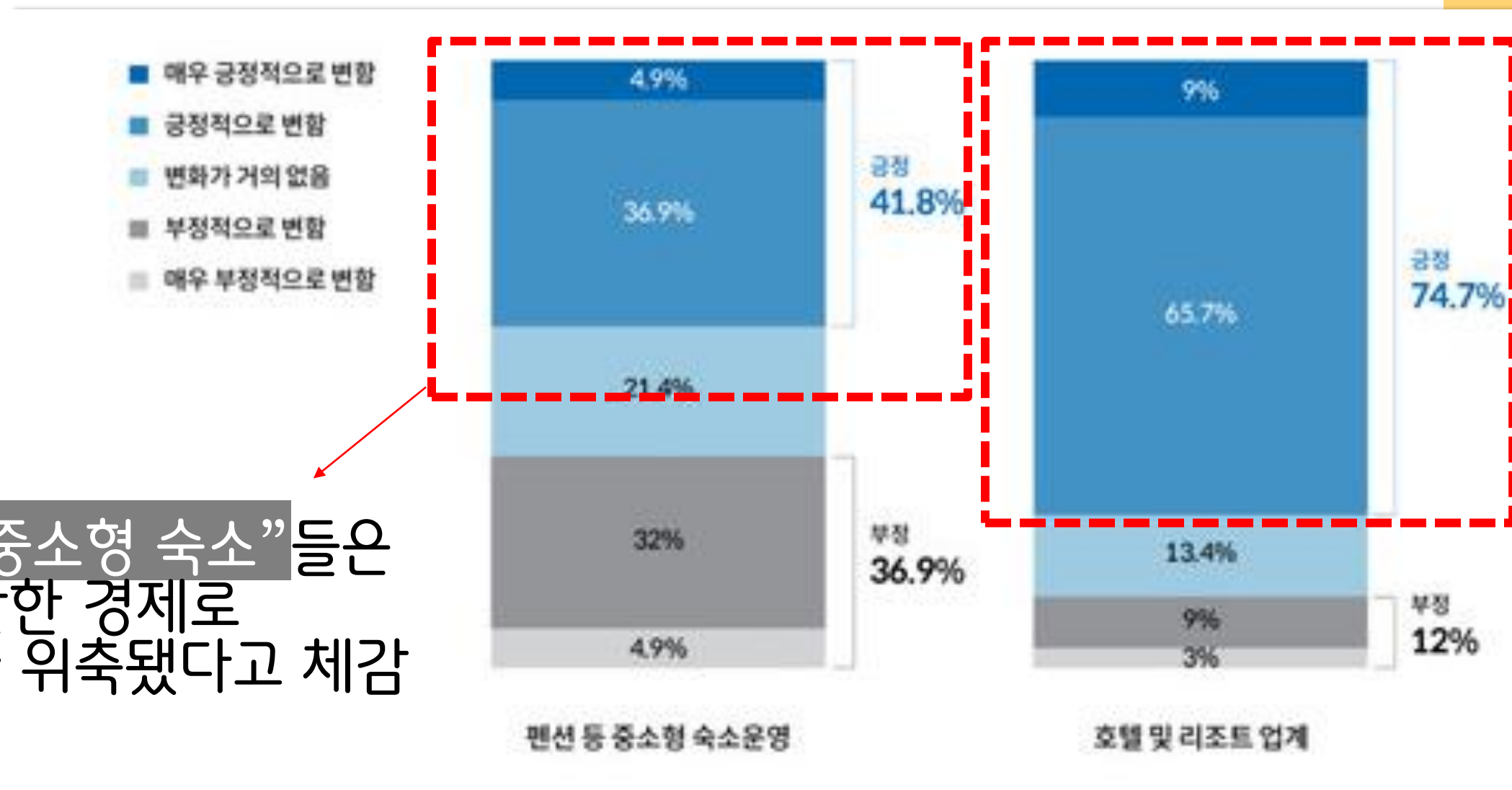
관광지 투어 54.8%

배낭 여행 50.0%

출처: 여기어때 설문조사

“중소형 숙소보다 호텔/리조트가 긍정적 영업 분위기”

“펜션 등 중소형 숙소”들은
불안한 경제로
여행소비가 위축됐다고 체감



반면,
“호텔 및 리조트 업계”는
보복여행 및 인바운드 여행 수요
증가로 긍정적 변화 인식

왜 제주도 인가?

1) 코로나 이후 가장 방문하고 싶은 도시

35%

제주

23%

부산

21%

인천

출처: 트립닷컴

2) 지역별 관광지 검색 순위 1위 도시

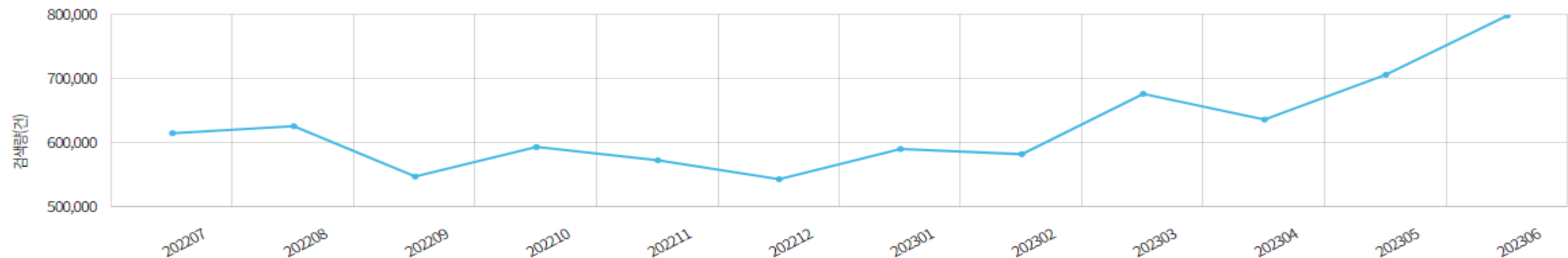
지역별 관광 검색 순위에 따르면
제주 공항이 46만 5천여 건으로 가장 많음

출처: 한국관광공사

3) 꾸준히 증가하는 관심도

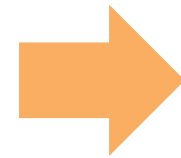
출처: 한국관광공사

SNS 언급량 ⓘ



주제 선정 배경

호텔 예약 객실의 대부분은 “온라인”으로 진행



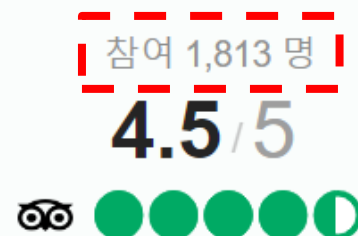
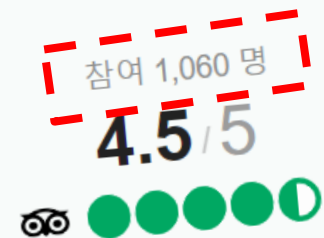
고객은 온라인 예약 시 리뷰와 평점 의존도가 높음

BUT!

1) 평점만으로 평가하기에는 어려우며,
일일이 읽어보며 판단하기에 **리뷰수가 많음**

2) 추천 호텔이더라도 **불만사항**이 존재

• Traveller's Choice 선정 호텔이더라도 불만 사항이 존재



●○○○○1

"5성급이라 불리기에 너무나 허술했던 호텔 서비스"

롯데 제주에서는 신혼여행으로 왔다고 미리 귀뜸을 드려도 전혀 미동하지 않더군요. 하물며 작은 카드 하나에 "happy honeymoon" 메세지 조차도 준비할 수 없었던 호텔측에 심히 실망했습니다. 그나마 F&B에 계신 분들은 상당히 상냥하시고 서비스도 좋았습니다만 호텔측에 대한 실망은 감히 어떤 단어로도 형용이 되지 않네요. 시설도 너무나 낙후되어 있고, 녹슨 발코니 방충망이며, 납작해져버린 카페트, 그 사이에 끼있는 머리카락들, 오랜기간 사용된 뻗뻗한 배스룸과 타월까지 하나부터 열까지 맘에 드는게 하나도 없었습니다. 아주 작은 제스처로도 큰 감동을 줄 수 있었던 호텔측의 방치가 애석하기 짝이 없네요. 절대 재방문, 친구추천 없습니다. 2023년에 이런 시설이 5성급이라는게 믿겨지지 않습니다. 돈 쓸 맛 안나요.

2023-01-12 | 커플여행 | nay

소비자가 원하는 호텔 속성에 대한 정보를 제공

+

Travellers' Choice 를 보조하는 수단 제공

데이터 수집

1) 호텔 필터링


- '제주도 호텔' 검색 상위 30개 중
- Travelers' Choice 호텔만 선정
- 총 11개의 호텔


2) 웹 크롤링

- 트립어드바이저와 부킹닷컴의 리뷰
- 한국어 리뷰
- 호텔명, 평점, 리뷰 제목, 리뷰 내용 등


3) 데이터 프레임 확인


- 11개 호텔의 리뷰 크롤링
- 10723개의 리뷰 데이터 확인






롯데 호텔 제주
5성급 | 서귀포, 제주도
★ 8.8 | 수영장이 매우 좋음
250,222원~


 1% 적립 호텔 직접 운영 객실패키지 보기 >




라마다
4성급 | 서귀포, 제주도
★ 8.1
65,596원~





파르나스 호텔 제주
5성급 | 서귀포, 제주도
★ 9.2
271,033원~

 1% 적립 호텔 직접 운영 객실패키지 보기 >





베스트웨스턴 제주호텔
4성급 | 제주, 제주도
★ 7.6 | 객실상태 깨끗함
56,868원~

 1% 적립 호텔 직접 운영 객실패키지 보기 >



랜딩관 제주신화월드 호...
5성급 | 서귀포, 제주도
★ 8.3
105,880원~

 1% 적립 호텔 직접 운영 객실패키지 보기 >



메종 글래드 제주
5성급 | 제주, 제주도
★ 8.1 | 직원들이 친절함
109,503원~

	name	rate	detail_p	detail_n
0	롯데 호텔 제주	4.5	침대보 매일 교체 등 매우 청결. 수영장 및 편의시설 등 매우 우수	높은 호텔 이용료. 원하지 않는 콜택시 음식점 호객행위 매우 불쾌. 맛 없음
1	롯데 호텔 제주	4.5	중문이어서 근처에 볼 것도 많았고 좋았습니다. 서비스도 좋아서 편안히 있다가 갔습니다.	별로 없었던 것 같습니다.
2	롯데 호텔 제주	4.5	- 접근성_x000D_₩n중문관광단지 내 위치하여 여기저기 접근이 용이함_x000D...	- 룸 컨디션_x000D_₩n침대가 너무 높아서 아이 있는 가족은_x000D_₩n침...
3	롯데 호텔 제주	4.5	기대이상으로 멋진 뷰에₩n넓은 룸에 ₩n너무너무 만족스러워요^^	없었어요^^*
4	롯데 호텔 제주	4.5	조식 석식 다 맛있고 좋아요	없습
...
10718	더 그랜드 섬오름	4.5	NaN	NaN
10719	더 그랜드 섬오름	4.5	NaN	NaN
10720	더 그랜드 섬오름	4.5	NaN	NaN
10721	더 그랜드 섬오름	4.5	NaN	NaN
10722	더 그랜드 섬오름	4.5	NaN	NaN

10723 rows x 4 columns

데이터 분석 배경

EDA	감성 사전 구축	속성 분석	추천 시스템
<div>데이터 수집<ul style="list-style-type: none">· 웹 스크래핑</div> <div>전처리<ul style="list-style-type: none">· 불용어 제거· 단어 토큰화 등</div> <div>EDA<ul style="list-style-type: none">· 리뷰 당 형태소 수· 워드 클라우드 등</div>	<div>로지스틱 회귀<ul style="list-style-type: none">· 긍/부정 단어 분류</div> <div>LDA K-Mens<ul style="list-style-type: none">· 감성사전 카테고리 분류· 카테고리 별 단어 선정</div>	<div>나이브 베이즈<ul style="list-style-type: none">· $P(\text{단어} 긍정)$, $P(\text{단어} 부정)$· 확률 값을 감성사전에 매칭</div> <div>평가함수 구축<ul style="list-style-type: none">· 제주 호텔 속성별 평가</div>	<div><ul style="list-style-type: none">· 고객 우선순위에 따라 가중치 적용하여 호텔 추천</div> <div></div> <div>“고객 맞춤형 추천서비스 제공”</div>

전처리

- ① 특수 문자 제거 : (, [, !, ♥ ...
- ② 이모티콘 제거 : 😊, 😍, ♥ ...
- ③ 자음, 모음 제거 : ㄱ, ㄴ, ㄷ...

before

수영하면서 공연도 너무 좋았고 입구에서 부터
친절하게 안내해주신 모든 분들 덕분에
좋았습니다! 너무 좋음 아이들과 부모님 모두
만족스러운 여행이었어요! 최고! 감사합니다!
ㅎㅎ ♥♥♥ 😊

after

수영하면서 공연도 너무 좋았고 입구에서 부터
친절하게 안내해주신 모든 분들 덕분에
좋았습니다 너무 좋음 아이들과 부모님 모두
만족스러운 여행이었어요 최고 감사합니다

형태소 분석

데이터 토큰화

텍스트 데이터를 하나의 특정 기본 단위로 자르는 것
KoNLPy의 Kkma, Okt 와 Mecab 라이브러리를 시도

비교적 빠르고. 품사 태그가 많은
“Mecab” 형태소 분석기 사용

프론트나 레스토랑 수영장 호텔 내 모든
직원분들이 다 친절하셨습니다

('프론트', 'NNP'), ('나', 'JC'), ('레스토랑', 'NNG'),
('수영장', 'NNG'), ('호텔', 'NNG'), ('내', 'NNB'),
('모든', 'MM'), ('직원', 'NNG'), ('분', 'XSN'), ('들',
'XSN'), ('이', 'JKS'), ('다', 'MAG'), ('친절', 'NNG'),
('하', 'XSA'), ('셨', 'EP+EP'), ('어요', 'EF')

[illegible]

유의미한 단어들 추출

전처리 후, 총 고유 형태소 수



청와대 국민청원 말뭉치

6,477개 VS 15,310개

전처리 후, 리뷰 당 평균 형태소 수

약 56개 VS 약 154개

부킹닷컴 리뷰 추가 결정

로지스틱 회귀

: 범주형 변수를 예측하는 모델

호텔 속성을 카테고리화 하여 감성 사전 구축

긍정 감성 사전
인적 서비스 = ["친절", "직원", "...]

부정 감성사전
불청결성 = ["냄새", "더럽", "...]

긍정 부정 단어 분류

로지스틱 회귀를 통하여

추정된 "회귀 계수"로

금/부정 단어를 분류

긍정

트립어드바이저 : 평점기준 4~5점
부킹닷컴 : 손가락 이모티콘기준 👍

1값 부여

부정

트립어드바이저 : 평점기준 1~3점
부킹닷컴 : 손가락 이모티콘기준 👎

-1값 부여

회귀 계수 '0' 기준

회귀계수가 0보다 큰 단어 중
상위 400개 단어
긍정 단어 질 분류

긍정 단어	
word	coef
친절	3.592519
위치	2.915053
깨끗	2.746333
여행	2.615250
깔끔	2.496709
...	...
특가	0.114228
아름답	0.113045
박일동	0.113027
커피	0.112197
마시	0.112125
400 rows × 1 columns	

부정 단어	
word	coef
소리	-1.716323
불편	-1.523443
주차장	-1.492967
방음	-1.487046
냄새	-1.400883
...	...
고치	-0.088328
미숙	-0.087957
모양	-0.087551
진동	-0.087462
악취	-0.087337
400 rows × 1 columns	

회귀계수가 0보다 작은 단어 중 절대값
기준상위 400개 단어
부정 단어 질 분류

데이터 모델링

LDA

: 문서들은 토픽들의 혼합으로 구성되어져 있으며, 토픽들은 확률 분포에 기반하여 단어들을 생성한다고 가정. LDA는 문서가 생성되던 과정을 역추적 토픽을 추출

**토픽의 수는 10개 내외에서
응집도가 가장 높은 수를 선택함**

① gensim

긍정 단어

(0, '0.211*"직원" + 0.181*"친절" + 0.056*"서비스" + 0.030*"청결" + 0.029*"깨끗" + 0.022*"시설" + 0.022*"기분" + 0.022*"깔끔" + 0.021*"응대" + 0.021*"체크인"')
(1, '0.115*"여행" + 0.061*"가족" + 0.041*"방문" + 0.038*"이용" + 0.030*"워터" + 0.029*"파크" + 0.027*"만족" + 0.024*"숙박" + 0.020*"예약" + 0.020*"숙소"')
(2, '0.205*"조식" + 0.065*"맛있" + 0.059*"아침" + 0.056*"바다" + 0.049*"식사" + 0.046*"음식" + 0.030*"전망" + 0.024*"뷰페" + 0.021*"산책" + 0.019*"시장"')
(3, '0.139*"위치" + 0.090*"침대" + 0.076*"공항" + 0.044*"깨끗" + 0.040*"숙소" + 0.037*"근처" + 0.037*"성비" + 0.034*"깔끔" + 0.027*"시내" + 0.026*"침구"')
(4, '0.116*"이용" + 0.116*"시설" + 0.083*"가격" + 0.072*"수영장" + 0.055*"만족" + 0.038*"생각" + 0.033*"대비" + 0.027*"전반" + 0.023*"부대시설" + 0.020*"서비스"')
(5, '0.260*"객실" + 0.067*"예약" + 0.036*"주변" + 0.036*"거리" + 0.035*"편의점" + 0.030*"식당" + 0.026*"내부" + 0.025*"하루" + 0.019*"해결" + 0.017*"이동"')

부정 단어

(0, '0.063*"불편" + 0.052*"주차" + 0.049*"주차장" + 0.030*"체크" + 0.028*"부족" + 0.027*"사용" + 0.025*"필요" + 0.024*"공간" + 0.019*"프론트" + 0.017*"요청"')
(1, '0.044*"청소" + 0.042*"소리" + 0.041*"화장실" + 0.034*"느낌" + 0.033*"방음" + 0.028*"냄새" + 0.027*"사람" + 0.025*"소음" + 0.024*"오래" + 0.020*"아쉽"')

토픽의 수 : 10개

② sklearn

긍정 단어

Topic 1: [('위치', 380.22), ('주변', 175.24), ('공항', 155.57), ('편의점', 138.21), ('성비', 114.24), ('거리', 90.26), ('시내', 89.33), ('근처', 88.9), ('이동', 75.4), ('편리', 55.65)]
Topic 2: [('아침', 193.87), ('식사', 131.48), ('예약', 111.02), ('저녁', 72.05), ('테라스', 54.25), ('준비', 43.85), ('도착', 38.95), ('조식', 28.99), ('쾌적', 27.72), ('일정', 27.05)]
Topic 3: [('이용', 147.4), ('워터', 72.08), ('파크', 71.35), ('응대', 69.14), ('수영장', 68.64), ('체크아웃', 67.66), ('야외', 56.03), ('수영', 50.36), ('나가', 47.62), ('가족', 43.88)]
Topic 4: [('직원', 324.84), ('친절', 197.93), ('숙소', 139.75), ('서비스', 92.26), ('로비', 90.49), ('만족', 85.98), ('깔끔', 70.26), ('컨디션', 61.69), ('전반', 52.33), ('방문', 38.93)]
Topic 5: [('청결', 270.24), ('바다', 122.22), ('침구', 108.6), ('전망', 86.14), ('편안', 81.76), ('객실', 50.47), ('위치', 46.32), ('매일', 42.26), ('분위기', 41.27), ('마음', 40.81)]
Topic 6: [('가격', 276.11), ('대비', 165.55), ('관찰', 124.09), ('조식', 82.92), ('조용', 62.13), ('저렴', 50.65), ('아기', 49.57), ('객실', 43.38), ('만족', 41.35), ('할인', 36.04)]
Topic 7: [('가성', 103.04), ('체크인', 82.62), ('처음', 55.29), ('커피', 51.84), ('부대시설', 47.48), ('무료', 44.21), ('세요', 38.18), ('업그레이드', 34.37), ('이용', 31.99), ('객실', 24.8)]
Topic 8: [('조식', 479.51), ('수영장', 256.01), ('깨끗', 113.38), ('음식', 106.9), ('맛있', 106.82), ('전체', 81.2), ('이용', 76.49), ('뷰페', 68.46), ('아주', 68.11), ('사우나', 52.3)]
Topic 9: [('침대', 481.57), ('객실', 375.64), ('생각', 114.39), ('식당', 50.99), ('오션', 39.34), ('분리', 31.54), ('욕실', 29.4), ('장점', 29.27), ('다양', 29.01), ('키즈', 28.84)]
Topic 10: [('시설', 343.2), ('성급', 81.85), ('내부', 72.22), ('숙박', 52.59), ('가능', 52.11), ('이용', 37.86), ('영장', 37.55), ('생각', 35.7), ('편의', 35.35), ('서비스', 30.58)]

부정 단어

Topic 1: [('별로', 205.2), ('필요', 136.7), ('화장실', 128.9), ('코로나', 100.98), ('룸서비스', 89.41), ('신경', 81.61), ('그대로', 57.66), ('도로', 43.72), ('마스크', 37.96), ('아무래도', 35.27)]
Topic 2: [('냄새', 242.34), ('느낌', 239.41), ('노후', 118.26), ('샤워', 98.64), ('칫솔', 88.03), ('에어컨', 83.61), ('테이블', 83.02), ('화장실', 68.91), ('치약', 65.55), ('담배', 51.79)]
Topic 3: [('소리', 322.98), ('오래', 224.08), ('아쉽', 186.28), ('욕실', 162.6), ('딱히', 157.53), ('바닥', 153.74), ('먼지', 143.4), ('방음', 129.91), ('욕조', 120.26), ('옆방', 110.2)]
Topic 4: [('주차', 426.29), ('부족', 252.43), ('공간', 199.72), ('주차장', 189.43), ('협소', 129.13), ('아쉬움', 125.32), ('어메니티', 102.9), ('살짝', 78.58), ('불편', 72.57), ('모르', 57.51)]
Topic 5: [('불편', 314.07), ('사용', 162.23), ('수압', 85.79), ('샤워기', 57.48), ('비싸', 48.98), ('와이파이', 35.05), ('성수기', 33.53), ('잘못', 32.12), ('니스', 31.91), ('피트', 27.69)]
Topic 6: [('체크', 138.85), ('프론트', 111.68), ('종류', 89.14), ('변기', 53.17), ('들어가', 50.88), ('입구', 44.37), ('복잡', 41.65), ('최악', 40.84), ('확인', 37.43), ('배치', 31.15)]
Topic 7: [('주차장', 261.94), ('방음', 239.02), ('소음', 212.06), ('공사', 75.54), ('투숙객', 71.8), ('조명', 61.79), ('그릴', 49.81), ('비치', 47.51), ('삼푸', 42.02), ('모기', 40.57)]
Topic 8: [('청소', 215.42), ('사람', 134.27), ('건물', 128.83), ('연결', 75.88), ('안내', 74.74), ('특별히', 64.98), ('외부', 63.32), ('따로', 63.14), ('운영', 62.75), ('동선', 56.11)]
Topic 9: [('힘들', 107.12), ('개선', 97.03), ('엘리베이터', 73.74), ('문제', 73.17), ('요청', 69.48), ('고객', 64.62), ('온도', 64.44), ('굳이', 57.75), ('불친절', 56.08), ('수준', 54.43)]
Topic 10: [('메뉴', 120.94), ('수건', 96.05), ('추가', 69.92), ('세면대', 54.39), ('머리카락', 53.05), ('실망', 52.68), ('전화', 50.41), ('당황', 49.04), ('매트리스', 47.23), ('오픈', 45.85)]

Gensim의 부정 단어들은 잘 분류되는 듯 보이나, 나머지 형태소들은 잘 분류되지 않음

2차 카테고리 선정 시도

Fasttext

: n-gram 모델을 사용하여 언어의 형태학적 구조를 반영한 임베딩 모델

재미있다	<ㅈㅍ	ㅈㅍ_	ㅍ_ㅁ	_ㅁㅣ	ㅁㅣ_	ㅣ_ㅇ	_ㅇㅣ	ㅇㅣㅅ	ㅣㅅㅁ	ㅁㅅㅈ	ㅈㅈ_	ㅈ>
재미있다	<ㅈㅍ	ㅈㅍ_	ㅍ_ㅁ	_ㅁㅣ	ㅁㅣ_	ㅣ_ㅇ	_ㅇㅣ	ㅇㅣㅅ	ㅣㅅㅁ	ㅁㅅㅈ	ㅈㅈ_	ㅈ>
잼있다	<ㅈㅍ	ㅈㅍㅁ	ㅍㅁㅇ				ㅁㅇㅣ	ㅇㅣㅅ	ㅣㅅㅁ	ㅁㅅㅈ	ㅈㅈ_	ㅈ>

[그림 6] 재미있다와 재미있다, 잼있다의 3-gram subword

‘재미있다’, ‘잼있다’의 경우, ‘재미있다’와 동일한 단어 형태가 아님에도 불구하고 자모 단위로 분리했을 경우 겹치는 subword가 많기에 학습과정에 나타나지 않아도 의미를 추론할 수 있음
효과적으로 임베딩이 가능

K-Means

군집 결과 특성을 추출하여
카테고리화

* 축소한 후와 축소하지 않은 뒤 군집분석
결과가 비슷하게 나왔음, 축소한 뒤 진행

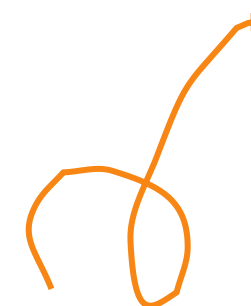
데이터 모델링

K-Means

Fasstext로 100차원에 임베딩 된 값을 t-sne를 활용하여, 2차원으로 축소

t-sne

높은 차원의 복잡한 데이터를 2차원에 차원 축소하는 방법 중 하나



	0	1	2	3	4	5	6	7	8	9	...	90	91	92	93	94	95	96	97	98	99
word_pd																					
친절	0.412975	0.163120	-0.223239	0.066492	-0.064201	0.198817	-0.323929	-0.024202	0.238423	0.091111	...	0.005331	-0.357637	0.021944	0.054601	0.107391	-0.169066	-0.091920	0.467110	-0.232517	-0.171546
위치	0.279903	-0.087060	-0.077252	0.053126	0.043666	0.016810	-0.373488	0.041704	-0.005134	-0.206486	...	-0.268356	0.153645	0.017890	-0.061612	-0.185415	-0.072886	-0.139744	0.414140	0.005186	-0.331578
깨끗	0.317015	0.002039	-0.231544	0.160634	-0.125178	0.279845	0.072771	0.044835	-0.111417	-0.033298	...	-0.017997	-0.128959	0.199624	0.238615	-0.160689	-0.174266	-0.286479	0.303363	-0.061260	-0.117981
여행	0.210757	-0.160539	-0.020765	0.049103	-0.473084	0.201031	-0.022779	-0.019959	0.169826	-0.010401	...	-0.292089	0.083591	0.034250	-0.095327	0.086768	-0.093218	0.021369	0.407927	-0.136184	-0.002916
깔끔	0.315674	0.142510	-0.265222	0.176078	-0.098717	0.329886	-0.098867	0.003970	-0.031321	0.075141	...	-0.081685	-0.127140	0.117239	-0.020281	0.224508	-0.328746	-0.315816	0.479009	-0.266403	-0.317797
...
특가	0.101822	0.560838	0.125468	0.423492	-0.209200	-0.251334	0.617375	-0.284378	0.518242	-0.683028	...	-0.173581	-0.589339	0.826544	-0.465194	-0.543840	0.704674	-1.760975	1.274195	-0.287777	-0.018855
아름답	0.638452	0.291839	-0.142175	-0.224074	-0.390814	0.386794	0.309388	-0.124255	1.056396	-0.424924	...	-0.209788	-0.572671	0.829462	-0.209843	-0.614742	-0.809706	0.774021	0.251472	0.191426	-0.812770
박일동	-0.310653	0.995964	-0.608055	-0.855739	-0.299477	0.066967	-0.206548	0.096283	0.189993	-0.402724	...	-0.264043	-0.165298	0.393771	0.037345	-0.493009	-0.497254	0.099733	0.562390	-0.276476	-0.949150
커피	0.627640	0.108316	-0.155869	0.609876	0.470559	0.227662	0.337127	-0.410303	0.022284	-0.627532	...	0.147230	-0.339829	-0.076411	-0.045578	-0.611533	-0.361905	-0.065302	0.498208	0.623748	-0.303857
마시	0.448112	0.710389	-0.089040	-0.299883	-0.033719	0.251515	-0.709340	-0.096308	-0.268366	-0.425255	...	-0.063454	0.495738	0.444990	0.594602	-0.419743	-0.146883	0.309292	0.766861	0.158570	-0.568967

400 rows × 100 columns

	x	y
word_pd		
쾌적	8.466201	-15.146723
최적	7.337820	-15.487930
서귀포	-7.688689	-40.262054
불꽃놀이	3.549763	-70.534874
갑자기	1.109239	1.467487
도움	9.196942	-27.925711
운행	20.187927	0.675724
전하	2.551301	-27.209047
착하	53.956436	-46.316475
서쪽	37.888916	-19.720255
식음료	-5.406144	-25.602116
타입	25.599373	-31.573040
오설록	43.702534	-28.627272
인프라	5.823172	-57.174946
제주시	44.655174	-16.451084
유용	10.584465	-19.846632
신나	47.966507	-50.645153
발렛	17.760536	-22.312078
서울	15.684736	-39.568932

데이터 모델링

K-Means

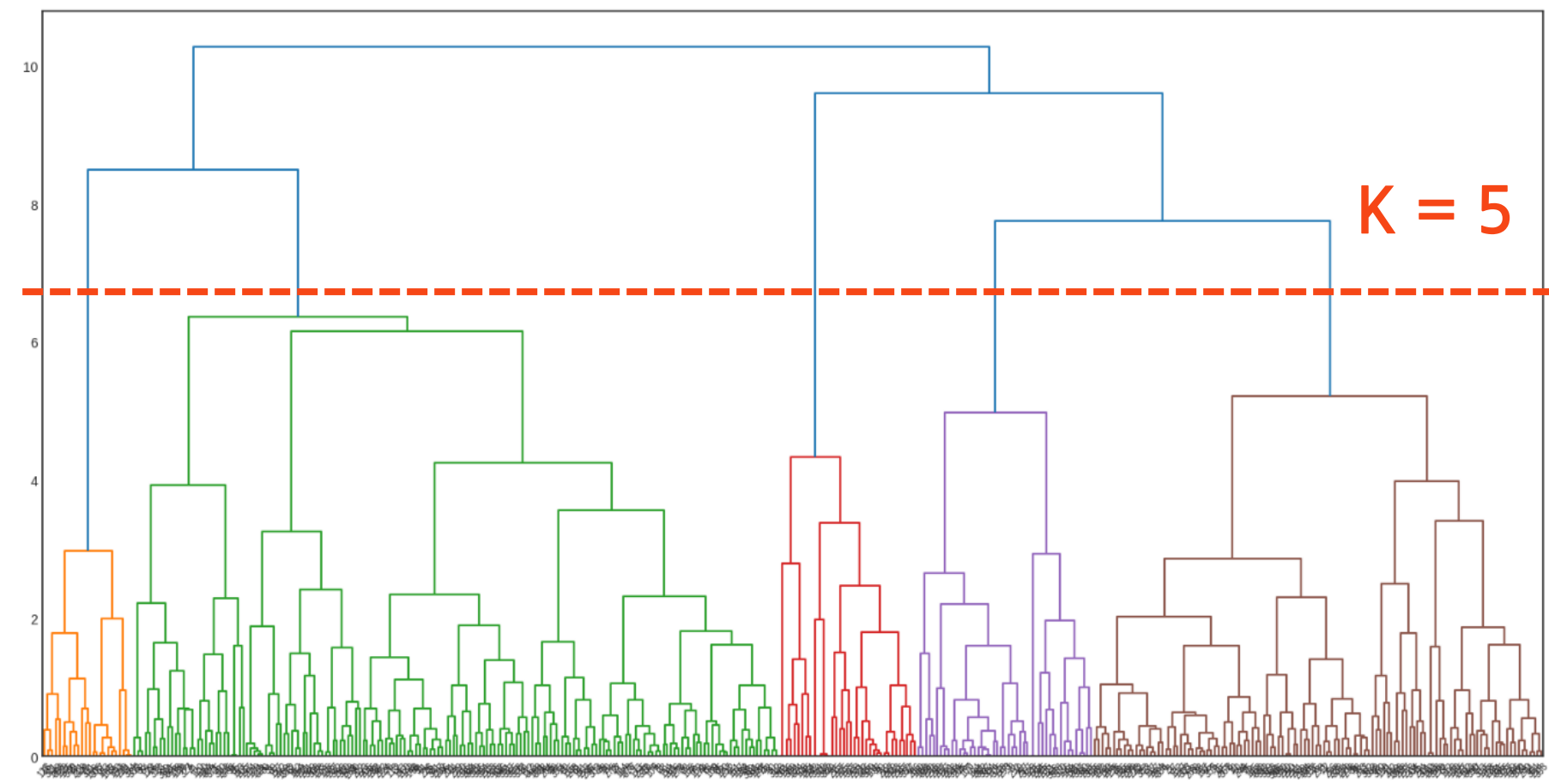
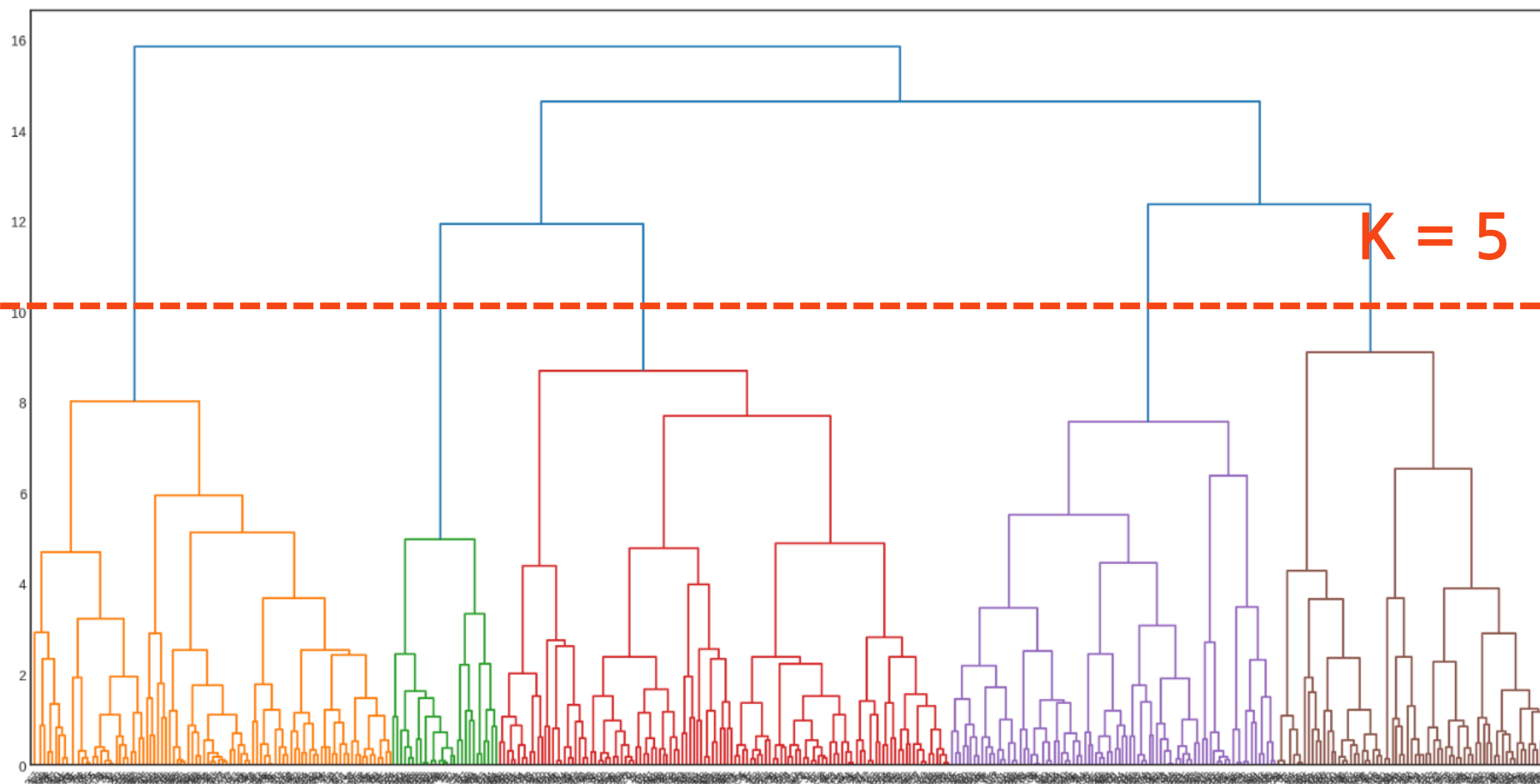
덴드로그램

각 단계에서 관측치의 군집화를 통해 형성된 그룹과 이들의 유사성 수준을 표시하는 트리 다이어그램

최장 연결법 (Complete linkage method) 사용

긍정 단어

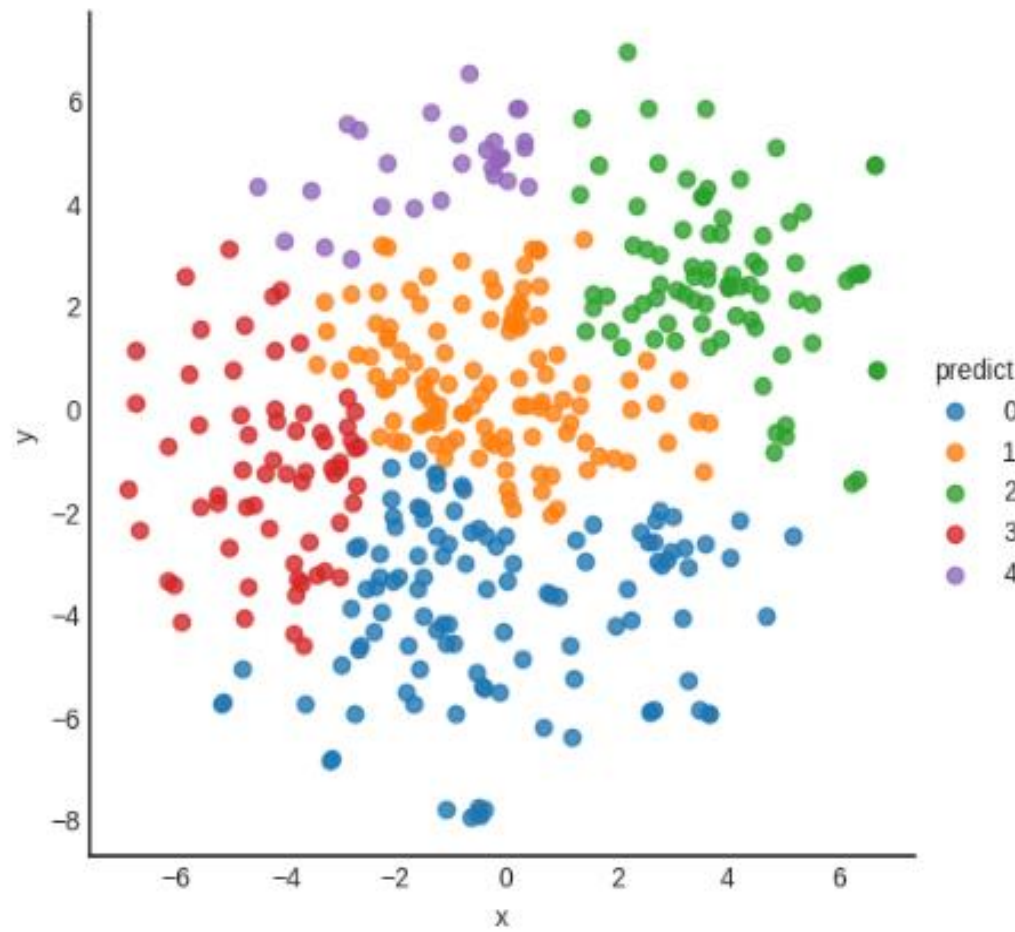
부정 단어



데이터 모델링

K-Means

긍정 단어



0군집 : 111개, 1군집 : 116개

2군집 : 80개, 3군집 : 65개

4군집 : 28개

0군집

	x	y	predict
word_pd			
여행	-2.288728	-3.270089	0
최고	-1.284962	-1.263407	0
가족	-1.504519	-4.036815	0
추천	-0.539081	-5.164546	0
숙소	-1.259033	-1.479066	0
업그레이드	0.119766	-3.037092	0
수영장	2.571931	-2.617976	0
테마파크	2.671039	-5.880286	0
영장	2.715649	-2.200134	0
투숙	-1.623682	-2.870042	0
따뜻	2.749801	-2.990728	0
친구	-2.647427	-4.620243	0
워터	2.593987	-5.894961	0
지내	-2.111364	-1.778379	0
파크	2.587938	-5.903023	0
숙박	-1.277033	-2.502567	0
부모	-3.155648	-6.813128	0
강추	0.292771	-4.898151	0
걱정	-1.308960	-1.278727	0
즐기	1.256158	-2.568225	0
처음	-2.050274	-2.094370	0
키즈	-0.497048	-7.766230	0
정원	4.044734	-2.918504	0
박일	-2.046193	-3.346092	0
계획	-2.814451	-3.884204	0

#동반여행 시, 만족도

1군집

	x	y	predict
word_pd			
깨끗	-1.360770	0.139584	1
깔끔	-1.463939	0.301764	1
만족	-1.311623	-0.715299	1
편안	-2.382185	0.639015	1
맛있	0.073158	1.701272	1
아주	-1.500057	-0.301770	1
편하	-1.591521	-0.337670	1
이용	-0.002459	-1.591743	1
가성	3.460317	-0.229213	1
조용	-1.591634	2.048918	1
시설	-0.401328	-0.342610	1
괜찮	-0.462350	0.279854	1
전반	-0.835280	-0.099423	1
서비스	-1.906140	-0.656900	1
컨디션	-1.253967	-0.145893	1
훌륭	0.210536	0.572823	1
성비	1.019827	0.160497	1
청결	-1.223168	0.062120	1
식사	0.241645	1.997161	1
침구	-1.932738	1.112863	1
저렴	1.433324	-0.662864	1
부대시설	-0.216976	-0.691076	1
해결	0.587492	-1.260664	1
조식	0.037709	1.489552	1
분위기	-2.176599	3.129666	1
가능	0.635173	-1.623827	1

#시설 만족도

2군집

	x	y	predict
word_pd			
위치	3.385018	2.110520	2
공항	4.218362	2.376433	2
편리	2.903400	1.670448	2
시내	4.010431	2.322418	2
근처	3.638706	2.713635	2
가깝	4.013386	2.380793	2
산책	4.864521	-0.456161	2
접근성	4.021554	2.409978	2
시장	3.534722	4.107871	2
맛집	3.353633	2.763039	2
거리	2.779164	2.972862	2
가까워	4.085110	2.614881	2
주변	3.084072	2.287308	2
다양	1.415323	1.504159	2
스타	6.709372	0.745000	2
박스	6.714360	0.744681	2
동문	3.551806	4.127768	2
버스	6.277709	2.592660	2
편의	2.656987	1.354365	2
교통	3.655617	1.215061	2
산책로	4.835573	-0.865091	2
중문	5.129155	3.646678	2
바닷가	2.747350	4.787230	2
음식점	2.444478	2.016921	2
편의점	2.695895	2.164220	2

#접근 편의성

3군집

	x	y	predict
word_pd			
친절	-2.681353	-0.740438	3
직원	-2.760523	-0.763442	3
방문	-3.315636	-3.182171	3
감사	-4.365000	-1.284634	3
마음	-2.760245	-0.036926	3
쾌적	-2.883330	0.211104	3
항상	-2.782957	-1.843780	3
덕분	-4.580976	-1.870628	3
감동	-4.684335	-0.501186	3
추억	-6.043048	-3.424189	3
행복	-5.238870	-1.832146	3
기분	-3.639813	-1.226986	3
응대	-3.306162	-0.621948	3
보내	-3.858363	-4.375122	3
편히	-6.713282	0.088822	3
즐겁	-5.558212	-1.932434	3
예정	-3.733296	-3.418829	3
기회	-3.454740	-3.260369	3
도착	-2.707917	-1.504339	3
체크인	-3.023739	-1.145589	3
친절히	-4.182400	-0.252327	3
기억	-4.303791	-2.341367	3
너무나	-3.667900	-0.100376	3
휴식	-5.560843	1.556033	3
고민	-6.856137	-1.561547	3
충분히	-3.746559	1.286073	3

#인적 서비스성

4군집

	x	y	predict
word_pd			
바다	-0.138757	4.850344	4
오션	-0.099996	4.901392	4
고급	-2.812271	2.913676	4
경치	-0.379859	5.052758	4
힐링	-2.670061	5.404701	4
전망	-0.179669	4.843182	4
풍경	-0.809103	4.766814	4
한라산	0.213855	5.829239	4
범섬	-0.236283	5.199297	4
야경	-1.697012	3.908620	4
바라보	0.017945	4.435916	4
라산	0.170773	5.821544	4
무척	-4.021578	3.235060	4
대박	-1.361416	5.747177	4
테라스	-0.235531	4.538319	4
일출	0.398639	4.327383	4
전경	0.326763	5.090323	4
멋지	-0.875814	5.333804	4
경관	-0.668544	6.539497	4
넓적	-1.196741	4.043449	4
끝내	0.318047	5.172130	4
소소	-2.877985	5.517834	4
외국	-2.166097	4.782688	4
부담	-3.547580	4.235342	4
지루	-4.525369	4.319495	4
감상	-0.278991	4.685592	4

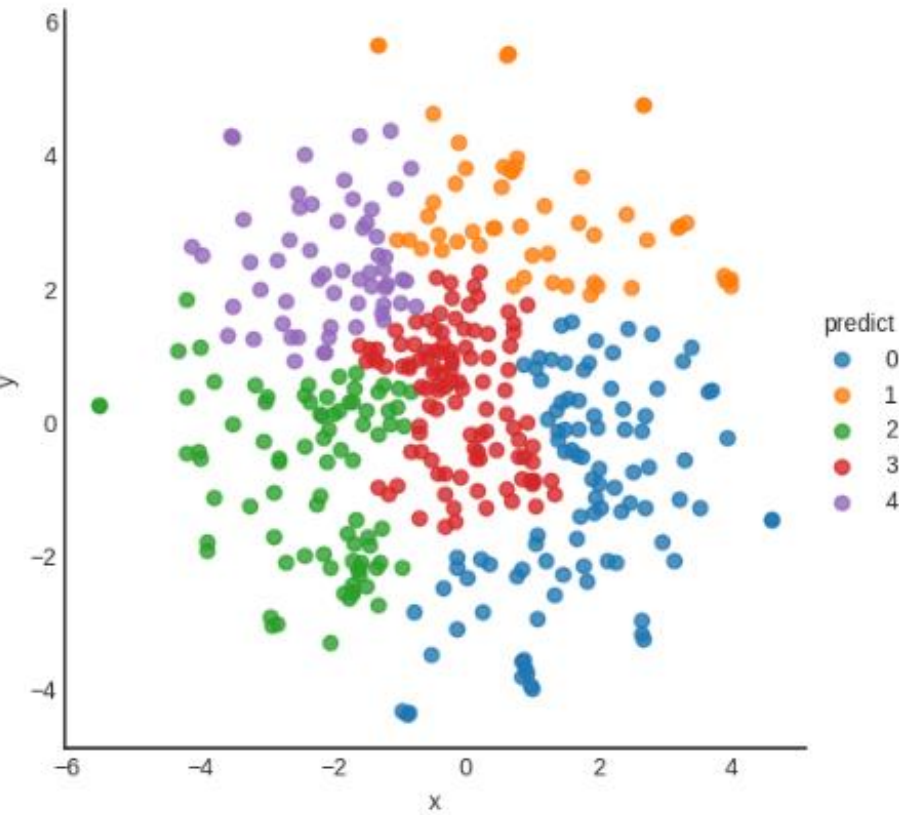
#주변 경관

LDA에 비해 잘 분류되어 나타남

데이터 모델링

K-Means

부정 단어



0군집 : 111개, 1군집 : 116개
2군집 : 80개, 3군집 : 65개
4군집 : 28개

0군집

	x	y	predict
word_pd			
냄새	1.053107	-1.834739	0
먼지	1.939811	-1.151850	0
수건	1.537370	0.328575	0
바닥	1.664043	-0.511863	0
창문	-0.131697	-2.050212	0
냉장고	0.244510	-2.858409	0
머리카락	1.572128	-0.441368	0
모기	3.117368	-2.107105	0
불쾌	1.078172	-1.726193	0
그대로	1.742373	-0.542014	0
샤워	1.222931	0.035723	0
시트	1.685270	-0.124043	0
벌레	-0.869274	-4.386503	0
난방	0.904406	-3.786297	0
벽지	1.720944	-1.440570	0
에어컨	0.869883	-3.576537	0
커버	1.483040	0.878958	0
곰팡	2.132654	-2.088724	0
카펫	1.994858	-0.705167	0
화장	2.649900	-3.199334	0
교체	1.693965	0.305477	0
담배	0.837011	-2.238472	0
곰팡이	1.756609	-2.167387	0
실문	2.665859	-3.265654	0
이불	2.186097	0.494772	0
흔적	4.605882	-1.485912	0

#불청결성

1군집

	x	y	predict
word_pd			
칫솔	3.185877	2.887169	1
치약	3.211904	2.921751	1
부실	-0.493374	4.605948	1
당황	-0.430253	2.785431	1
면도기	3.312655	2.975778	1
유료	0.702540	2.014704	1
비용	0.688820	3.739484	1
요금	0.653143	3.771211	1
지불	0.727046	3.816968	1
평범	1.746178	3.673774	1
삼푸	3.912282	2.103939	1
추가	0.530425	3.510203	1
확인	-0.689274	2.593521	1
숙박비	1.173087	3.222566	1
요구	-1.056483	2.719465	1
비누	3.874291	2.173360	1
생수	0.987597	2.494472	1
일회용품	2.713526	2.722020	1
싱글	1.955854	2.031013	1
린스	3.981254	2.130025	1
위험	1.690204	2.963051	1
건조	0.869795	2.164436	1
불가	0.084878	2.857823	1
음료	-0.124519	4.189207	1
사전	-0.503535	3.280504	1
금액	0.431544	2.897355	1

#어메니티 및
비용 불만족도

2군집

	x	y	predict
word_pd			
소리	-1.597753	-2.123939	2
주차장	-1.761706	0.289160	2
방음	-1.290084	-1.613215	2
부족	-1.030818	0.490467	2
소음	-1.481931	-1.741057	2
딱히	-0.850345	0.429319	2
주차	-1.682105	0.486712	2
협소	-1.915677	0.145060	2
공사	-1.661424	-1.493164	2
애매	-0.931789	-0.086188	2
옆방	-1.619573	-2.200340	2
힘들	-1.038571	0.218553	2
복도	-1.306821	-2.129083	2
작음	-4.355714	1.050884	2
그리	-1.161155	-0.046278	2
엘리베이터	-2.156593	-0.262804	2
복잡	-2.197616	0.094462	2
동선	-2.177718	0.070431	2
건물	-1.562917	-0.055961	2
새벽	-1.724393	-2.109784	2
밤새	-1.647762	-2.226793	2
도로	-2.249501	-1.253193	2
일부	-3.188302	0.537070	2
바람	-0.965054	-2.211166	2
이터	-5.527936	0.228886	2
자리	-2.383826	0.542397	2

#시설 불만족도

3군집

	x	y	predict
word_pd			
불편	-0.694735	0.525181	3
별로	-0.251719	0.759777	3
노후	0.175051	-0.448312	3
오래	0.055199	-0.204580	3
청소	0.924607	-0.030261	3
최악	-1.053068	-0.961549	3
아쉽	-0.382589	0.714717	3
개선	-0.432616	0.471404	3
화장실	0.872961	-0.534873	3
아쉬움	-0.596212	0.493848	3
사람	-0.979868	0.815293	3
욕실	0.979013	-0.381732	3
문제	-1.355484	1.088113	3
특별히	-0.127136	0.592634	3
이해	-0.210201	1.841550	3
모르	-0.393163	1.023903	3
살짝	-0.442141	0.188223	3
미흡	0.153862	-1.048299	3
수업	0.609695	-1.011693	3
모텔	-0.685619	-0.459442	3
세면대	0.985786	-0.610615	3
메뉴	-0.610322	1.048305	3
테이블	-0.104171	-0.835706	3
실망	0.063271	0.949646	3
가구	0.129278	-0.551583	3
느낌	-0.324214	0.512924	3

#객실 불만족도

4군집

	x	y	predict
word_pd			
불친절	-2.362140	2.551237	4
전화	-1.327523	2.477102	4
고객	-0.976319	2.133139	4
태도	-2.341747	3.252722	4
연결	-1.879943	2.248530	4
비추	-3.596183	1.278917	4
마스크	-3.520873	4.251216	4
상황	-1.002313	1.778187	4
이야기	-2.229538	2.119178	4
체크	-1.245805	1.505701	4
잘못	-2.144472	2.203893	4
죄송	-1.570474	2.907739	4
연락	-1.364923	2.778700	4
말투	-2.508543	3.203054	4
데스크	-1.435809	2.032063	4
프론트	-1.245627	1.750683	4
반복	-1.956214	2.999480	4
형편없	-1.138181	4.364532	4
황당	-1.422436	3.185087	4
전기	-3.974382	2.497224	4
카운터	-1.245758	2.269670	4
연박	-0.838139	3.802135	4
입실	-3.258430	2.386457	4
한참	-2.162829	1.031945	4
이러	-1.188076	2.059722	4
일행	-2.725469	1.797204	4

#불친절성

LDA에 비해 잘 분류되어 나타남

호텔 속성 감성사전 구축

긍정 단어 사전

동반여행 시, 만족도 = ['여행', '최고', '가족', '추천', '숙소', '투숙', '따뜻', '친구', '부모', '강추',...]

시설 만족도 = ['깨끗', '깔끔', '만족', '편안', '조용', '시설', '관찰', '전반', '서비스', '컨디션', '훌륭', '침구',...]

접근 편의성 = ['위치', '공항', '편리', '시내', '근처', '가깝', '산책', '접근성', '시장', '맛집', '거리', '가까워',...]

인적 서비스성 = ['친절', '직원', '방문', '감사', '마음', '항상', '덕분', '감동', '추억', '행복', '기분', '응대', '보내', '편히', '즐겁',...]

주변 경관 = ['바다', '오션', '고급', '경치', '힐링', '전망', '풍경', '한라산', '범섬', '야경', '바라보', '대박',...]

부정 단어 사전

불청결성 = ['냄새', '먼지', '수건', '바닥', '창문', '냉장고', '머리카락', '모기', '불쾌', '그대로', '샤워', '시트', '벌레', '난방',...]

어메니티 및 비용 불만족도 = ['칫솔', '치약', '부실', '당황', '면도기', '유료', '비용', '요금', '지불', '평범', '샴푸', '추가',...]

시설 불만족도 = ['소리', '주차장', '방음', '부족', '소음', '딱히', '주차', '협소', '공사', '애매', '옆방', '힘들', '복도',...]

객실 불만족도 = ['불편', '별로', '노후', '오래', '청소', '최악', '아쉽', '개선', '화장실', '아쉬움', '사람', '욕실', '문제',...]

불친절성 = ['불친절', '전화', '고객', '태도', '연결', '비추', '마스크', '상황', '이야기', '체크', '잘못', '죄송', '연락', '말투',...]

리스트(list)형태로 구축
“호텔 속성”이 감성사전의 카테고리

나이브베이즈

베이즈 정리에 따르면
사전확률 $P(\text{긍정})$, $P(\text{부정})$ 이 “**동일**”한 경우,

$$\begin{aligned} P(\text{단어}|\text{긍정}) &= P(\text{긍정}|\text{단어}) \\ P(\text{단어}|\text{부정}) &= P(\text{부정}|\text{단어}) \end{aligned}$$



$$\hat{P}(\text{긍정}|\text{단어}) = P(\text{단어}|\text{긍정})$$

$$\hat{P}(\text{부정}|\text{단어}) = P(\text{단어}|\text{부정})$$

데이터 모델링

나이브베이지스

나이브 베이지스를 클래스 함수로
구현한 코드를 이용하여, 학습

```
# 모델 결과 반환.  
def word_probabilities(self, counts, total_class0, total_class1, k):  
    # 단어의 빈도수를 [단어, p(w|긍정), p(w|부정)] 형태로 반환  
    return [(w,  
             (class0 + k) / (total_class0 + 2*k),  
             (class1 + k) / (total_class1 + 2*k))  
            for w, (class0, class1) in counts.items()]
```

단어 별로,
P(긍정단어), P(부정단어) 확률을 추출

긍정 감성 사전
인적 서비스 = ["친절", "직원", ...]

부정 감성사전
불청결성 = ["냄새", "더럽", ...]

긍정 감성 사전의 단어에는 각 단어의 P(긍정단어)
부정 감성사전 단어에는 각 단어의 P(부정단어)
“매칭”

데이터 모델링

긍정,부정 리뷰 평가함수 구축

(긍정리뷰 평가의 예시)

리뷰 : “ 직원이 친절하고...”

긍정 감성 사전

인적 서비스 = [“친절”, “직원”, ...]

부정 감성사전

불청결성 = [“냄새“, “더럽“, ...]

인적서비스 dictionary

+P(긍정|단어)

리뷰에 해당하는 단어가 나올 때 마다,

긍정 단어의 경우 $P(\text{긍정}|\text{단어})$ 가 더해지고

부정단어의 경우 $P(\text{부정}|\text{단어})$ 가 더해진다



각 호텔의 모든 리뷰를 긍정/부정 평가함수에 대입

추천 시스템 알고리즘

최종 결과

	긍정						부정					
	동반 여행 시 만족도	시설 만족도	접근 편의성	인적 서비스	주변 경관	긍정 총합	불청결	어메니티 및 비용 불만족성	시설 불편성	객실 불편성	불친절 성	부정 총합
롯데호텔	13.29	24.88	2.33	8.80	0.64	49.95	0.40	0.10	0.45	2.64	0.33	3.93
제주 신라 호텔	7.75	16.43	1.80	5.35	0.06	31.40	0.33	0.14	0.74	2.68	0.15	4.04
호텔 휘슬락	9.09	16.22	9.20	5.03	1.75	41.29	1.22	0.38	2.85	3.88	0.20	8.53
메종 글래드 제주	17.56	31.85	8.61	14.10	0.18	72.31	0.95	0.40	3.25	5.11	0.65	10.35
라마다 제주 시티 호텔	9.19	29.49	11.82	7.49	0.25	58.24	1.29	0.54	5.58	3.59	0.88	11.87
라마다 프라자 제주 호텔	15.41	39.50	7.50	8.58	2.44	73.43	1.24	0.26	2.08	7.17	0.59	11.34
신라스테이 제주	2.83	21.13	6.77	6.79	0.19	37.70	1.10	0.17	5.35	3.70	0.44	10.76
신화관 제주신화월드 호텔앤리조트	28.99	73.02	6.46	18.78	0.24	127.50	2.46	0.45	3.18	8.76	0.77	15.61
더 그랜드 섬오름	7.32	21.22	3.03	5.60	2.14	39.31	0.71	0.15	1.85	4.13	0.38	7.23
랜딩관 제주신화월드 호텔앤리조트	25.45	71.17	7.11	21.17	0.23	125.13	2.54	0.54	3.71	9.47	0.88	17.14

추천 시스템 알고리즘

STEP 1

긍정/부정 카테고리에서
1~3순위 속성을 선택

긍정

“접근 편의성“, “인적 서비스성”,
“시설 만족도“,...

부정

“불청결성“, “불친절성”,
“객실 불만족도“,...

STEP 2

순위별 가점 부여



긍정 총합이 가장 높은 호텔 추천
부정 총합 가장 높은 호텔 비추천

STEP 3

속성별 수치 제공
+

추천/비추천 호텔 시각화
(WordCloud)



기대효과 01

호텔 품질 및 서비스 분석에 기여,
개선 방향을 제시

기업에게 감성 분석 데이터 제공



고객서비스 품질

+

리뷰 관리

+

경쟁사 분석

+

긍정적 경험 강조

기대효과 02

감성분석의 점수 합계를 통해
이용자가 원하는 호텔 추천

이용자에게 감성 분석 데이터 제공

이용자의 선호 토픽 선정

이용자 선호 토픽과 가장 점수가 높은 호텔 추천

개선방안

1) 토픽의 단어 추가

- 토픽을 구성하는 단어 수를 맞춰 늘려 성능 향상을 기대

2) 불용어 추가 제거

- 분석 단계별 나타나는 불용어를 추가로 제거하여
성능 향상을 기대

긍정 토픽

▪ 동반 여행 시 만족도	111 개
▪ 시설 만족도	116 개
▪ 접근 편의성	80 개
▪ 인적 서비스	65 개
▪ 주변 경관	28 개

부정 토픽

▪ 불청결	96 개
▪ 메니티 및 비용 불만족성	54 개
▪ 시설 불편성	84 개
▪ 객실 불편	108 개
▪ 불친절	58 개

개선방안

1) 토픽의 단어 추가

- 토픽을 구성하는 단어의 수를 비슷하게 맞춰 성능 향상을 기대

긍정 토픽

- 동반 여행 시 만족도 111 개
- 시설 만족도 116 개
- 접근 편의성 80 개
- 인적 서비스 65 개
- 주변 경관 28 개

부정 토픽

- 불청결 96 개
- 어메니티 및 비용 불만족성 54 개
- 시설 불편성 84 개
- 객실 불편 108 개
- 불친절 58 개

2) 불용어 추가 제거

- 분석 단계별 나타나는 불용어를 추가로 제거하여 성능 향상을 기대

'힐링',
'라면',
'에스'],
['위치',
['공항',
'기차']

'무섭',
'고치',
'미숙',
'무양'

'최고',
'스타',
'박스',
'에코',
'매장'



감사합니다

