제주도 여행 루트 추천 시스템

이수경 이은서 이승우 주혜인 홍현경

Jeju 주제 선정 배경

최근 국내 여행의 트렌드 파악

1
2
3
4

국내 여행자들은 패키지 보다는
자유 여행을 선호하는 경향이 있음

'호캉스·한달살이' 패키지 대신 개별 여행 선호...여행업계 지각변동

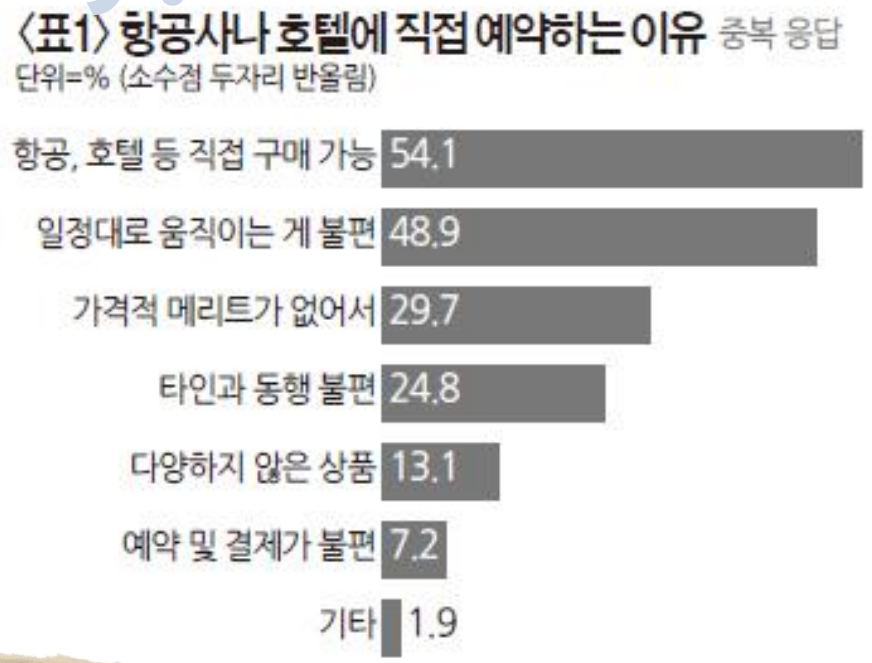
김은영 기자

입력 2019.10.27 10:00

'호캉스·한달살이' 등 개별 여행 인기에 패키지 여행객 줄어
여행업계, 맞춤형 여행 상품으로 체질 개선 시도

why?

출처: 여행 신문



해외여행과 달리 짧은 일정으로 가볍게 떠나는
여행이 주를 이루는 국내여행의 특징이 반영됨

Jeju 주제 선정 배경

자유 여행의 특징

1
2
3
4

개인의 취향과 이동 거리 등을 기준으로 웹사이트, SNS 등에서 여행지에 대한 정보를 조사하여 설계 필요

트래블 테크 (travel-tech)

기존 관광객들의 방문 장소 수집 및 분석

→ 개인의 취향 및 이동거리를 고려한 여행 경로를 설계하는 서비스의 필요성 증대

Jeju 데이터 수집

데이터 선정

여행 루트 계획 시 고려 할 사항

1
2
3
4

관광지

- 어디 가지? ➔ 관광지 Info
- 어떤 곳이지? ➔ 관광지 Review

with 크롤링 !!

VISIT JEJU

VISIT JEJU
+ Trip advisor

식당

- 어디 가지? ➔ 식당 Info
- 어떤 곳이지? ➔ 식당 Review

Jeju 데이터 수집

최종 관광지 데이터셋

PLC_INFO

VISIT JEJU에서 크롤링 한
1089개의 관광지 정보

PLC_name	Difficulty
Sogae	Convenient
Detail	b_parking
Eyong_time	b_elevator
Fee	b_toilet
Preoperty	b_toilet_space
Objectivty	b_seat
Etc	b_anneso
Soyo_sigan	b_dolbom

PLC_REVIEW

VISIT JEJU & Tripadvisor에서
크롤링 한 18310개 리뷰

plc_name
Tag
Review
ratings

Jeju 데이터 수집

최종 식당 데이터셋

1
2
3
4

FOOD_INFO

VISIT JEJU에서 크롤링 한
1326개의 음식점 정보

PLC_name

Sogae

Detail

Eyong_time

Convenient

b_active

address

FOOD_REVIEW

VISIT JEJU & Tripadvisor에서
크롤링 한 9693개 리뷰

plc_name

Title

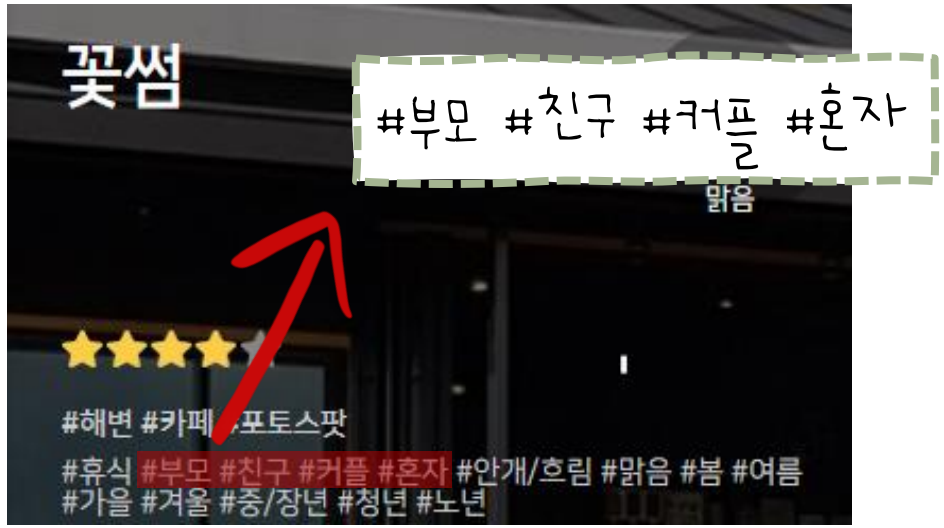
Review

Ratings

Visit_date

태그 정제

가장 많은 태그를 지닌 관광지 확인!



카페가 왜....관광지...?

태그의 신뢰성에 의문 발생

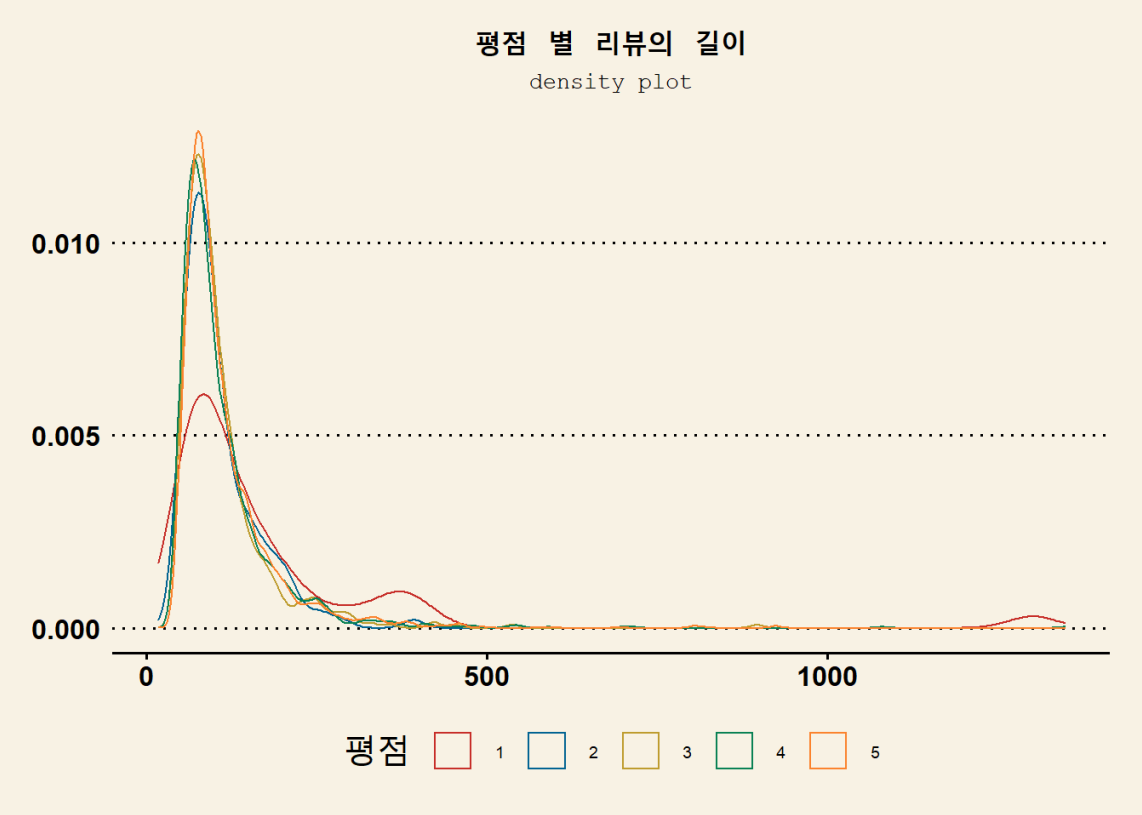
```
for i in data['tags']:
    if "커플" in i and "친구" in i:
        i.remove('커플')
        i.remove('친구')
        i.append('함께')
```

```
for i in data['tags']:
    if "함께" in i and "혼자" in i:
        i.remove('함께')
        i.remove('혼자')
```

'커플'과 '친구'가 동시에 있으면 → '함께'

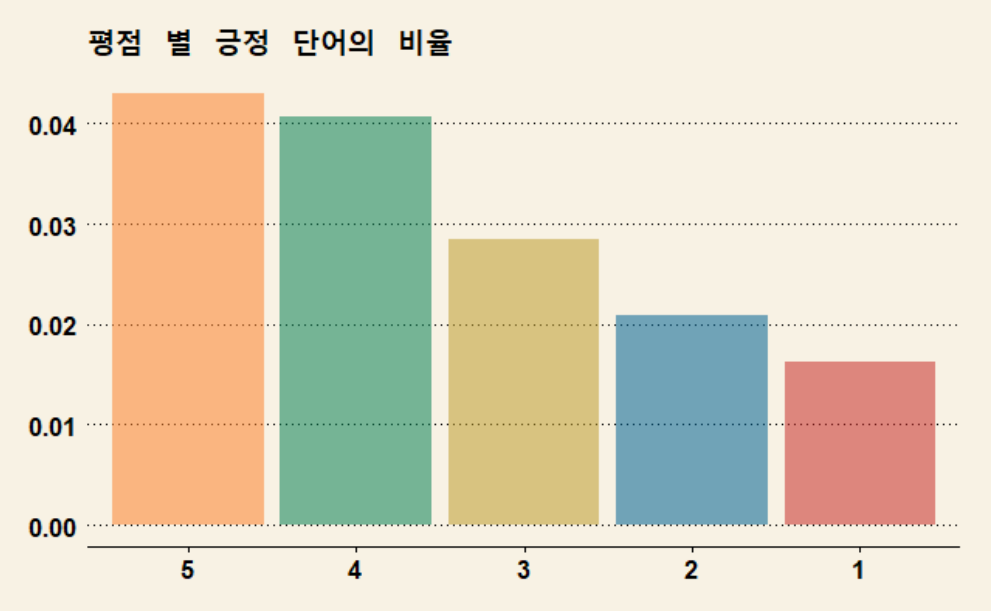
'함께'와 '혼자'가 동시에 있으면 → 모두 제거

리뷰의 길이와 평점의 관계



평점이 1인 경우 다른 평점을 줄 때에 비해 긴 편

Tripadvisor 리뷰 평점 5점 명사



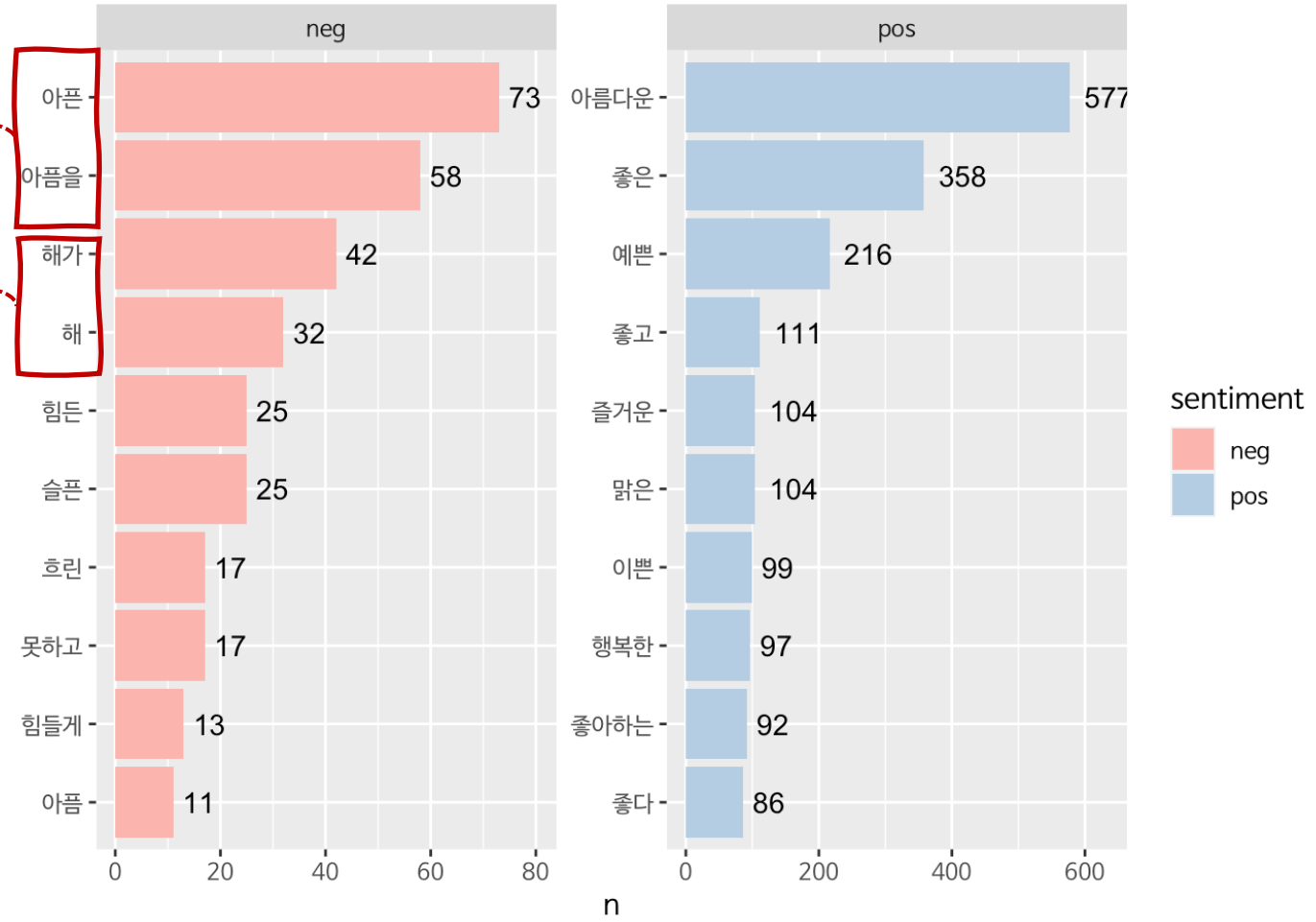
평점 5점을 준 리뷰의 경우 긍정적인 단어를 많이 사용한다

➡ 긍정적인 단어가 많으면 높은 평점을 주는데, 부정적인 단어가 많으면 낮은 평점을 받았을까?

비짓제주 리뷰의 감성분석

4.3 사건의 아픔을 담은 리뷰 내용

sun과 hurt의 중의적 의미



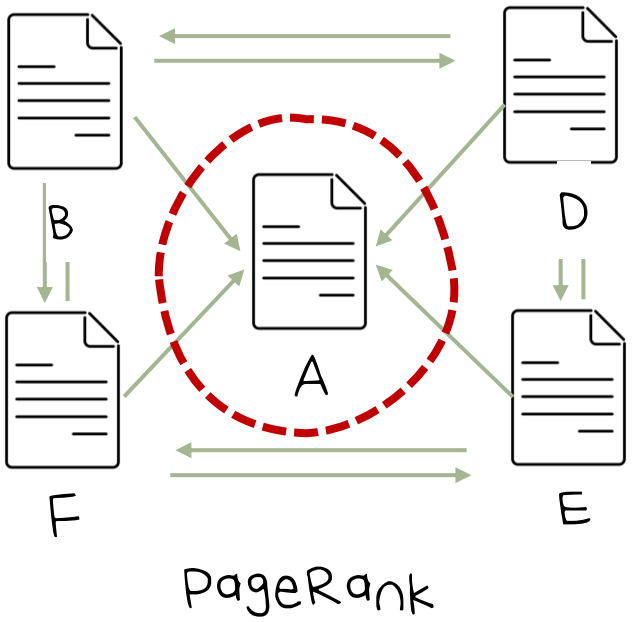
부정적인 단어가 등장한다는 것이 부정적인 리뷰를 의미하는 것은 아님

비짓제주 리뷰 워드클라우드

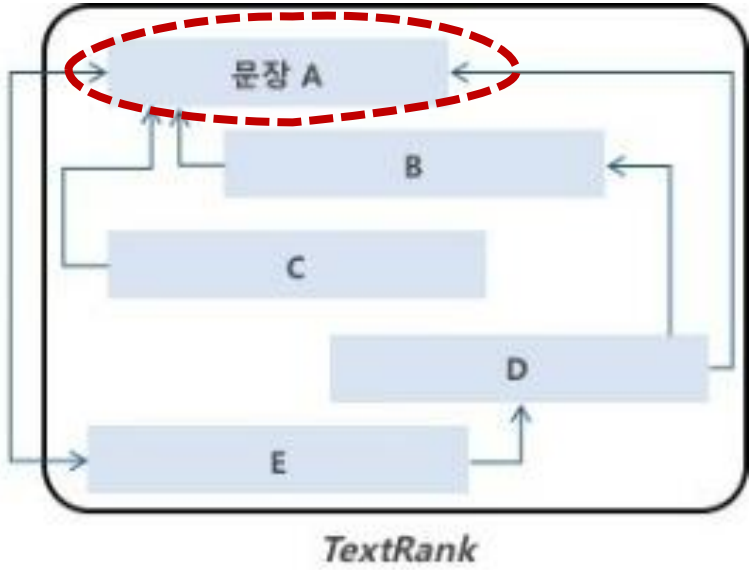
바다, 해변 사진 등 태그와 관련된 단어가 리뷰에 자주 등장함

하지만 관광지의 이름 역시 리뷰에 자주 등장함 => 리뷰에 나오는 관광지 명을 불용어로 처리!

WordRank란?



문서 내
문장에 적용



웹 페이지의 중요도를 따지는 방법인 PageRank를
텍스트에 적용하여 단어의 중요도를 계산함

WordRank 결과

관광지	result
백록담	('무척', 0.5238167593410135), ('올라가기', 0.0), ('너무', 0.0), ('험했지만', 0.0), ('눈에', 0.0), ('덮인', 0.0), ('백록담의', 0.0), ('설경은', 0.0), ('아름다웠습니다', 0.0)
성널오름(성판악)	('한라산', 0.7817498727584), ('코스', 0.4708824094344274), ('성판악', 0.0)

키워드 추출

TF-IDF란?

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

TF(단어 빈도, term frequency) – 특정 단어가 문서 내에 얼마나 자주 등장하는지 나타내는 값

DF(문서 빈도, document frequency) – 단어가 다른 문서에서 등장하는 빈도

IDF(역문서 빈도, inverse document frequency) – DF의 역수

TF-IDF – TF와 IDF를 곱한 값으로 특정 단어가 다른 문서에는 적고 해당 문서에 등장하는 정도

Jeju 키워드 추출

TF-IDF 결과

관광지	result
백록담	('한라산', 19.101401635251147), ('코스', 5.246934633318522), ('성판악', 3.7967730860385673), ('합니다', 3.2017959485449636), ('파노라마로', 2.98101223032703 94), ('정상', 2.6648947443866633), ('그저', 2.5569953267434116), ('곳이었다', 2.5201248528802584), ('대피소', 2.469460180941927 4), ('이번', 2.399706708973907 3)
성널오름(성판악)	('한라산', 9.206410624507523), ('성판악', 5.883576230513387), ('코스', 4.197417332899637), ('가장', 3.271814397846293 7), ('설경이', 2.616042258735787), ('겨울', 2.359624231112910 8), ('시점이라고', 2.323282708889734 8), ('해발고도', 2.251057615564641), ('생각합니다', 2.046292560712449 6), ('육개장을', 2.042434703142214)