

아무데나 가는건 아닌거잖아

제주도 여행객들을 위한 맛집 / 명소 추천서비스

[3팀] 어디가맨





01

프로젝트 배경

02

팀 구성 및 역할

03

수행절차 및 방법

04

프로젝트 수행결과

05

결론 및 향후 과제

01 프로젝트 배경



이 출시 배경

소믈리에타임즈 | 2019.11.28.

한국인 여행객 10명 중 9명, '즉흥 여행 경험'... 제주도부터 호강...

집중되었던 즉흥여행의 범주가 확장되고 있다. 보다 간편한 예약 시스템과 여행 계획에 유용한 다양한 서비스들도 등장했다. 세계적인 온라인 여행사 익스피디아는...

스포츠동아 | 2020.08.17. | 네이버뉴스

"올 여름 즉흥여행 늘었다", 2일전 숙소예약이 절반

없던 즉흥여행 트렌드가 강했던 것으로 나타났다. 심지어 투숙 2일 전에 숙소를 예약하는 비율이 전체... 코로나19나 집중호우 등 휴가일정을 정하기 어려운 불가항...

스포츠월드 | 2021.02.05. | 네이버뉴스

제주도 여행욕 샘솟네...효연의 '탐나호' 눈길

그래서 '탐나호'의 전반적인 분위기는 국내의 대표 여행지인 '제주도'에서 벌어지는 '즉흥적'이고, '일상적'이며 외부 접촉을 최소한 한 '언택트형 감성 여행'을 따르고...



제주도 오긴 왔는데...

가까운 맛집은 어떻게 찾지?
걸기 좋은 곳에 가고 싶은데
어떻게 찾지?



이런 여행자들을 위해
'어디가맨'이 출시한 서비스

01-1 출시 컨셉



02

팀 구성 및 역할



02 역할 분담

김남경 (팀장)

- 데이터 크롤링
- 추천 시스템 모델링

김용호

- 데이터 크롤링
- 추천 시스템 모델링

이종민

- 데이터 크롤링
- 추천 시스템 모델링

여정문

- 데이터 수집 및 전처리
- 시각화 및 웹 배포

문준영

- 데이터 수집 및 전처리
- 시각화 및 웹 배포

이주남

- 데이터 수집 및 전처리
- 시각화 및 웹 배포

03

수행절차



03-1 work-flow

10/5 ~ 10/11 : 주제 선정 및 일정 수립

10/12 ~ 10/17 : 데이터 수집(크롤링)

10/18 : ERD 구성 및 AWS 파이프라인 구축

10/18 ~ 10/22 : 데이터 전처리 및 모형화

10/18 ~ 11/5 : 추천시스템 모델링

10/26 ~ 11/7 : 웹 배포

11/8 ~ 11/11 : 발표 포트폴리오 작성 및 발표 준비

11/12 : 프로젝트 경진대회



04

프로젝트 수행결과



04-1 데이터 명세

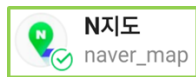
1. 제주 음식점 데이터

- 상호명
- 도로명주소
- 업종소분류



2. 네이버 맵 크롤링 데이터 (음식점)

- 음식점 평점
- 음식점 사진
- 리뷰 개수



3. 비짓 제주 크롤링 데이터 (관광지)

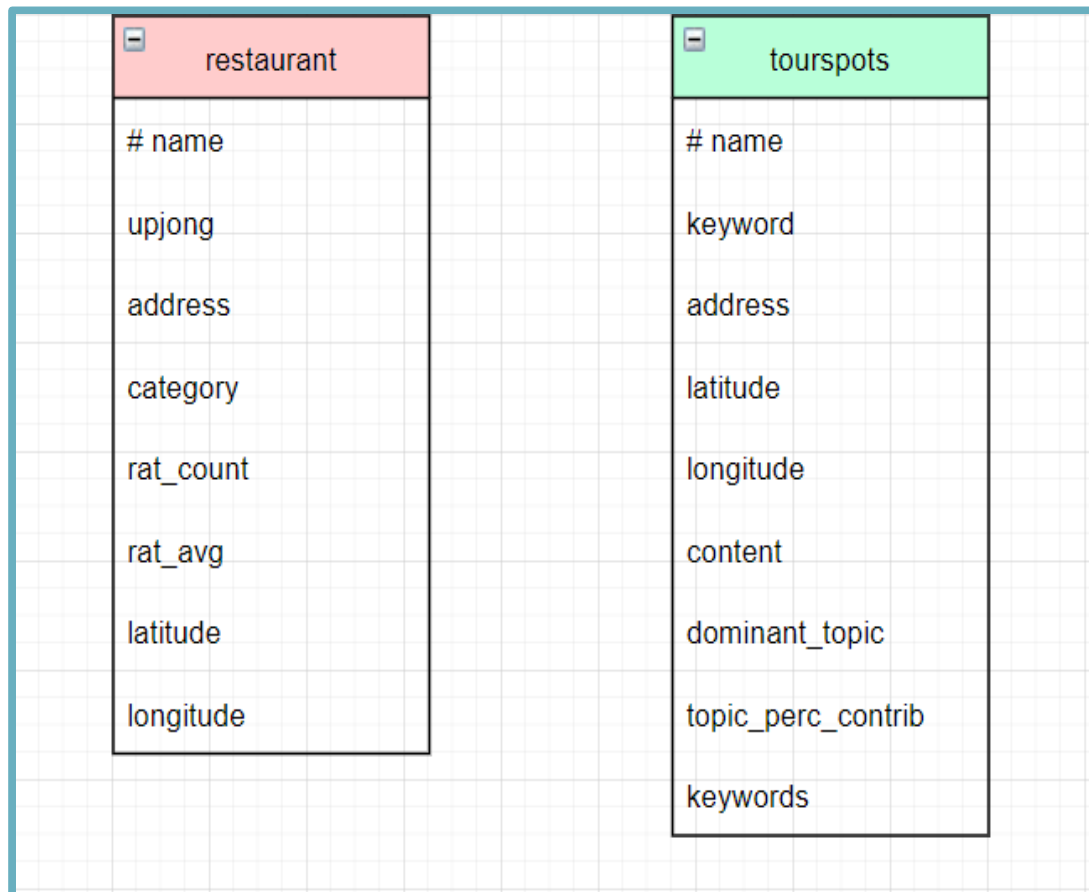
- 관광지명
- 주소
- 관광지별 리뷰 내용



04-2 데이터 파이프라인



04-2 ERD



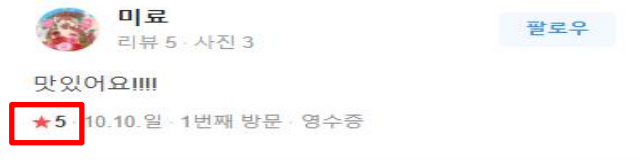
04-3 데이터 수집 및 전처리



- 음식점의 상세 페이지 url 검색 시
상호명과 주소를 혼합하여 검색 키워드로 설정

⇒ 동명 음식점 혹은 소재지가 제주가 아닌 음식점 데이터 필터링

- 상호명, 리뷰 작성자 ID, 평점 및 리뷰 사진 크롤링



	name	upjong		address	rating	id
0	통복리해녀장수촌	회집		제주특별자치도 제주시 구좌읍 구좌해안로 32	4.5	60e668059f1eace64860294b
1	통복리해녀장수촌	회집		제주특별자치도 제주시 구좌읍 구좌해안로 32	4.0	5bf276974dbe68446e4337b2
2	통복리해녀장수촌	회집		제주특별자치도 제주시 구좌읍 구좌해안로 32	5.0	5d269e0ed9a77f8490cbfafb
3	통복리해녀장수촌	회집		제주특별자치도 제주시 구좌읍 구좌해안로 32	3.0	603ef197844400635a88a8e1
4	통복리해녀장수촌	회집		제주특별자치도 제주시 구좌읍 구좌해안로 32	5.0	5e0fc4c98f87a842bc9deec0
...
950261	한옥마루카페	커피숍		제주특별자치도 서귀포시 성산읍 한도르 269-23, 부속동 1층	4.0	5e27bd1c8f87a842bcae972c
950262	한옥마루카페	커피숍		제주특별자치도 서귀포시 성산읍 한도르 269-23, 부속동 1층	5.0	5da09bf28f87a842bca15a74
950263	한옥마루카페	커피숍		제주특별자치도 서귀포시 성산읍 한도르 269-23, 부속동 1층	5.0	5e5cb4848f87a842bcf753a4
950264	한옥마루카페	커피숍		제주특별자치도 서귀포시 성산읍 한도르 269-23, 부속동 1층	3.0	5e4c6ac98f87a842bc30aab8
950265	한옥마루카페	커피숍		제주특별자치도 서귀포시 성산읍 한도르 269-23, 부속동 1층	5.0	5e7628e68f87a842bcfb6dbc


950266 rows × 5 columns

04-3 데이터 수집 및 전처리

▶ 비짓제주 크롤링

공지 | '탐나는전'으로 공영관광지를 알뜰하게 이용하세요!

로그인 | 한국어 ▾

VISIT JEJU

관광지

음식점

숙박

쇼핑

면세점

제주이야기

여행필수정보


제주여행추천

나의 여행

🔍

인기 지도 우도 애월 카페

JEJU DUTY FREE
제주관공공사 인타넷만점

제주여행공유

제주시

서귀포

🏠 > 관광지 > 전체 ▾

단축 URL

TALK

f

t

🖨

📱

제주도 모든 여행지를 한 눈에...

내가 가본 제주는 어디까지일까? 수많은 제주의 아름다운 여행지를 취향에 맞게 선택해보자. 368개의 크고 작은 오름을 비롯하여 눈 돌리면 어디에서나 마주치는 한라산 그리고 푸른 바다.... 제주의 보석 같은 여행지가 여러분의 선택을 기다린다.

전체	자연	문화관광	레저/체험	테마관광지
	섬속의 섬	도보	포토스팟	템플스테이
	제주 4·3	의료관광	웰니스	

▶ 비짓제주 크롤링

방문했어요 #5월엔제주 #우도의봄 #반짝반짝우도 리뷰 번역

우도섬에 봄이 왔습니다. 유채와 청보리 파란 하늘과 그보다 더 푸른섬 우도가 제주의 반짝반짝한 바다에서 빛나고 있어요. 우도에서 이 봄을 만끽하시길...



댓글(0)

댓글 쓰기

방문했어요 ##우도일몰 #우도1박2일 #우도에서인생샷 ##낮과다른우도의밤 ##우도해안

리뷰 번역

일몰이 제일 아름답다는 우도 낮에 사람들로 가득한 우도는 마지막 배가 떠나고 나면 동그랗고 핑크빛의 해가 지는 동안 몇몇 우도에 남은 사람들은 30여분동안 조용한 해변에서 해가 지는 모습만 보는 것만으로 감동적인 영화한편을 보는 것보다 더한 감동을 느낍니다. 첫배가 들어오기전 밀물과 바람으로 가득한 해변은 전날 저녁과는 다른 모습입니다.



	name	keyword	content	tags	address	address1
0	더리조트학교(구, 더리본고)	미와	초등학교 해안들과 만경화 가을 여행 속 방문한 애플 더리본고	['#제주힐러러학', '#더리본고', '#초등학교', '#제주힐러러학', '#무지개빛우도']	제주특별자치도 제주시 애월읍 하가로 195	애월읍
1	더리조트학교(구, 더리본고)	미와	무지개빛 아름다운 더리본고	['#더리본고', '#더리본고', '#미와', '#미와']	제주특별자치도 제주시 애월읍 하가로 195	애월읍
2	더리조트학교(구, 더리본고)	미와	무지개빛 아름다운 더리본고	['#더리본고', '#더리본고', '#무지개빛', '#미와']	제주특별자치도 제주시 애월읍 하가로 197	애월읍
3	더리조트학교(구, 더리본고)	미와	일출명소 너무 예쁜 초등학교에서 아내와 함께 좋은 추억과 사진을 남겼어요!	['#5월엔제주', '#더리본고', '#더리본고', '#초등학교']	제주특별자치도 제주시 애월읍 하가로 198	애월읍
4	더리조트학교(구, 더리본고)	미와	#제주도 #더리본고 #연꽃 #연꽃마을 정말 원가 작은 무늬가 자꾸 생각나고 웃음이...	['#제주도', '#더리본고', '#연꽃', '#연꽃마을']	제주특별자치도 제주시 애월읍 하가로 199	애월읍
...
13063	누워마루거리(구 바오전)	도트	#바오전거리 #신라스테이제주 #돌하루방	['#바오전거리', '#신라스테이제주', '#돌하루방']	제주특별자치도 제주시 신광로 47	제주시
13064	누워마루거리(구 바오전)	도트	물국인들이 많이 찾다는 제주시 연동 바오전거리.	[]	제주특별자치도 제주시 신광로 47	제주시
13065	숫모르편백나무숲	도트	제주생태공원에서 풀들자연휴림으로 이어지는 길기 좋은 일이다. #숫모르편백나무숲	['#숫모르편백나무숲', '#숫모르편백나무숲', '#숫모르편백나무숲']	제주특별자치도 제주시 516로 2508	제주시
13066	숫모르편백나무숲	도트	#숫모르편백나무숲	['#숫모르편백나무숲']	제주특별자치도 제주시 516로 2508	제주시
13067	삼의악트레킹코스	도트	#산과마을 #삼의악마을	['#산과마을', '#삼의악마을']	제주특별자치도 제주시 아라동 산 24-2	제주시

13068 rows x 6 columns

04-3 데이터 수집 및 전처리

▶ 네이버맵, 비짓제주 데이터의 위·경도 수집

```
import googlemaps
import pandas as pd

my_key = "AIzaSyDa9AwBq_wIY41raiCjJvF7CWgbHYyYyy0"
maps = googlemaps.Client(key=my_key) # my key값 입력
lat = [] #위도
lng = [] #경도

# 위치를 찾을 장소나 주소를 넣어준다.
places = data2["address"]

i=0
for place in places:
    i = i + 1
    try:
        print("%d번 인덱스에서 %s의 위치를 찾고있습니다"%(i, place))
        geo_location = maps.geocode(place)[0].get('geometry')
        lat.append(geo_location['location']['lat'])
        lng.append(geo_location['location']['lng'])

    except:
        lat.append('')
        lng.append('')
        print("%d번 인덱스 위치를 찾는데 실패했습니다."%(i))

# 데이터프레임만들어 출력하기
data3 = pd.DataFrame({'위도':lat, '경도':lng}, index=places)
print(data3)
```

	name	keyword		content	tags	address	address1	위도	경도
0	더럭초등학교 (구, 더럭분교)	문화	초등학교 동창들과 함께한 가을 여행 속 방문한 애월 더럭분교	[#제주컬러어택', '#더럭분교', '#초등동창가을여행', '#무지개빛우정]	제주특별자치도 제주시 애월읍 하가로 195	애월읍	33.453459	126.345275	
1	방주교회	문화	좋아요좋아요좋아요	[#봄', '#미식', '#애인', '#혼자', '#직장동료]	제주특별자치도 서귀포시 안덕면 산록남로762번길 113	안덕면	33.305073	126.387664	
2	김영갑갤러리 두모악	문화	오렌지빛 가득한 김영갑갤러리두모악. 그는 가도 그의 영혼은 생기있게 빛나고 있다.	[#제주컬러어택', '#김영갑갤러리두모악', '#오렌지빛이눈에들여왔다]	제주특별자치도 서귀포시 성산읍 삼달로 137	성산읍	33.372065	126.854180	
3	너른송이 4.3 기념관	문화	운영시간이 6시까지인데 미리 알아보지 못하고 5시 58분에 도착했습니다 아쉽게도 발걸...	[#너른송이', '#4.3길', '#북촌리', '#순이삼촌]	제주특별자치도 제주시 조천읍 북촌3길 3	조천읍	33.545988	126.688764	
4	제주민속촌	문화	친동생 군입대와 가족여행 겸 다녀온 제주여행입니다. 입대로 인한 슬픔과 가족여행이라...	[#제주민속촌', '#제주도', '#가족여행', '#군입대]	제주특별자치도 서귀포시 표선면 민속해안로 631-34	표선면	33.322345	126.841487	
...
551	제주올레 18코스	도보	#올레길 18코스 #우리동네 #급트레킹 #완벽한 날씨 #노을 #사라봉 #별도봉 #화...	[#올레길', '#우리동네', '#급트레킹', '#완벽한', '#노을', '#사라...	제주특별자치도 제주시 삼양1동 1131-2	제주시	33.528676	126.596536	
552	제주올레 6코스	도보	#제주가를#올레6코스#제주올레걷기축제#참가#하포항#서귀포질서시공원#소전지#이중섭길...	[#제주가를', '#올레6코스', '#제주올레걷기축제', '#참가', '#하포항', '#이중섭길', '#소전지', '#이중섭길']	제주특별자치도 서귀포시 보목동 1377-4	서귀포시	33.237120	126.596068	
553	제주올레 17코스	도보	#제주보리#도두마을#올레길17코스#도두마을을 걷노라면 즐겁다#...	[#제주보리', '#도두마을', '#올레길17코스']	제주특별자치도 제주시 도두1동 2612-1	제주시	33.505224	126.467967	
554	숫모르편백나무숲길	도보	제주생태공원에서 절물자연휴양림으로 이어지는 걷기 좋은 길이다.#숫모르편백나무...	[#숫모르편백나무숲길', '#숫모르숲길', '#편백나무', '#숲길']	제주특별자치도 제주시 516로 2596	제주시	33.430924	126.595761	
555	삼의악트레킹코스	도보	#산과오름 #삼의악오름	[#산과오름', '#삼의악오름]	제주특별자치도 제주시 아라동 산 24-2	제주시	33.440020	126.561964	

04-3 데이터 수집 및 전처리

▶ 네이버맵 크롤링 데이터 전처리

- 업종소분류가 음식점으로 등록되어 있지만 음식점이 아닌 경우와 프랜차이즈 음식점인 경우 데이터 삭제
- 음식점 주소를 바탕으로 행정구역 컬럼 생성 (사용자의 실시간 위도 경도를 사용하는 서비스로 변경하면서 추후 삭제)
- 전처리 결과 음식점 데이터 로우 개수 변화 : 7312 → 약 4400 개

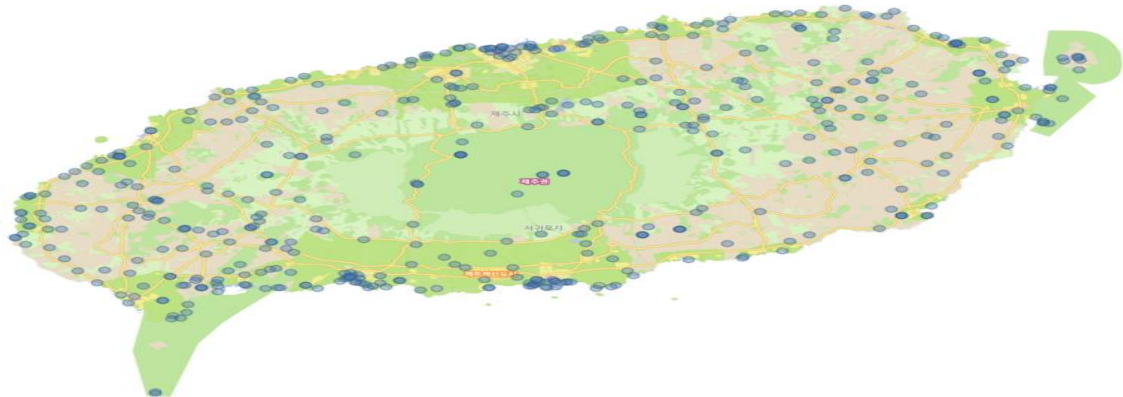
▶ 비짓제주 크롤링 데이터 전처리

- 자연어 모델에 사용되는 데이터이고 내국인을 대상으로 하는 서비스이므로
 1. 리뷰가 존재하지 않는 관광지
 2. 외국어로 작성된 리뷰⇒ 결측치 처리
- 리뷰 데이터를 spark에서 불러왔을 때 텍스트에 들어 있는 엔터 때문에 텍스트들이 한 컬럼 안에 속하지 못하는 문제 발생
⇒ 엔터값 삭제
- 관광지 주소를 바탕으로 행정구역 컬럼 생성 (사용자의 실시간 위도 경도를 사용하는 서비스로 변경하면서 추후 삭제)

< 제주시 일반음식점 분포 >



< 제주시 관광지 분포 >



04-4 데이터 시각화

- 식당들이 제주시내, 서귀포 시내에 밀집되어 있음
- 해변가를 따라 위치하고 있음
- 최근 떠오르는 애월읍 부근도 식당이 밀집되어 있음

- 관광지는 시내뿐만 아니라 고르게 분포되어 있음
- 명소가 밀집된 지역의 경우 중문 관광단지 등의 다양한 관광단지가 산재되어 있음

04-5 맛집 추천 시스템

네이버 리뷰 데이터를
이용한 리뷰수 가중 평점 +
식당 간 코사인 유사도
이용한 유사한 식당 추천

- 간단하고 단순한 필터링
>>> 폐기

회원이 가입한 회원의
식당 평가를 이용한
협업 필터링 추천 시스템

- 회원 가입과 평점을 기입하는
번거로움 때문에 불편함 예상
>>> 폐기

리뷰 수 가중 평점
사용자의 위치를 이용하여
새로운 지표인

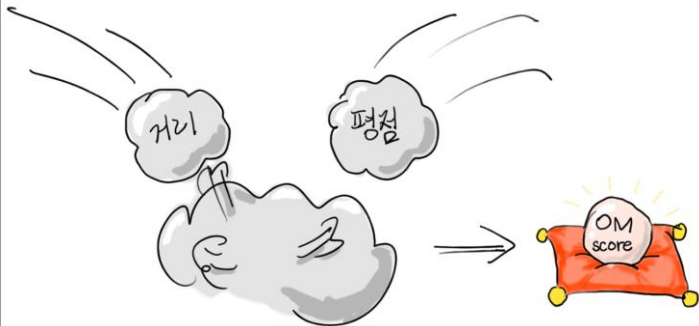
OMscore(어디가맨 지수)를
생성하여 추천

- 자동으로 위경도를 가져오기 때
문에 사용자는 맞춤 추천 서비스
를 제공 받을 수 있음

04-6 맛집 추천 시스템

OMscore (어디가맨 지수)란?

식당의 **평점**과 **리뷰 수**,
사용자와 식당간 **거리**를
모두 고려하기 위해
어디가맨이 만들어낸 점수



리뷰수 가중 평점

- 리뷰수에 가중치가 붙어 리뷰수가 많으면 높은 점수, 적으면 낮은 점수가 나오도록 설정

거리 점수

- 사용자와 식당간의 거리를 haversine 패키지를 이용해 구한 뒤 로그함수를 사용해 가까운 거리는 완만하게, 너무 먼 거리는 급격하게 점수가 떨어지도록 설정

OMscore (어디가맨 지수)

- 리뷰수 가중 평점 + 거리 점수 = OMscore
- 가중 평점과 거리점수에 각각 적당한 비율을 설정하여 두 가지 모두 반영된 지표로 생성

04-6 맛집 추천 시스템

리뷰 수 가중 평점

$$\frac{c}{m+c} \times R + \frac{m}{m+c} \times \text{AVG}$$

m = 기준 리뷰 수(100으로 설정)(상위 12.2%)

c = 식당의 리뷰 수

R = 식당의 평균 평점

AVG = 전체 3200개 식당의 평균 평점(4.415)

배달의 민족을 보면 리뷰가 100개 이상이면 똑같이 100+로 표시함 -
리뷰가 100개 이상이면 충분히 많은 것이라 판단
서비스의 목적이 **사람들에게 검증된 맛집**을 추천해주는 것이기 때문
에 **적어도 100개**의 리뷰가 있어야 사람들이 많이 찾는 식당이라 판
단하여 기준 리뷰 수를 100개로 지정

- 리뷰 수 100개 이상의 식당은 식당의 평점에 영향을 더 받음
- 100개 미만의 식당은 전체 식당 평점에 영향을 더 받음

<리뷰 수 가중 평점 계산 결과>

	name	rat_count	rat_avg	weighted_rating
4396	흑돼지해물삼합	1634	4.927785	4.898214
1472	제주공항본점공항공복배기	509	4.952849	4.864538
951	반디파스타	793	4.916772	4.860586
3756	제주순풍해장국함덕점	2961	4.873691	4.858708
2203	중문색달통갈치	952	4.902311	4.855992
99	호커센터	559	4.920394	4.843708
3923	대한민육	361	4.959834	4.841657
2329	중문흑돼지전국	780	4.890385	4.836368
3992	만족한상회	670	4.890299	4.828576
2703	말젖은	716	4.879888	4.822921

<리뷰수에 따른 가중평점 영향>

	name	rat_count	rat_avg	weighted_rating
2657	월정해변식당	50	4.72	4.516691
2724	하도핑크	150	4.72	4.598015

<배달의 민족 리뷰>

←

한식

한식

분식

돈까스·회·일식

치킨

피자

아시아

배달 빠른 순

배달팁 낮은 순

기본순

주문 많은 순

손찬

반찬백화점

4.8(100+)

간장불고기, 고등어구이(자반고등어)

최소주문 13,000원, 배달팁 0원~

40~55분

위생정보

돼지네 매운갈비찜

매운 돼지갈비찜, 소불고기

4.9(100+)

매운 돼지갈비찜, 소불고기

최소주문 11,000원, 배달팁 2,000원~

72~87분

04-6 맛집 추천 시스템

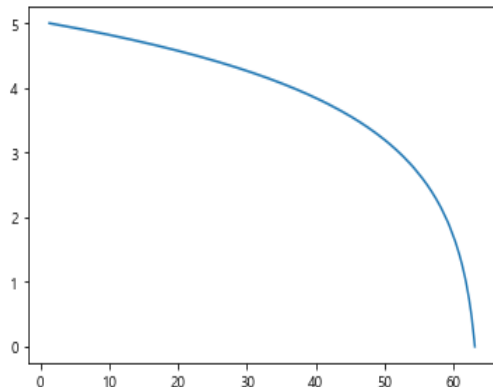
거리 점수

haversine 패키지를 이용해 사용자의 위경도와 식당의 위경도로 거리를 구한 뒤

거리에 대한 점수를 할당하기 위해
로그 함수를 사용하여

- 가장 먼 식당에 0점
- 가장 가까운 식당에 5점 부여

가까운 거리부터 중간정도까지는
완만하게 점수가 감소하지만
중간부터 후반까지는
급격하게 점수가 감소하여
너무 먼 식당은 추천하지 않도록 점수 책정



	name	Latitude	Longitude	distance	distance_log
2852	오데뜨	33.410870	126.295600	1.275832	5.000000
2782	돈내코순두부한림점	33.409700	126.282500	1.949924	4.986978
119	통성식당	33.418670	126.312681	2.386440	4.978471
447	해담은	33.409500	126.275472	2.509936	4.976053
2853	흑백돈한림점	33.398090	126.273000	2.515446	4.975945
...
3682	해와달그리고섬	33.510756	126.965529	62.958426	0.184836
1966	우도특별시	33.506700	126.967000	63.007241	0.133167
1268	제주삼촌네우도점	33.497653	126.969065	63.017432	0.122095
1572	바위소리	33.497690	126.969200	63.030508	0.107739
1838	아비알또	33.498760	126.970000	63.123833	0.000000

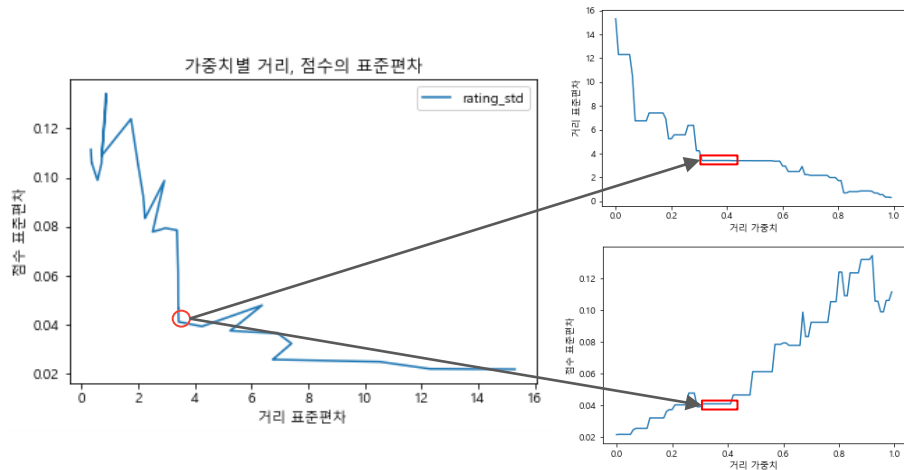
OMscore(어디가맨 지수)

앞서 구한 리뷰 수 가중 평점과 거리점수에 각각 비율을 부여하여 새로운 지표인

OMscore(어디가맨 지수)를 생성

- 특정 위경도 [33.3,126.3]에서 거리 가중치를 0~1까지 0.01단위로 부여하면서 OMscore 측정
- 구한 OMscore 상위 10개 식당의 거리 가중치 별 가중평점과 거리점수의 표준편차를 측정
- 특정위경도 [33.3,126.3]에서 표준편차를 구했을 때 오른쪽과 같은 그림
- 너무 멀거나 너무 낮은 평점인 식당은 추천하지 않는 것이 목적이므로 $y=x$ 직선에 가깝고 원점에 가까우면(빨간 동그라미) 적절한 가중치라고 판단
- 5개의 위경도에서 측정하였을 때 거리 가중치가 0.41일 때 가장 성능이 좋음

04-6 맛집 추천 시스템



거리 가중치
0.31~0.41에서
적절한 표준편차를
구할 수 있음

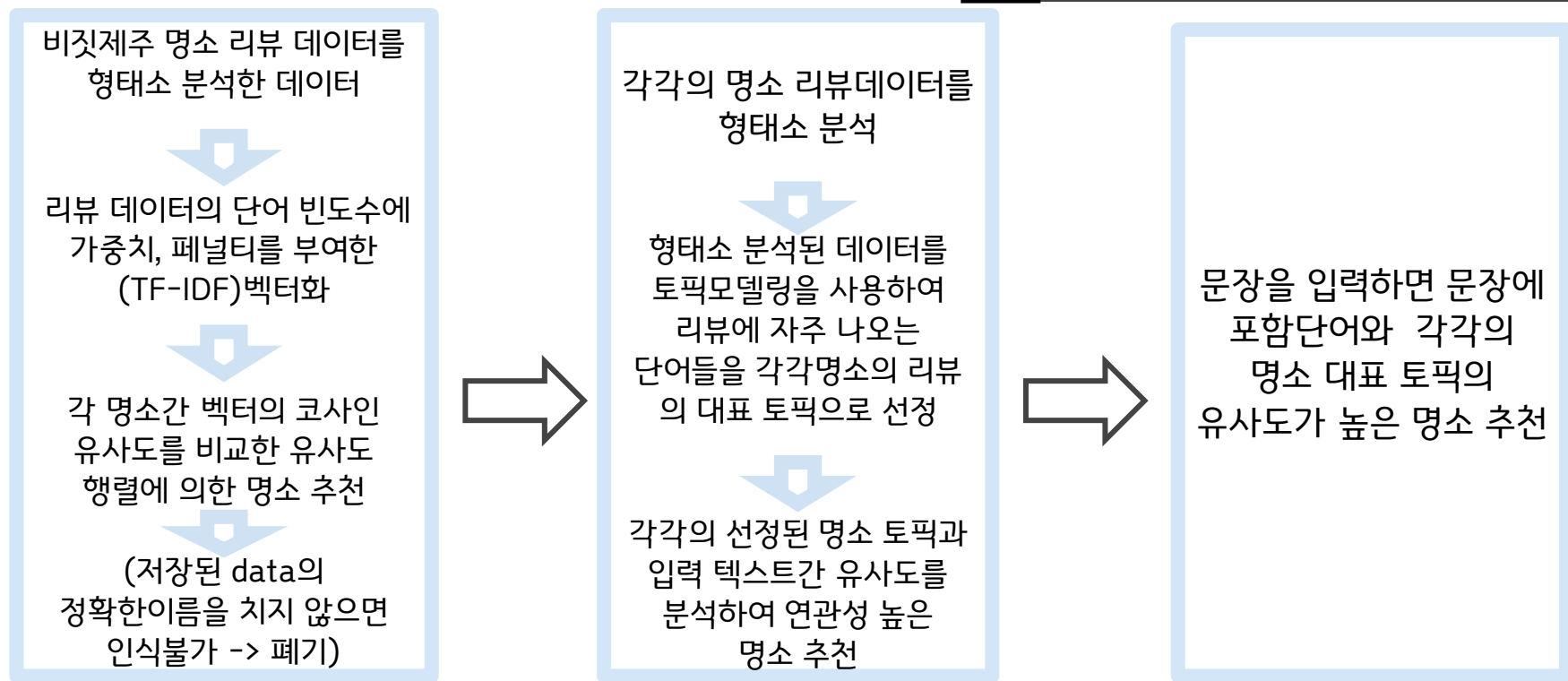
5곳의 위경도에서 구해보았을 때

- [33.4, 126.4] - 0.37~0.39 - 0.38
- [33.3, 126.35] - 0.34~0.4 - 0.37
- [33.3, 126.4] - 0.44~0.64 - 0.54
- [33.4, 126.3] - 0.36~0.44 - 0.4
- [33.3, 126.3] - 0.31~0.41 - 0.36
- 평균 - 0.41

	name	address	weighted_rating	distance	OMscore
2203	중문색달물갈치	제주특별자치도 서귀포시 일주서로 993, 1층 101호 (석달동)	4.855992	10.761324	4.846509
2522	만복돼지	제주특별자치도 서귀포시 안덕면 산방로 53	4.777784	6.754739	4.829278
349	물동식당	제주특별자치도 제주시 한경면 명이5길 20	4.735152	3.442542	4.826769
507	오만정상	제주특별자치도 제주시 한림읍 일주서로 5083, 1층	4.812421	10.922898	4.819598
2359	고집돌우럭 중문점	제주특별자치도 서귀포시 일주서로 879, 2층 (석달동)	4.821784	11.811861	4.818445
165	올레마당	제주특별자치도 서귀포시 안덕면 사계남로 224	4.762493	7.478784	4.815160
3992	만복한상회	제주특별자치도 서귀포시 중문상로 58-5, 1층 (중문동)	4.828576	12.863284	4.814433
160	알뜰네집화순점	제주특별자치도 서귀포시 안덕면 화순로 6, 1층	4.744410	6.681430	4.810101
2329	중문록화지천국	제주특별자치도 서귀포시 이머도로 137, 1층 (대포동)	4.836368	14.390308	4.807137
2228	전통가들문분점	제주특별자치도 서귀포시 일주서로 873, 1층 (석달동)	4.784129	11.868588	4.795799

<거리 가중치 0.41로 설정한 OMscore결과>

04-7 명소 추천 시스템



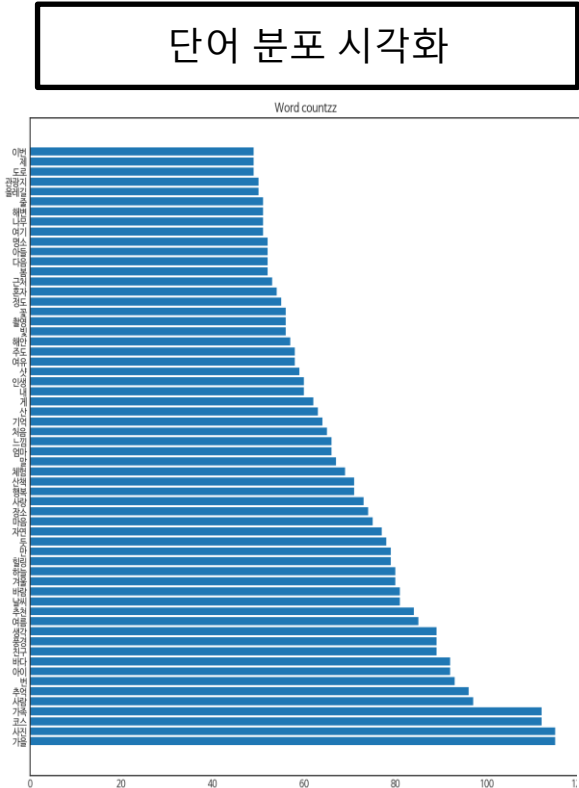
텍스트 데이터 전처리

- 출시배경에 기반하여 즉흥적으로 여행 오는 한국인들을 위한 서비스이므로 명소 리뷰 데이터에서 불필요한 단어 (영어,숫자,특수문자)들을 제거하여 한국어 명사만 추출
- 리뷰가 없는 명소들은 자연어 처리가 안되기에 삭제하고 각각의 명소 별 리뷰 데이터를 통합
- 리뷰에서 나오는 단어들 중에 **Counter** 함수를 이용하여 단어 빈도수를 확인하고 불필요한 단어(조사,특이값,잘못된 형태소 분석)들을 **Stopwords**에 포함시켜 제거
- **Konlpy** 패키지에서 빠른 연산속도와 상위품질성을 보장하는 **Mecab**모듈을 사용하여 각 리뷰데이터에서 명사를 추출

- 출시배경에 기반하여 즉흥적으로 여행 오는 한국인들을 위한 서비스이므로 명소 리뷰 데이터에서 불필요한 단어 (영어,숫자,특수문자)들을 제거하여 한국어 명사만 추출
- 리뷰가 없는 명소들은 자연어 처리가 안되기에 삭제하고 각각의 명소 별 리뷰 데이터를 통합
- 리뷰에서 나오는 단어들 중에 **Counter** 함수를 이용하여 단어 빈도수를 확인하고 불필요한 단어(조사,특이값,잘못된 형태소 분석)들을 **Stopwords**에 포함시켜 제거
- **Konlpy** 패키지에서 빠른 연산속도와 상위품질성을 보장하는 **Mecab**모듈을 사용하여 각 리뷰데이터에서 명사를 추출

04-7 명소 추천 시스템

단어 분포 시각화



추억	마음	체험	샷	혼자	해변	올레길	제	이번
사람	자연	산책	인생	정도	나무	줄	관광지	도로
가족	دت	행복	내	꽃	아들	명소	여기	
코스	만	사랑	계	촬영	근처	봄	다음	
	힐링	장소	산	여유	주도	해안	빛	
사진			말	엄마	느낌	처음	기억	
가을	여름	추천	날씨	바람	겨울	하늘		
	번	아이	바다	친구	풍경	생각		

04-7 명소 추천 시스템

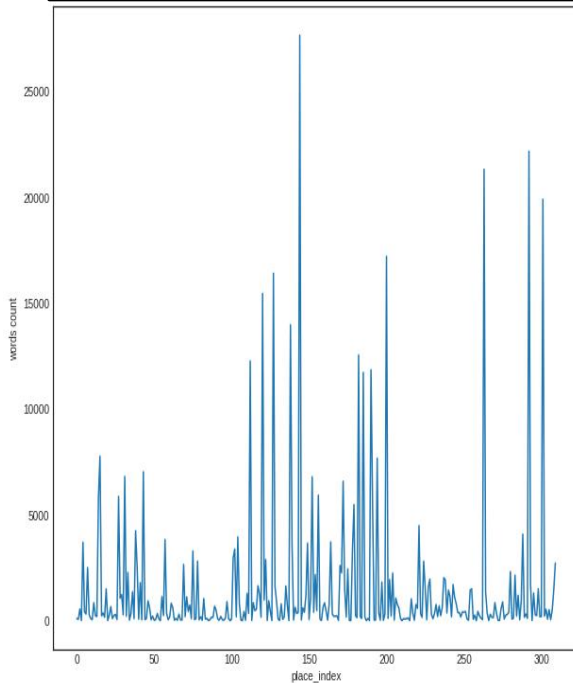
모델링 데이터 전처리

-문서 내에 주제토픽을 분류하는데 있어서 리뷰내용이 너무 적은 데이터를 포함시키면 적절한 토픽을 구하는데 차질이 생겨 명소 별 문자열 길이 분석

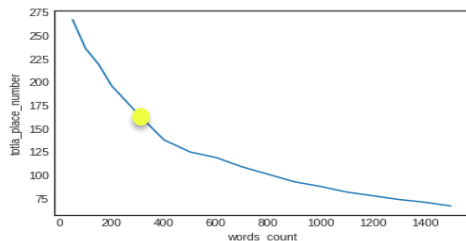
- 리뷰내용이 적을수록 인기가 없는 곳이라고 생각되고 추천을 해 줄만한 곳은 아니라고 판단하여 상위 50%이상의 데이터만 추출

-상위 50% 대략 문자열길이 350 이상 의 데이터 160개의 추출 후 모델링 진행

리뷰 데이터 문자열 길이 분포



문자열 길이 별 명소 개수



문자열 길이의 사분위수

count	310.000000
mean	1514.964516
std	3520.601827
min	2.000000
25%	92.250000
50%	332.500000
75%	1210.500000
max	27640.000000

토픽 모델링(LDA 모델링)

- 모델링 하기 전에 앞서 전처리 한 리뷰 내용들을 doc 2 bow(bag of word) 메소드를 이용하여 문서 데이터를 수치화
- LSI,LDA 모델링을 사용하여 추출된 토픽과 원본 리뷰데이터를 명사만 추출한 데이터들과 비교하여 연관성 파악 후 추출 성능이 더 좋은 LDA 모델링으로 선정
- LDA 모델에서 명소별 Topic의 일관성 점수가 가장 높은 최적의 토픽 집합의 수를 구하기 위한 함수를 생성하여 최적의 토픽집단 142개 선정
- 선정된 주제토픽 단어들을 word2vec를 사용하여 벡터화 한 후 토픽집단끼리의 분포와 응집성을 확인하여 모델링이 잘됐는지 판단

- 모델링 하기 전에 앞서 전처리 한 리뷰 내용들을 doc 2 bow(bag of word) 메소드를 이용하여 문서 데이터를 수치화

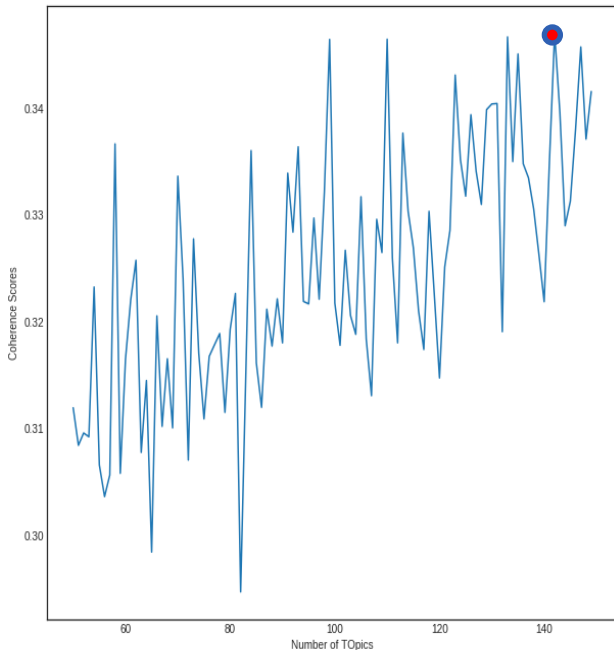
-LSI, LDA 모델링을 사용하여 추출된 토픽과
원본 리뷰데이터를 명사만 추출한 데이터들과
비교하여 연관성 파악 후 추출 성능이 더 좋은
LDA 모델링으로 선정

-LDA 모델에서 명소별 Topic의 일관성 점수가 가장 높은 최적의 토픽 집합의 수를 구하기 위한 함수를 생성하여 최적의 토픽집단 142개 선정

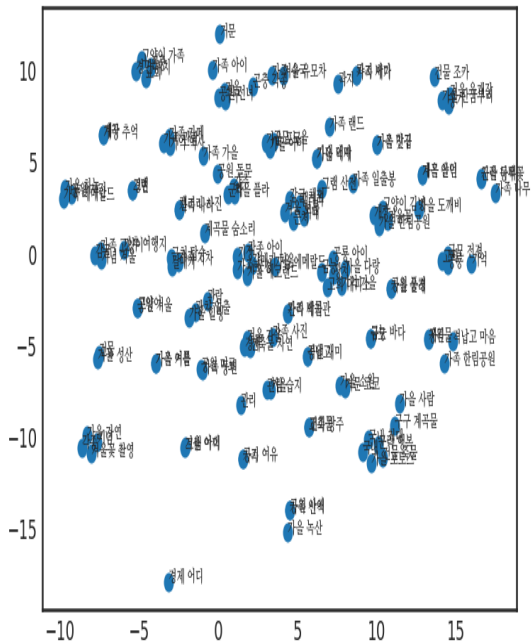
-선정된 주제를 토픽 단어들을 word2vec를 사용하여 벡터화 한 후 토픽집단끼리의 분포와 응집성을 확인하여 모델링이 잘됐는지 판단

04-7 명소 추천 시스템

생성된 TF-IDF벡터의 표본



토픽 시각화



04-7 명소 추천 시스템

유사도 분석 및 추천

-추출된 각각의 142개의 토픽 집단과 유사한 각각의 명소끼리 매칭시켜 데이터 프레임화 하여 저장

-입력된 문장에서 형태소 분석을 거쳐 추출된 명사와 저장된 각각의 명소 별 주제토픽 과의 자카드 유사도식을 사용하여 명사를 비교한 후 가장 유사도가 높은 명소들을 추천

명소 별 선정된 토픽

	name	keyword	address1	address	Latitude	Longitude	content	Topic_Perc_Contrib	Dominant_Topic	Keywords
0	가미물	레저 체험	남원읍	제주특별자치도 서귀포시 남원읍 남원리	33.289973	126.693528	월 감귤대산블루베리 감귤밭 모노레일 감귤 비누만들기 동물먹이주기 알찬체...	0.9549	94.0	감귤, 체험, 동물, 토끼, 농장, 곤충, 각종, 글, 아이, 밥, 비누, 택배, ...
1	거문오름 세계자연유산	자연	조천읍	제주특별자치도 제주시 조천읍 선흘리	33.457031	126.714300	유네스코 자연유산을 직접 체험하는 기회가 있어 좋습니다. 딱 좋은 기분은 거기까지 ...	0.8898	16.0	거문, 코코, 롱, 자연, 아이, 코스, 유산, 파크, 코, 예약, 유네스코, 용암...
2	거센새미오름	자연	구좌읍	제주특별자치도 제주시 구좌읍 송당리 산	33.450575	126.759182	비가 와서 오름 입구조차 찾기 힘들었고 쌍둥이형과 해냈지만 어려움속에서 만난 거센새...	0.9852	32.0	카드, 새미, 라인, 짚, 느낌, 비, 송당리, 천절, 직원, 서바이벌, 규모, 전...
3	검멀레해변	자연	우도면	제주특별자치도 제주시 우도면 우도해안길	33.507434	126.954269	우도에서 보이는 검멀레해안입니다 보는 순간 입이 떡 벌어졌던 장관 바다와의 조화도...	0.3359	58.0	함덕, 바다, 우봉, 해변, 해수욕장, 우도, 아이, 선인, 사진, 여름, 랍지, ...
4	공천포	자연	남원읍	제주특별자치도 서귀포시 남원읍 신례리	33.299995	126.629794	공천포 남원 공천포올레나 홀로 제주에서 일주일 도보 버스 ...	0.8481	48.0	천포, 용머리, 바다, 코스, 남원, 모래, 영, 하늘, 해안, 산책, 자연, 버스...

04-7 명소 추천 시스템

모델링 결과 값 예시

-입력 문장을 아이와 함께 하는 추억으로 했을시 대표적으로 에코랜드 테마파크



-일제시대때 잔해가 있는 곳으로 입력시 일제시대때 일본이 전초기지로 삼은 알뜨르 비행장과 태평양전쟁을 준비할때 만들어 놓은 갯도 진지가 있는 생이기정이 추천

추천된 명소

```
1 def Jaccard_similarity(doc1, doc2):
2     doc1=set(doc1)
3     doc2=set(doc2)
4     doc2=[x.strip() for x in doc2 if x.strip()]
5     doc2=set(doc2)
6     #d = [i for i,j in zip(doc1,doc2) if i==j]
7     list_2=[]
8     for i in doc1:
9         for j in doc2:
10             #print(i,j)
11             if i==j:
12                 list_2.append(i)
13             #print(list_2)
14     return len(list_2) / len(doc1 | doc2)
15 a=['아이와 함께하는 추억']
16 b=text_preprocessing(a)
17 c=sum(b,[])
18 cb=[]
19 for i in range(0,len(bbb)):
20     Jaccard_similarity(c,bbb['Keywords'][i].split(','))
21     cb.append(Jaccard_similarity(c,bbb['Keywords'][i].split(',')))
22 cb
23
24 hi=max(cb)
25 [i for i, v in enumerate(cb) if v==hi]
26 index_1=cb.index(hi)
27
28 max_value = [i for i, value in enumerate(cb) if value == hi]
29
30 hi=2
31 if hi==1:
32     print(bbb.iloc[index_1])
33 elif hi>1:
34     print(bbb.iloc[max_value])
35 elif hi==0:
36     print("비슷한 관광지가 없습니다.")
```

	name	Keywords
22	덕적초등학교 구 덕적본교	본교, 덕적섬오름, 사진, 가을, 유적지, 학교, 역사, 항물, 추억, 친구, 해바...
26	마라도성일	가을, 배교랜드, 선녀, 나무, 추억, 아이, 가족, 기차, 한일공원, 섬, 마라도...
58	선녀와나무꾼	가을, 배교랜드, 선녀, 나무, 추억, 아이, 가족, 기차, 한일공원, 섬, 마라도...
62	에코랜드 테마파크	에코랜드, 기차, 사진, 가을, 아이, 테마파크, 데마, 열매, 자갈, 역, 추억...
77	재작고남미로공원	공원, 미로, 집념, 에코랜드, 미, 고령미, 가족, 사진, 가을, 커피, 아이...
110	제주여빈동산	여빈, 동산, 본교, 사진, 아간, 아이, 행복, 학교, 초승, 개장, 추억, 친구...

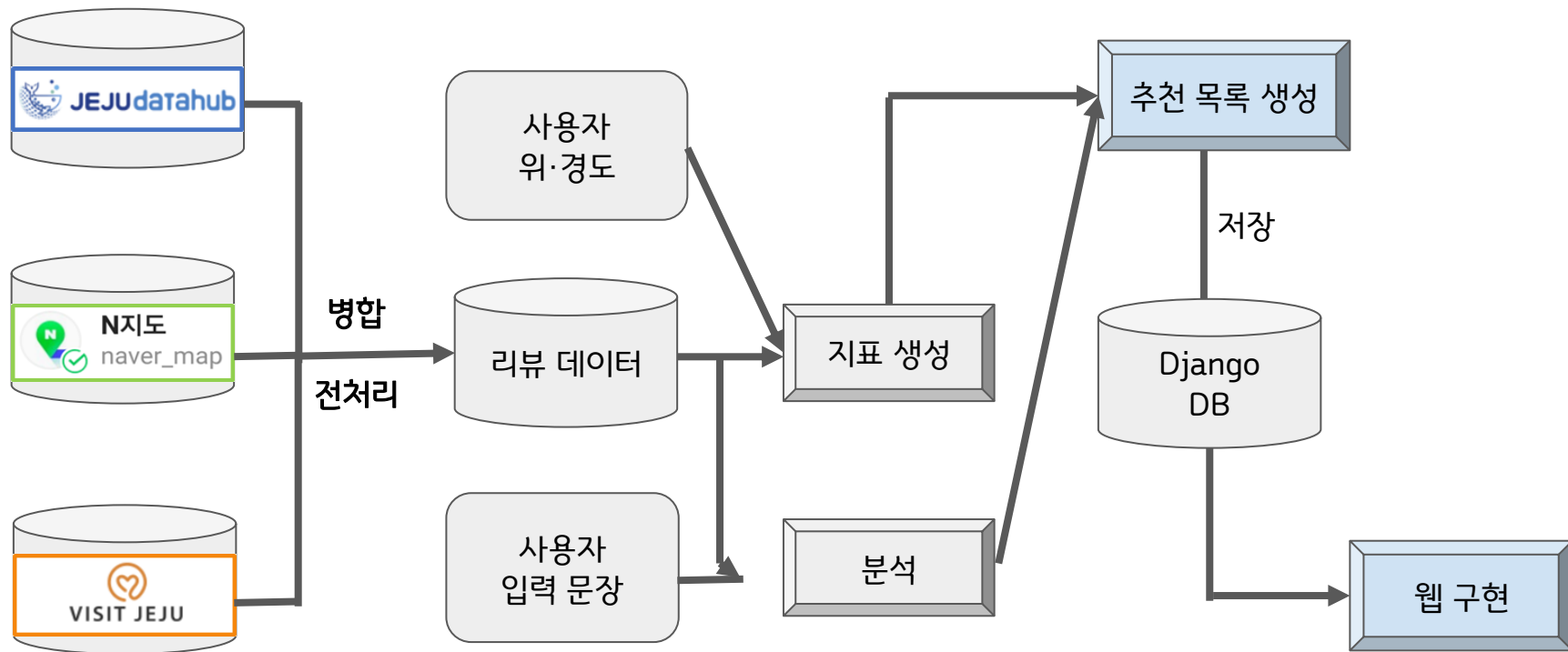
[6 rows x 10 columns]

```
1 def Jaccard_similarity(doc1, doc2):
2     doc1=set(doc1)
3     doc2=set(doc2)
4     doc2=[x.strip() for x in doc2 if x.strip()]
5     doc2=set(doc2)
6     #d = [i for i,j in zip(doc1,doc2) if i==j]
7     list_2=[]
8     for i in doc1:
9         for j in doc2:
10             #print(i,j)
11             if i==j:
12                 list_2.append(i)
13             #print(list_2)
14     return len(list_2) / len(doc1 | doc2)
15 a=['일제시대때 잔해가 있는곳']
16 b=text_preprocessing(a)
17 c=sum(b,[])
18 cb=[]
19 for i in range(0,len(bbb)):
20     Jaccard_similarity(c,bbb['Keywords'][i].split(','))
21     cb.append(Jaccard_similarity(c,bbb['Keywords'][i].split(',')))
22 cb
23
24 hi=max(cb)
25 [i for i, v in enumerate(cb) if v==hi]
26 index_1=cb.index(hi)
27
28 max_value = [i for i, value in enumerate(cb) if value == hi]
29
30 hi=2
31 if hi==1:
32     print(bbb.iloc[index_1])
33 elif hi>1:
34     print(bbb.iloc[max_value])
35 elif hi==0:
36     print("비슷한 관광지가 없습니다.")
```

	name	Keywords
54	생이기정	비행장, 알뜨르, 역사, 장소, 사건, 격납고, 마을, 기억, 가을, 포구, 일제...
79	알뜨르비행장	비행장, 알뜨르, 역사, 장소, 사건, 격납고, 마을, 기억, 가을, 포구, 일제...

[2 rows x 10 columns]

04-9 서비스 work-flow



3조 어디가맨 서비스 시연 영상입니다.

05

결론 및 향후과제



● 데이터 수집

제주데이터허브가 제공하는 제주도 음식점 목록 데이터의 갱신 주기를 알기 어려워
향후 각 음식점별 개/폐업 여부를 최신화하여 서비스

● 모델링

맛집 추천 시스템을 구현하는 데 머신러닝 모델에 대한 지식이 부족하여 사용하지 못해 아쉬웠음
향후 모델 가중치를 구하는 과정에서 좀 더 다양한 위치에서 가중치를 구해 성능을 높이는 작업진행

향후 계절적 요소나 여행목적 같은 요소들을 고려하여 모델링을 개선
맛집 추천 시스템과 자연어 모델을 결합하여 현재의 위치와 사용자의 입력 텍스트를 기반으로
맛집과 명소를 동시에 추천할 수 있는 모델링을 생성

● 웹 구현

AWS 서버에 로드밸런서를 생성할 수 없는 환경이라 https 포트를 사용할 수 없었고,
따라서 사용자의 실시간 위치를 수집하는 Geolocation API를 AWS 서버 내에서 사용할 수 없었음
향후 https 포트로 연결할 수 있는 서버에서 배포하여 서비스하는 것이 목표



Q & A





THANK YOU!



06

느린점





김남경

지금까지 해왔던 단편적으로 데이터를 분석하고 모델링 하는 것이 아닌 실제로 구현을 하고 웹페이지까지 만드는 과정에서 서로서로 의견이 조율이 잘돼야 하고 기술 공유가 필요하다는 것을 몸소 체험 하게 된 계기였습니다.

프로젝트를 하며 의사소통의 중요성을 알게되었다. 데이터 크롤링과 전처리를 기다리기만 할게 아니라 샘플 데이터를 받아서 모델링을 하는 것과, 프로토타입 모델을 넘겨줘서 웹 연결을 할 수 있도록 하는 것이 효율적인 진행 방식이고 시간에 쫓기지 않게 된다고 배웠다.



이종민



김용호

이번 프로젝트를 통해 ‘공유’의 중요성을 알게 되었습니다.
팀원 간의 업무 수행 정도를 제대로 파악하지 못해 많은 어려움을 겪었습니다.
다음번에는 협업 Tool을 적극적으로 사용해서 진행 사항을 공유하면서 진행하고 싶습니다.



문준영

깃을 적극적으로 사용해야 된다는 점을 느꼈다. 코드가 조금만 달라져도 실행이 안되는 경우가 많아서 너무 어려웠었다. 또한 백엔드와 프론트엔드끼리 협업의 중요성을 많이 느꼈다.

영상, 사진, 음성등의 빅데이터를 활용하고 싶었으나 실제로 관련 빅데이터들에 대해 접근하기는 매우 어려웠다. 기획, 디자인, 퍼블리싱, 프론트/백엔드 개발, 테스트 순으로 이루어지는 일반적인 SI업무에 대해 직간접적으로 모두 참여할 수 있어서 의미가 있었던 것 같다.



이주남



여정문

수집할 수 있는 데이터가 제한적이어서 아쉬웠습니다. 사용할 수 있는 데이터가 다양했다면 데이터 가공과 적재 과정이 더 풍부했을 것 같습니다. 그럼에도 불구하고 서비스 구현이라는 프로젝트에 참여할 수 있어서 값진 경험이었습니다.