



네이버 블로그 분석을 활용한 서울시 핫플레이스 키워드 분석

팀명 #찾아조

팀원 강지연, 박민재, 장수진

신청일자 2022.05.02

목차

1. 주제 및 선정 배경
2. 데이터 수집 및 전처리
3. 데이터 분석 및 대시보드
4. 분석 결과
5. 결론

01 #찾아조 - Team member

	학과	이름	역할
팀장	휴먼지능정보공학과	박민재	크롤러 개발, 데이터 수집 및 전처리, 분석, 대시보드 구상
팀원1	휴먼지능정보공학과	강지연	크롤러 개발, 데이터 수집 및 전처리, 분석
팀원2	경제금융학부	장수진	크롤러 개발, 데이터 수집 및 전처리, 분석

01 주제 및 선정배경

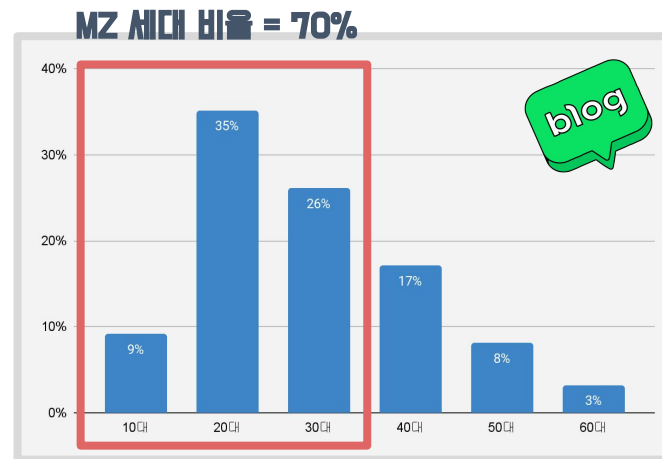
주제 : 네이버 블로그 분석을 활용한 서울시 핫플레이스 키워드 분석

분석 대상

- 분석 타겟 : MZ 세대
- 분석 대상 : 네이버 블로그
- 분석 지역 : 가로수길, 성수동, 연남동, 이태원, 익선동

분석 목표

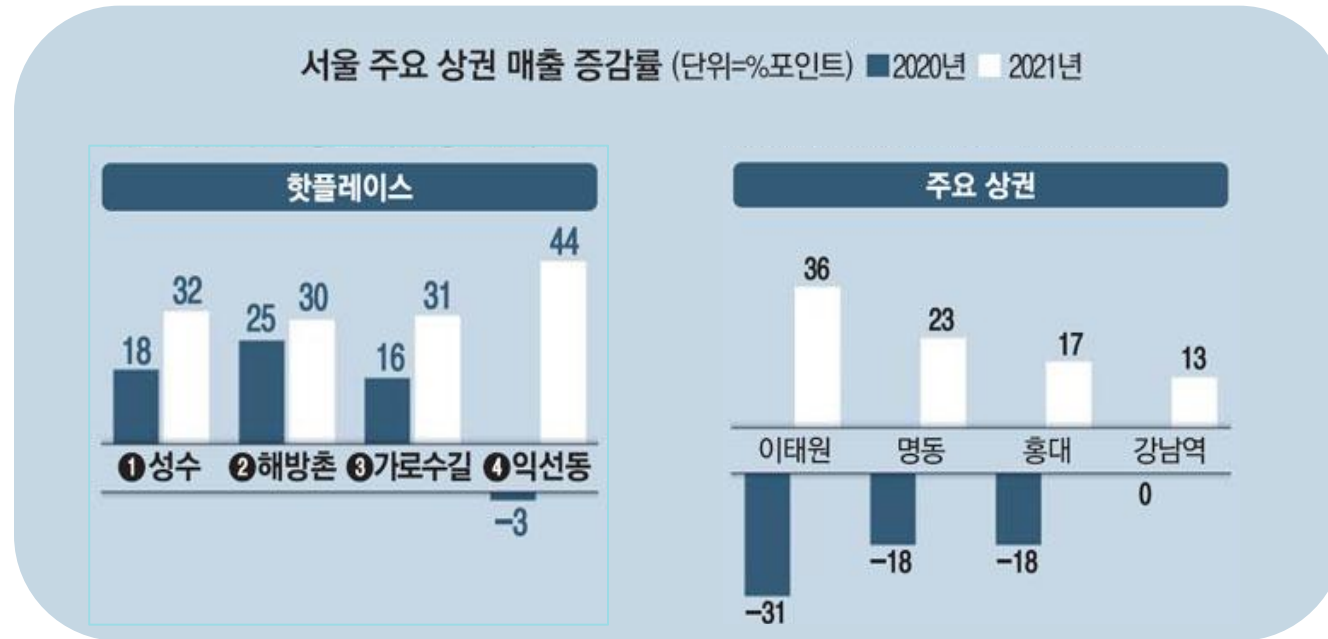
- 관련도가 높은 순으로 데이터를 수집하여 블로그 포스팅 내용을 통해 키워드 추출
- 핫플레이스 별 키워드 분석을 통한 트렌드 추출
- 사용자가 원하는 특정 정보 제공 => 대시보드를 통한 정보 필터링



〈 2021년 블로그 사용자 연령대 비율 〉

01 주제 및 선정배경

주제 선정 배경



코로나 19 완화 이후
주요 상권 지역의 매출 **급격히 증가**

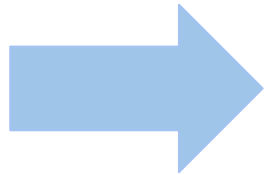


주요 상권 지역의
트렌드 or 상권 분석 필요

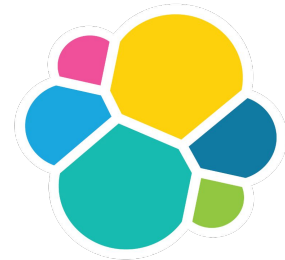
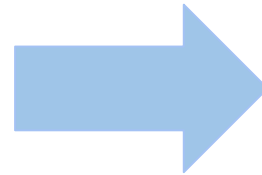
02 데이터 수집 및 전처리

데이터 수집

- 수집 기간 : 2021년 7월 ~ 2022년 7월
- 수집한 데이터양 : 총 234,769건

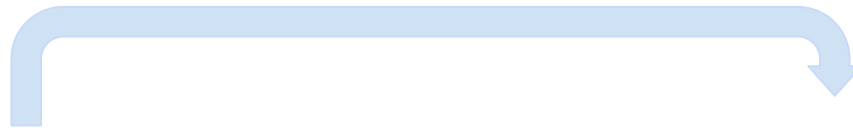


Selenium



elastic

연동



크롤링



포스팅 제목에 핫플레이스명이 포함된 경우만 크롤링

수집 및 적재

02 데이터 수집 및 전처리

데이터 전처리



elastic



Index



places

총 200,000건의
데이터 수집

필드 명(Field)	타입(Type)
place (핫플레이스명)	keyword
title (제목)	text
nickname (작성자)	keyword
date (작성일자)	datetime
content (내용)	text
spot (장소명)	keyword
address (주소)	keyword
location (위도,경도)	geo_point
like (공감 수)	integer
url (포스팅 링크)	keyword
tags (태그)	text
tag_list (태그 리스트)	keyword



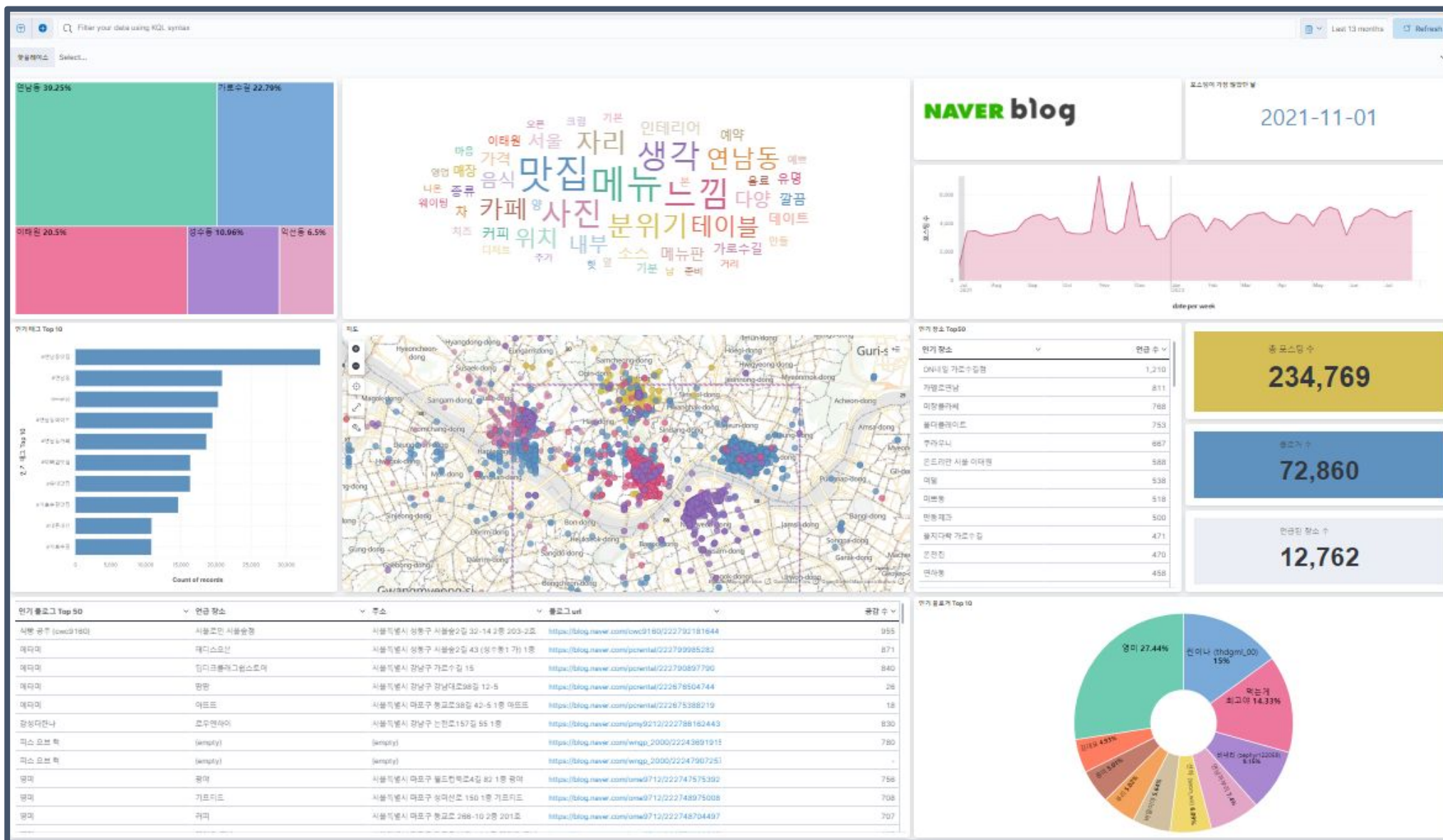
elasticsearch의
한글 형태소 분석기
nori tokenizer 적용



ingest pipeline 통해
list 형식으로 변환

03 테이터 분석 및 대시보드

대시보드

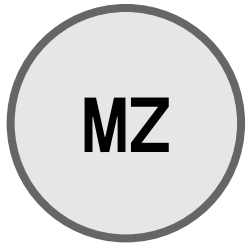


03 데이터 분석 및 대시보드

가설 설정



가설 1. 접근성이 좋을수록 방문객(=포스팅)이 많을 것이다.



가설 2. MZ 세대는 카페나 식당의 분위기, 인테리어를 중요시 할 것이다.

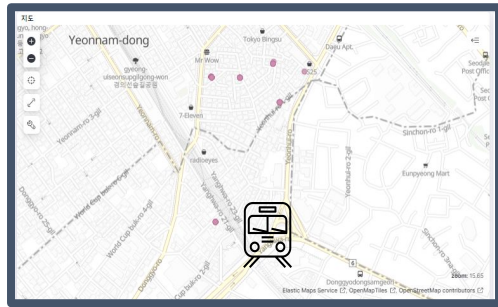


가설 3. 전년도 대비 올해 핫플레이스 방문객이 증가했을 것이다.

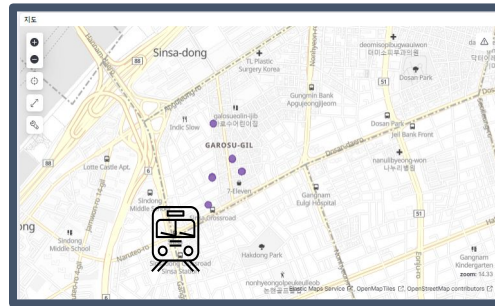
03 데이터 분석 및 대시보드

가설 검증

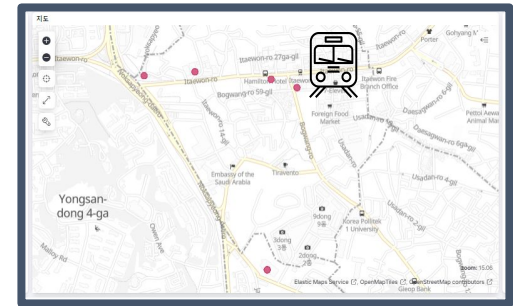
- 핫플레이스별 인기 장소 상위 5개의 위치 분석



〈 연남동 〉



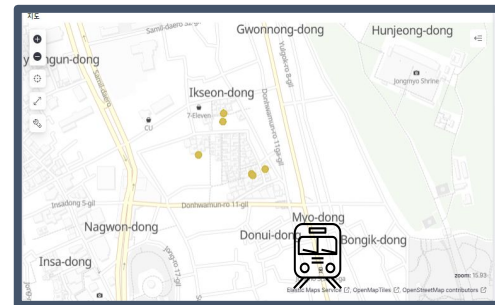
〈 가로수길 〉



〈 이태원 〉



〈 성수동 〉



〈 익선동 〉

인기 장소는 대부분 접근성이 좋은 곳에 위치한다.

03 데이터 분석 및 대시보드

가설 검증

- ◆ "런던 그대로 옮겨온 듯"...2030세대 SNS 인증샷 가득
- ◆ 음식 맛에 더해 '매장 디자인' 중요해져...색다른 경험 제공해야



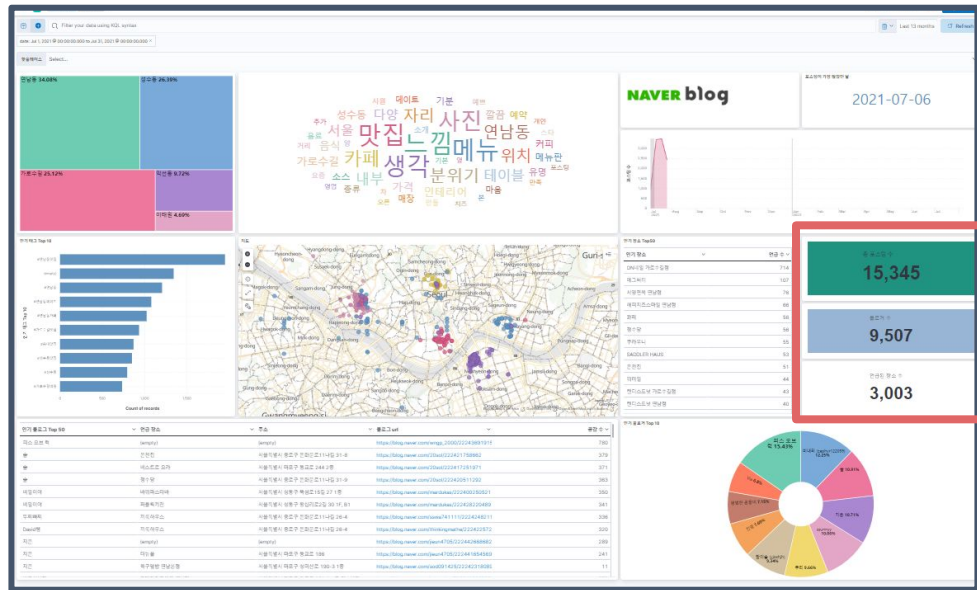
〈워드 클라우드 분석을 통한 주요 키워드〉

맛집, 느낌, 사진, 분위기, 인테리어

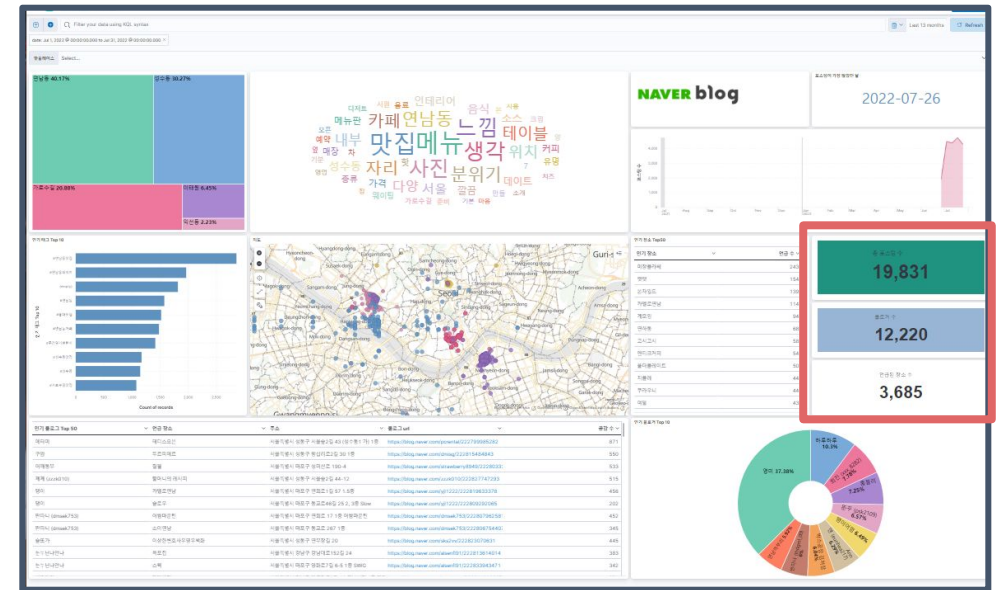
MZ 세대는 카페나 식당의 분위기나 인테리어를 중요시한다.

03 데이터 분석 및 대시보드

가설 검증



〈 2021년 7월 기준 대시보드 〉



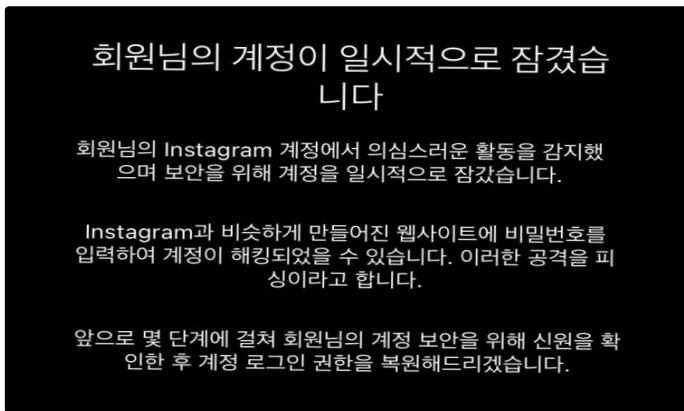
〈 2022년 7월 기준 대시보드 〉

포스팅 수
약 30% 증가

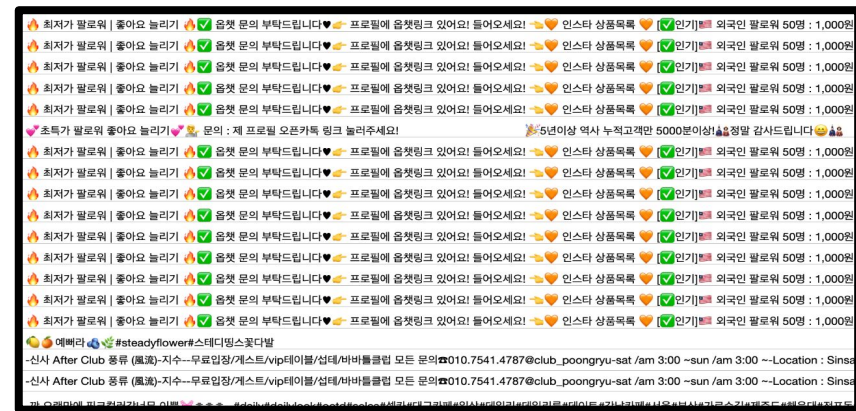
전년도 대비 올해 핫플레이스 방문객이 증가하였다.

트러블슈팅

- 기존 분석 대상이었던 인스타그램의 크롤링 과정에서 계정 차단, 광고성 게시물 문제가 발생하여 네이버 블로그로 대상 변경



〈 계정 차단 및 제한 〉



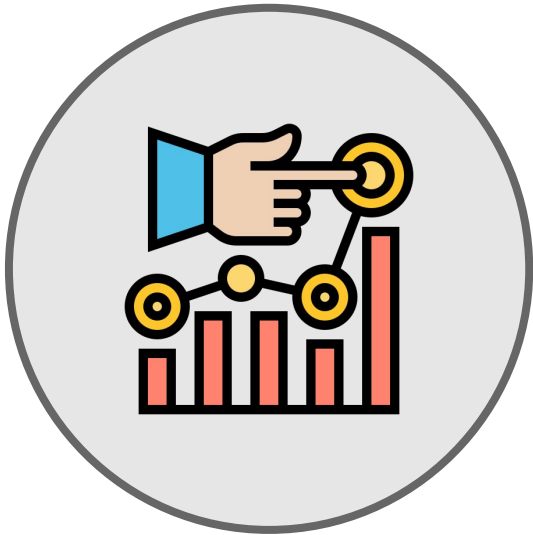
〈 수집한 데이터 내 다량의 광고성 게시물 〉

한계점

- 데이터의 양이 방대하여, 최근 1년 간의 데이터만 수집하였다.
 - 5개의 핫플레이스 지역 1년 데이터 크롤링 결과 => 약 233,449건

05 결론

향후 계획



과거 데이터 + 실시간 데이터 수집

MZ 세대가 추구하는 트렌드 정확히 추출



사용자의 데이터 수집

사용자 맞춤형 키워드, 장소 추천 제공

감사합니다.