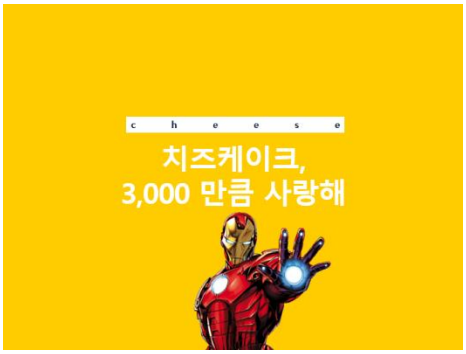


저희 조 이름은 치즈,
프로젝트명은 치즈케이크 3000만큼 사랑해 입니다.
여기서 왜 하필 3,000이라는 숫자인지는 발표를 보시면 알 수 있습니다.

발표는 다음과 같이 진행됩니다.
첫번째로 분석 목적과 데이터 출처를 소개하고,

두번째로 웹스크래핑, 분석/시각화와 결론,

세번째로 한계 및 보완점 으로 마치겠습니다

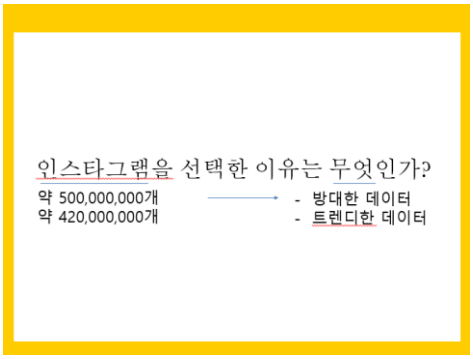
저희 치즈 조 팀원을 소개하겠습니다.
~
저희 팀원들이 치즈케이크를 좋아해서 분석 목적은 유명 치즈 케이크 맛집 찾기 입니다.

데이터출처는 인스타그램에서 #치즈케이크 해시태그로 검색하여 최근 게시물 데이터를 확보했습니다.

글 작성날짜, 해시태그, 좋아요 수, 위치, 사용자 아이디를 수집하였고요

데이터 샘플 수는 3159개,
데이터 날짜 범위는 최근 게시물만 검색되다보니 12월 9일부터 12월 31일까지의 데이터를 수집했습니다.

인스타그램을 선택한 이유는 무엇인가?



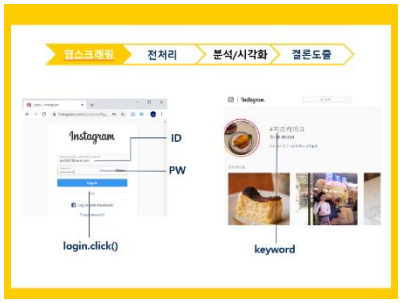
저희가 인스타그램을 선택한 이유는,
인스타그램에
하루에 올라오는 게시글 수 5억 개,
하루에 좋아요 수 4억 개로

방대하고 최신 경향을 반영하기 때문에 인스타그램을 선택

인스타그램의 트렌디하고 방대한 데이터를 수집하고,
웹스크래핑&전처리를 수행해서

크게 두 가지로 분석하려 합니다
치즈케이크 가게의 분포를 지도에 표시하는 것과
키워드를 분석한 뒤,
이를 시각화하고 결론을 도출할 것입니다

웹스크래핑_사전환경설정
웹스크래핑, 전처리, 분석 및 시각화하여 결론을 도출할 것인데요.
첫번째로 웹 스크래핑 단계입니다.
사전환경 설정 부분이구요
로그인에 필요한 아이디와 비밀번호, 검색 키워드, 검색량 등을 변수로 받습니다.



크롤링 함수를 크게 두 파트로 나누어 설명드리겠습니다.

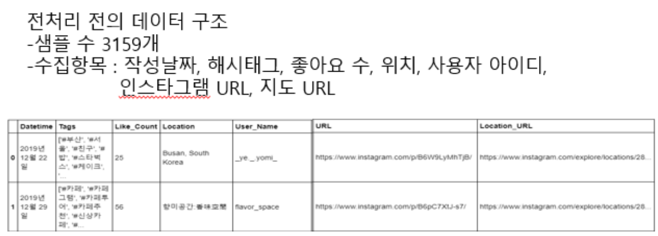
첫번째 함수는 InstagramUrlFromKeyword 라는 함수이고 브라우저, 키워드, 검색량을 변수로 받는 함수입니다.

로그인하고 해시태그 검색해서 나오는 목록에서 글 url을 가져오는 작업입니다.
로그인을 한 뒤 키워드로 검색하면 나오는 글 목록에서 글 주소를 가져왔습니다.

두번째 함수는 IdHashTagFromInstagram 라는 함수이고 브라우저, url을 변수로 받아 수집할 정보들을 리턴합니다.

첫번째 함수에서 수집한 글 url 리스트를 가져와 글을 열어서
사용자아이디, 태그된 위치, 해시태그, 좋아요 수를 수집했습니다.
사실 저희의 목표 샘플수는 10000개였으나,
인스타그램의 검색량 제한으로 중간에 차단되어 멈췄습니다.
따라서 크롤링 과정을 2단계로 나누어 네 명이서 분담하여 크롤링을 수행했습니다..
시간은 한명당 넉넉잡아 2시간 정도 걸렸던 것 같습니다.

#4단계:크롤링 시행
크롤링을 시행해서 새로운 데이터프레임에 데이터를 저장하는 과정입니다.
(+시연) 지금까지 설명한 크롤링 과정을 직접 보여드리겠습니다.
크롤링 과정이 시간이 오래 걸려서 지금 미리 실행하겠습니다.
((키워드 받아서 시연))



전처리 전의 데이터 구조
-샘플 수 3159개
-수집항목 : 작성날짜, 해시태그, 좋아요 수, 위치, 사용자 아이디,
인스타그램 URL, 지도 URL

	Datetime	Tags	Like_Count	Location	User_Name	URL	Location_URL
0	2019년 12월 22일	['#부산', '#서울', '#친구', '#밥', '#스타벅스', '#케이크', '...	25	Busan, South Korea	_ye._yomi_	https://www.instagram.com/p/B6W9LyMhTjB/	https://www.instagram.com/explore/locations/28...
1	2019년 12월 29일	['#카페', '#카페그램', '#카페투어', '#카페추천', '#신상카페', '#...	56	향미공간:[illegible]	flavor_space	https://www.instagram.com/p/B6pC7XtJ-s7/	https://www.instagram.com/explore/locations/28...

전처리 전의 데이터 구조입니다.
샘플 수는 3159개이구요
수집 항목 : 글 작성날짜, 해시태그, 좋아요 수, 위치, 사용자 아이디, 글 URL, 지도 URL
참고로) 수집할 때 위치 정보가 입력되지 않은 글은 저장되지 않습니다.

지도 API를 이용하여 위치정보 텍스트를 위도, 경도로 변환

```
#Location 데이터를 리스트로 가져오기
import pandas as pd
import folium
import numpy as np

insta_df_final.drop(columns = 'Unnamed: 0', inplace = True)

addr_list = []
for k in insta_df_final['Location']:
    addr_list.append(k)
```

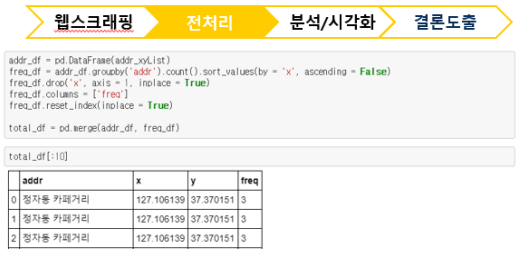
```
addr_list[:20]
```

```
['정자동 카페거리',
 '할리스커피(Hollys Coffee)',
 '안산 유달리',
 '한스오븐 HAN' s OVEN',
 '오지힐 한남',
 '소나비',
 'I Am Autumn',
 '버터힐',
 '휘게-HyggeCafe',
 '제주카페 스르륵',
```

밑에 보이는 리스트는 앞서 수행한 크롤링에서 수집한 위치정보 텍스트입니다.

이를 카카오맵 지도 API를 이용하여
위치정보 텍스트를 위도/경도로 변환하여 리스트로 저장했습니다.

#(코드페이지) APP 키를 카카오에서 발급받아서 입력하면 되고요




```
addr_df = pd.DataFrame(addr_xy_list)
freq_df = addr_df.groupby('addr').count().sort_values(by = 'x', ascending = False)
freq_df.drop('x', axis = 1, inplace = True)
freq_df.columns = ['freq']
freq_df.reset_index(inplace = True)

total_df = pd.merge(addr_df, freq_df)
```

```
total_df[:10]
```

	addr	x	y	freq
0	정자동 카페거리	127.106139	37.370151	3
1	정자동 카페거리	127.106139	37.370151	3
2	정자동 카페거리	127.106139	37.370151	3

위치정보 텍스트를 기준으로 위도, 경도, 빈도수를 정리하여 새로운 데이터프레임으로 저장

분석/시각화 과정에서는 크게 두 가지로 분석하려 합니다
지도를 이용한 방법으로 ~
키워드를 이용한 방법으로~

 # 클러스터마커
우선 지도를 이용한 방법으로 Folium을 사용했습니다.
폴리움의 마커클러스터 플러그인을 이용
(지도 링크 클릭)
높은 빈도수:주황색 --> 낮은 빈도수:연두색
지도를 확대하면 자동으로 세분화된다
(보여주기) 세종시 새롬동, 수원시 신동, 화성시 향남

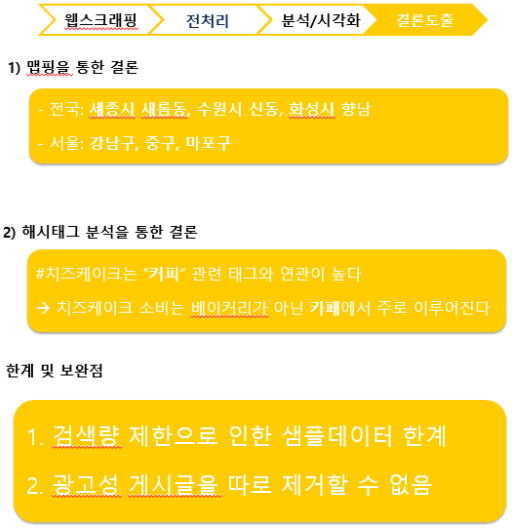
 # 서클마커
원지름의 길이를 빈도수를 반영하여 원을 그리도록 함. 샘플데이터의 한계로 빈도수가 작아서 *3 을 해서 보정했습니다
(지도 링크 클릭)
서울을 중심으로 그려봤는데. 강남, 이태원, 중구, 건대 등이 눈에 띄네요
이렇게 서클마커로 치즈케이크 위치정보의 분포를 한눈에 확인할 수 있었습니다.

 # 히트맵
이제 서울시를 구를 기준으로 구획화해보았습니다
앞서 본 서울에서의 치즈케이크 분포를 구별로 구획화하여 히트맵으로 표현한 것으로
빨간색일수록 빈도수가 높은 것입니다.
(빨간색) 마포구, 중구, 강남구

다음으로 해시태그에 대해 빈도분석을 진행해보았습니다.
수집한 태그 셀이 이렇게 되어있어서 이를 strip을 이용하여 태그와 빈도수를 리스트로 저장했습니다.
#(팝업) 보시면 카페, 마카롱, 딸기케이크, 스콘... 등이 보이고요
--> 대체로 유의미한 데이터로 보입니다.

이 데이터를 키워드와 관련된 단어를 분석하는 워드클라우드를 적용하겠습니다.
앞에서 수행했던 태그와 빈도수가 포함된 리스트이구요
워드클라우드 라이브러리를 이용하였습니다. 워드클라우드 모양을 그림파일로 설정할 수 있어서

치즈케이크 조각을 표현해보았고요
치즈케이크로 검색했기 때문에 치즈케이크 빈도가 가장 높고
카페, 커피 등이 보이고
또다른 디저트인 마카롱, 티라미수, 브라우니, 스콘, 마들렌 등도 보입니다.
검색 시기가 연말이라 크리스마스도 많았음



저희는 다음과 같은 결론을 얻었습니다.
~읽기~

[한계 및 보완점]
1 - 인스타그램의 검색량 제한으로 인해, 많은 양을 크롤링하려 하면 중간에 차단되어 멈춰서 3000개가 한계였다.
2 – 바이럴마케팅이 의심되는 결과도 있었는데요... 광고성 게시글을 따로 제거할 수 없는 한계가 있었다.

소감