## **Graduation Project**

Group 5 2021.10.27

201935006 고효진 201533631 김도균 201935144 최윤찬

## **Topics**

- 1) Instagram data crawling
- 2) Data use

## **Crawling target**



트위터 팔로워 크롤링4

https://beomi.github.io/2019/12/22/Crawling\_Twitter\_Following\_3/

정식으로 인증 받은 계정 + 개발자 계정이 아님 + 단일 프로세스←

User id + screen name 가져오기 가능←

1-1.5초에 한번의 요청↩

4

https://m.blog.naver.com/nonamed0000/220912854545

만약 이런식으로 API 등록을 해 키워드 정보를 받아온다면 헌혈 필요자들에게 홍보는 가능할듯..??<

https://m.blog.naver.com/nonamed0000/220912854545

User following – followers 동시에 가져오기←

한시간에 150개의 request←

Conclusion: 트위터 <u>크롤링</u> 역시 <u>팔로워-팔로잉</u> 아이디 이상의 정보를 가져올 수 없을 것 같아서 사용하게 된다면 더미 데이터를 만들때 1차적으로 사용은 가능하겠지만 부가적인 데이터 추가가 필요한 $\phi$ 



#### 크롤링 이슈

하루에 반복적으로 100~300번 크롤링 시, 페이지 차단, 이후 계정 정지될 수도...

- → 하루 100 request -> 864초당 1회←
- → 특정 시간대에 2~3명 분만 정보를 가져옴

#### instaloader <



- https://instaloader.github.io/as-module.html?highlight=followered
- followee(), followers()를 통해 친구망은 구축 가능↔
- Instagram 특성상, feature를 특정할만한 것이 포스트밖에 없음...
- Instagram을 통해서 facebook 아이디 알아낼 수 있는 방법 찾지 못함←

#### 한국어 예시<

https://kminito.tistory.com/13

#### facebook

Facebook api 사용시←

https://developers.facebook.com/docs/graph-api/reference/v12.0/user/friends

이거로 가지고 올수 있는정보가

https://developers.facebook.com/docs/graph-api/reference/user/

이거인데, 성별,위치, 종교 등등 여러 데이터 얻을수 있음↔

(user\_friends permission 획득한 상태로)←

4

근데 위에거는 제한적으로만 사용할수 있으니까 크롤링 해야 하는데,

https://www.bestproxyreviews.com/facebook-scraper/

상당히 어렵다고 함.←

페이스북은 <u>스크래핑에</u> 대처하는 전담 팀이 아주 큰 규모로 있어서 다른 소셜네트워크에 비해 <u>스</u>크래핑이 어렵다함.

4

https://www.octoparse.com/blog/5-things-you-need-to-know-before-scraping-data-from-facebook

4

이거에 의하면 person's name, physical address, email address, phone number, IP address, date of birth, employment info and even video/audio recording. 등 우리가 원하는 개인정보를 스크래핑하면 고소당한다고함.

원하는 feature 데이터를 수집할 때 완전히 불법적으로 해야함.

### Instagram data crawling

#### - Follows

The data of the Instagram account logged in through the Instagram crawler was crawled with a request time of 60 seconds.

As one cycle is performed, "depth" is added.

```
curr_target_list
{'adoles_0110'}

1
2
3
4
heeee_kite->adoles_0110
```

```
curr_target_list
{'heeee_kite', 'geurujam_', 'keepitlowssh'

2

3
4
adoles_0110->heeee_kite
oxoox.j->heeee_kite
```

## Instagram data crawling

### - Follows (Followee)

- id
  - : Target user's id
- name
  - : Target user's name on profile
- bio
  - : Target user's biography on profile
- from
  - : Target user (crawling start point)
- to
  - : Target user's following account
- is\_private
  - : Whether the Target user's account is private or not

	id	name	bio	from	to	is_private
0	adoles_0110	鎭	비공계개정	adoles_0110	heeee_kite	1
1	adoles_0110	鎭	비공계개정	adoles_0110	younghotyellow94	1
2	adoles_0110	鎭	비공계개정	adoles_0110	ph1boyyy	1
3	adoles_0110	鎭	비공계개정	adoles_0110	su_c.haere	1
4	adoles_0110	鎭	비공계개정	adoles_0110	painted_wood	1
5	adoles_0110	鎭	비공계개정	adoles_0110	kimmfall	1
6	adoles_0110	鎭	비공계개정	adoles_0110	nocontextid	1
7	adoles_0110	鎭	비공계개정	adoles_0110	lecinqmailunaire	1
8	adoles_0110	鎭	비공계개정	adoles_0110	youin_jung	1
9	adoles_0110	鎭	비공계개정	adoles_0110	kimyk10	1
10	adoles_0110	鎭	비공계개정	adoles_0110	_ss.ha	1
11	adoles_0110	鎭	비공계개정	adoles_0110	makesblaugh_j	1
12	adoles_0110	鎭	비공계개정	adoles_0110	v_toon_v	1
13	adoles_0110	鎭	비공계개정	adoles_0110	seoulforest_climbing	1
14	adoles_0110	鎭	비공계개정	adoles_0110	volleyball_korea	1
15	adoles_0110	鎭	비공계개정	adoles_0110	suji6620	1
16	adoles_0110	鎭	비공계개정	adoles_0110	kovopr_official	1
17	adoles_0110	鎭	비공계개정	adoles_0110	geurujam_	1
18	adoles_0110	鎭	비공계개정	adoles_0110	mi casa.kr	1
19	adoles_0110	鎭	비공계개정	adoles_0110	fxvnmm	1
20	adoles_0110	鎭	비공계개정	adoles_0110	itdew	1
21	adoles_0110	鎭	비공계개정	adoles_0110	hoohoohoo	1
22	adoles_0110	鎭	비공계개정	adoles_0110	jjuns.y22	1

### Instagram data crawling

### - Followers

- id
  - : Follower's id
- name
  - : Follower's name on profile
- bio
  - : Follower's biography on profile
- from
  - : Follower's account
- to
  - : Target user's account
- is\_private
  - : Whether the Follower's account is private or not

122	nocontextid		밥 먹어라!!!!	nocontextid	adoles_0110	0
123	lecinqmailuna	金旦午	supercalifra	lecinqmailuna	adoles_0110	0
<b>12</b> 4	_ss.ha			_ss.ha	adoles_0110	1
125	makesblaugh <sub>.</sub>	장슨아		makesblaugh_	adoles_0110	1
126	fxvnmm			fxvnmm	adoles_0110	1
127	ever_249	Noah		ever_249	adoles_0110	1
128	1hyn_dred	백현정		1hyn_dred	adoles_0110	1
129	rlo_vely			rlo_vely	adoles_0110	1
130	oesxnue_	은서		oesxnue_	adoles_0110	1
131	sonxgee	강송이		sonxgee	adoles_0110	1
132	ddoonngg_99	눌		ddoonngg_99	adoles_0110	0
133	iluv_yj2			iluv_yj2	adoles_0110	0
134	hy_enii			hy_enii	adoles_0110	1
135	o.zero_	은영		o.zero_	adoles_0110	1
136	holy_evening	정우		holy_evening	adoles_0110	0
137	aheunkim	은ㅏ	@taek_chu	aheunkim	adoles_0110	0
138	c.j.wan	지완		c.j.wan	adoles_0110	1
139	no_omoon			no_omoon	adoles_0110	1
140	aexxrim	愛		aexxrim	adoles_0110	0
141	ratherbe_cool	이윤진		ratherbe_cool	adoles_0110	1

### Data use

:Data collecting

- 1. Merge datasets that collected from each member
- 2. Add dummy data to the merged dataset
- 3. Dataset training

# Thank you