Graduation Project

Group 5
2021.11.18

201935006 고효진
201533631 김도균
201935144 최윤찬

Topics

1) Instagram data crawling
2) Usage of data

Ⅰ. Subject Review

Ⅰ. Overview

Directed Blood Donation

Be designated by someone
who get blood from donor

Ⅰ. Elicitation Result

SNS & Community Service for
Directed Blood Donation

Objective 1
Recommend the blood donor not even closest person but further human connections.

Objective 2
Sort recommendation with using various features.

Objective 3
UI and UX design for massive user and collecting data for ML.

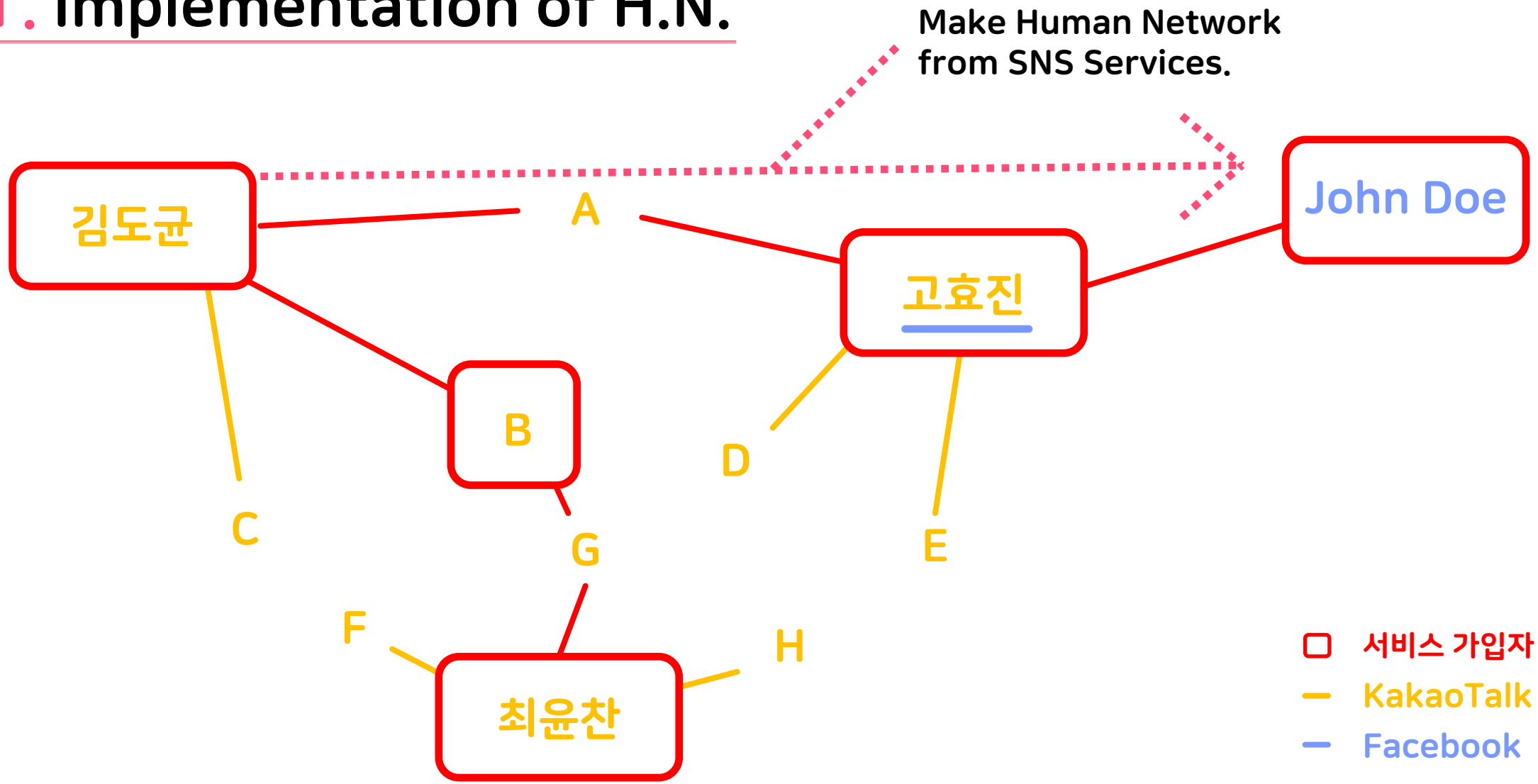
Ⅰ. Human Network

Human Network

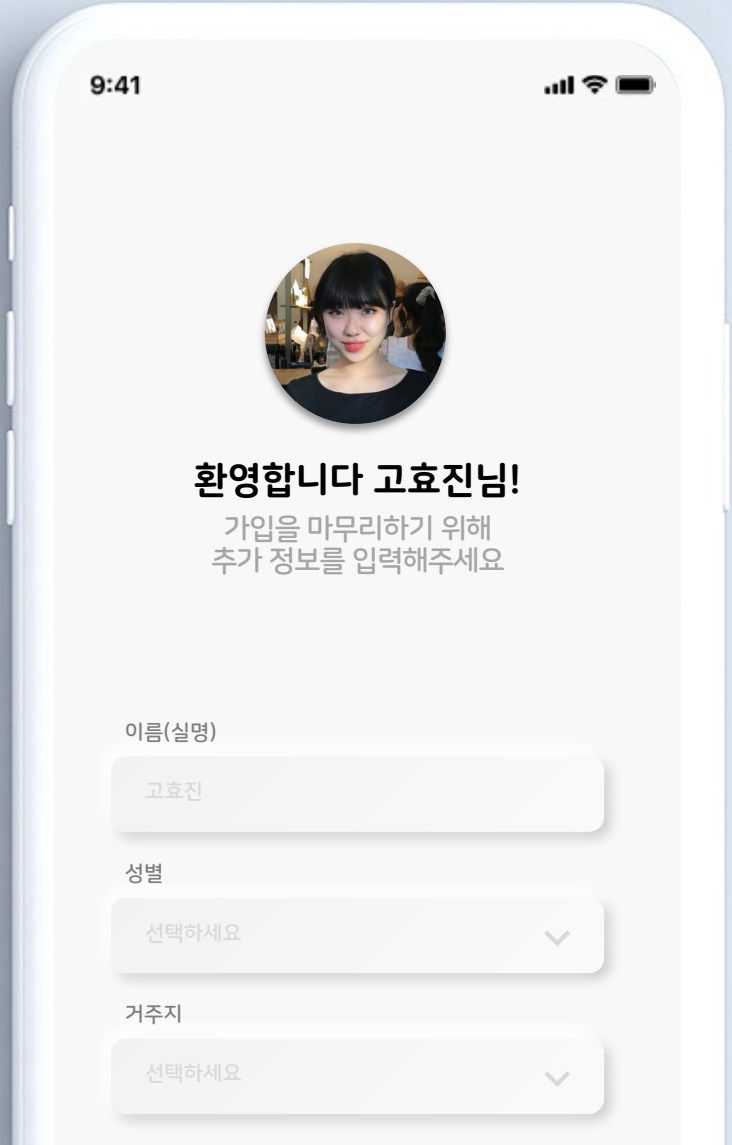
: Recommend the blood donor not even closest person but further human connections.

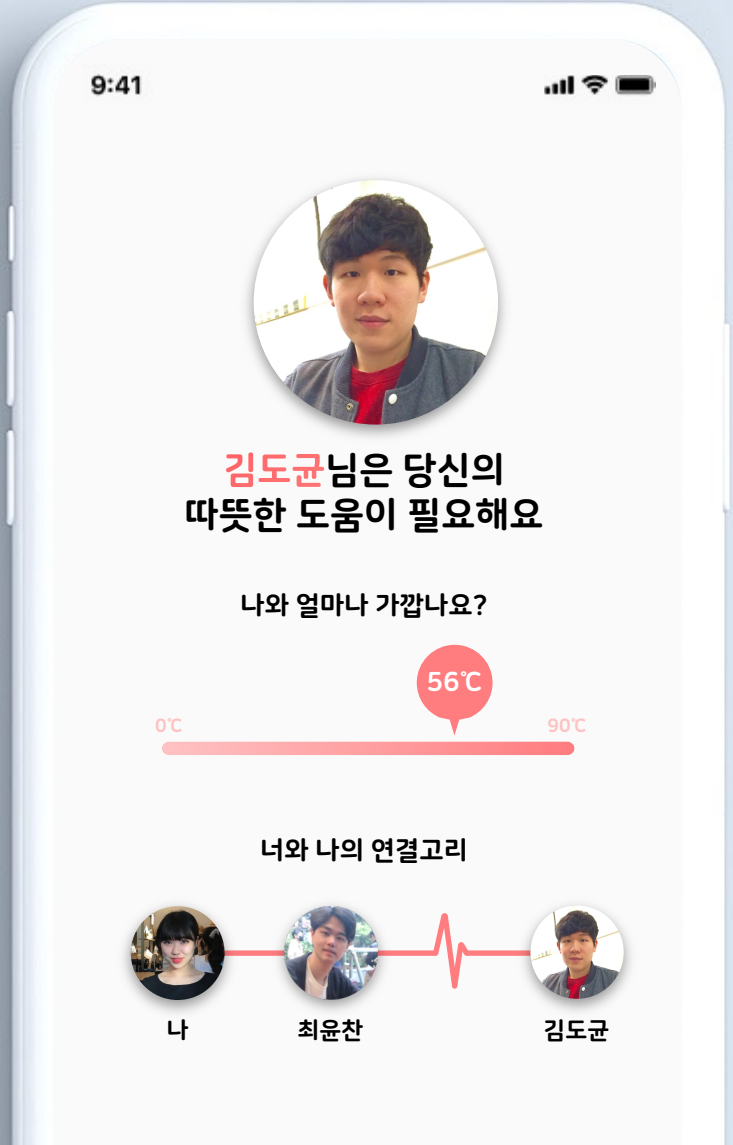
Six Degrees of Kevin Bacon

Ⅰ. Implementation of H.N.



Ⅰ. Samples

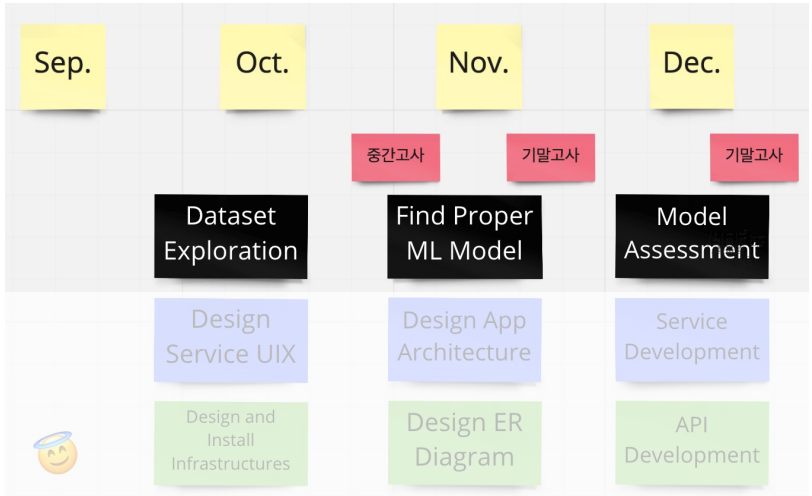

9:41
환영합니다 고효진님!
가입을 마무리하기 위해
추가 정보를 입력해주세요
이름(실명)
고효진
성별
선택하세요
거주지
선택하세요


9:41
김도균님은 당신의
따뜻한 도움이 필요해요
나와 얼마나 가깝나요?
56℃
0℃
90℃
너와 나의 연결고리
나
최윤찬
김도균

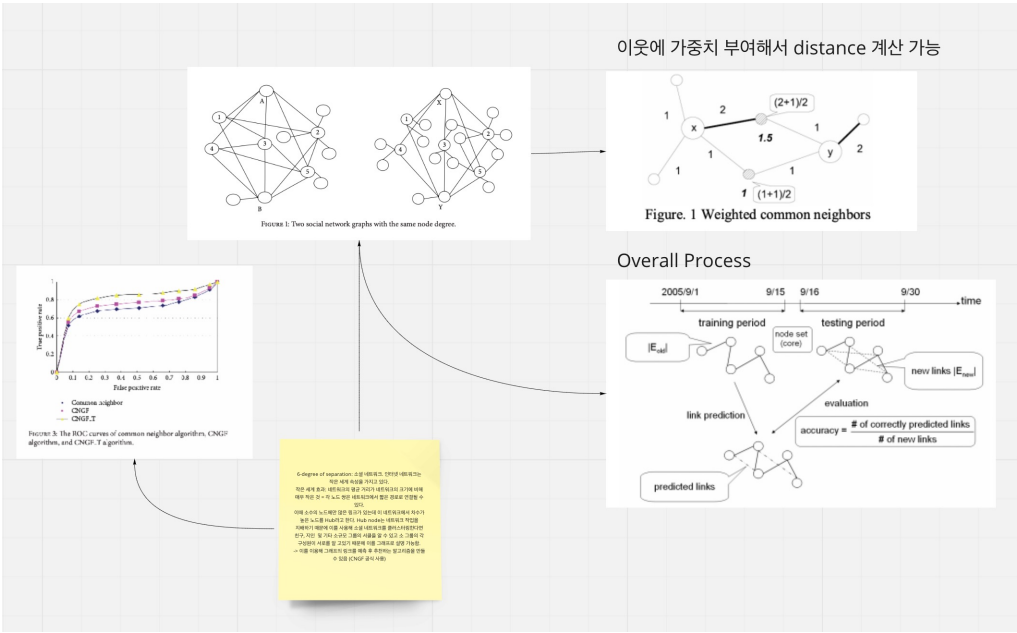
II . Roles

Ⅱ. Roles

고효진



Calculate distance through weights each nodes and predict proper patient.

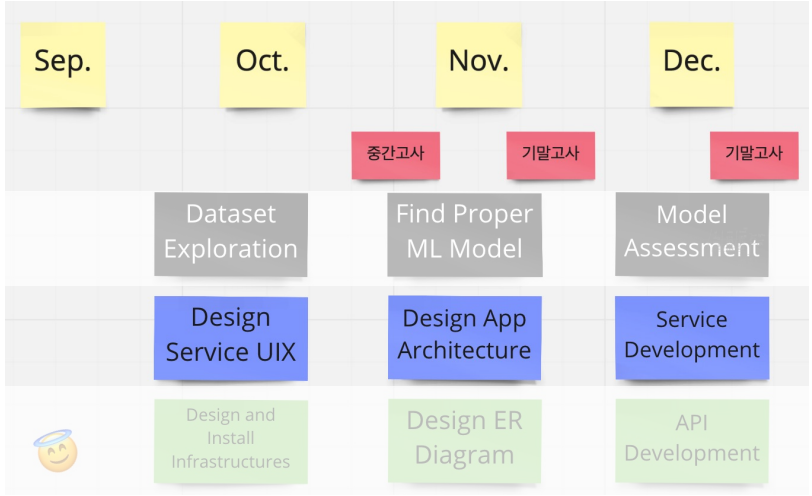


References (Research)

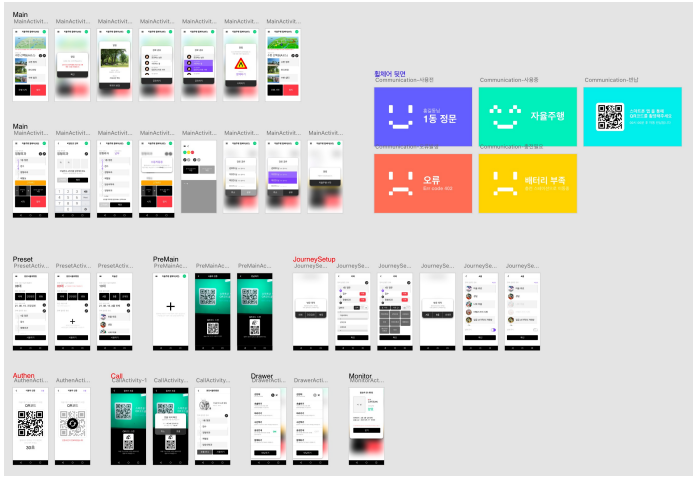
"The Algorithm of Link Prediction on Social Network"
"Link Prediction of Social Network Based on Weighted Proximity Measures"

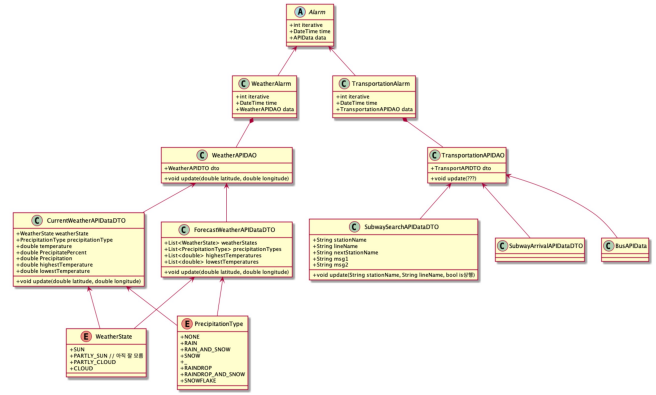
Ⅱ. Roles

김도균



Design Application UI, UX and business logic through Adobe XD and PlantUML.



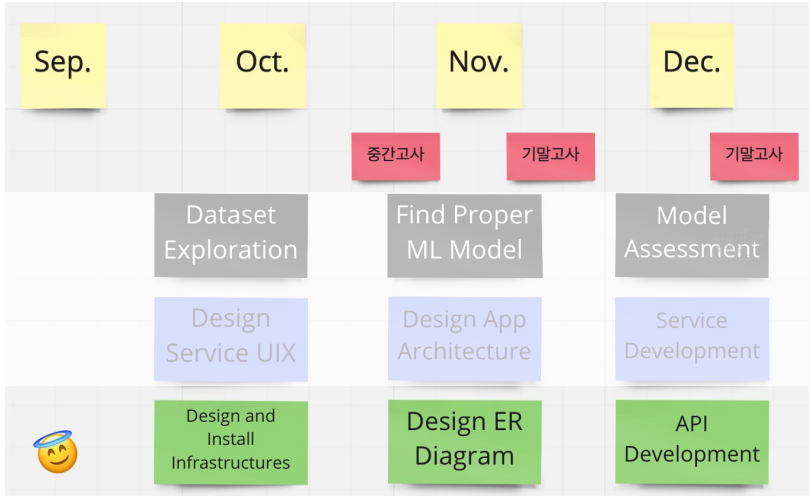


Implement application service with using Flutter cross-platform framework.

Flutter

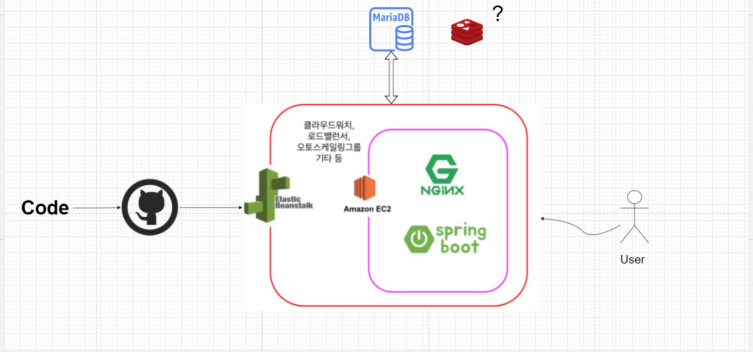
Ⅱ. Roles

최윤찬

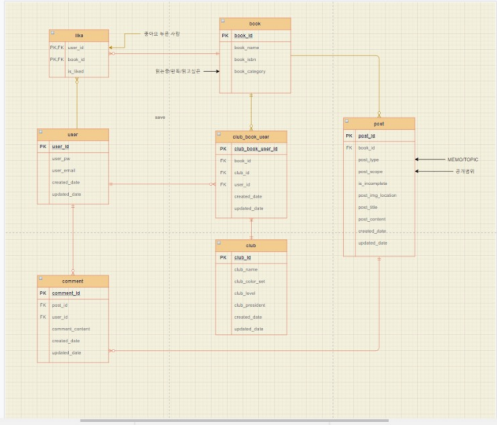


Implement CI/CD pipeline with using beanstalk(PaaS) in AWS.

Use NoSQL (in case)



API and Database design from ER-Diagram

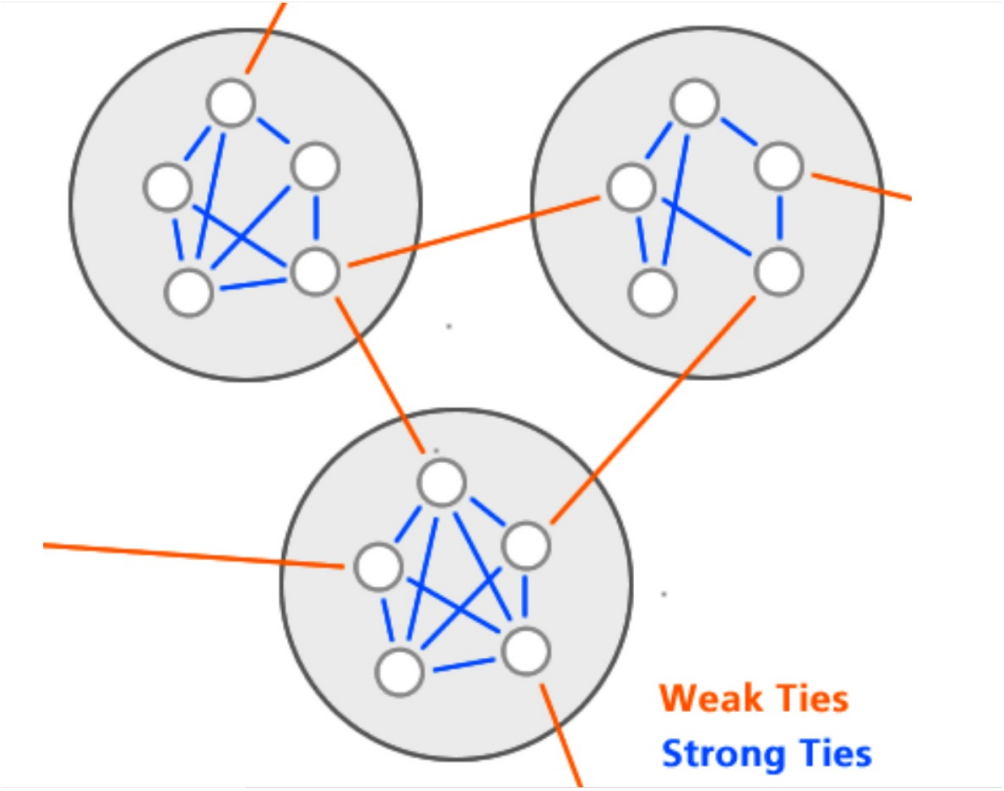


Implement back-end service with using Spring Framework.

III. Data Curation

Ⅲ. Data Structure

:Tie strength



Weak Ties

: Those edges that occur among nodes belonging to different communities.

Strong Ties

: Those edges occurring among nodes in the same community

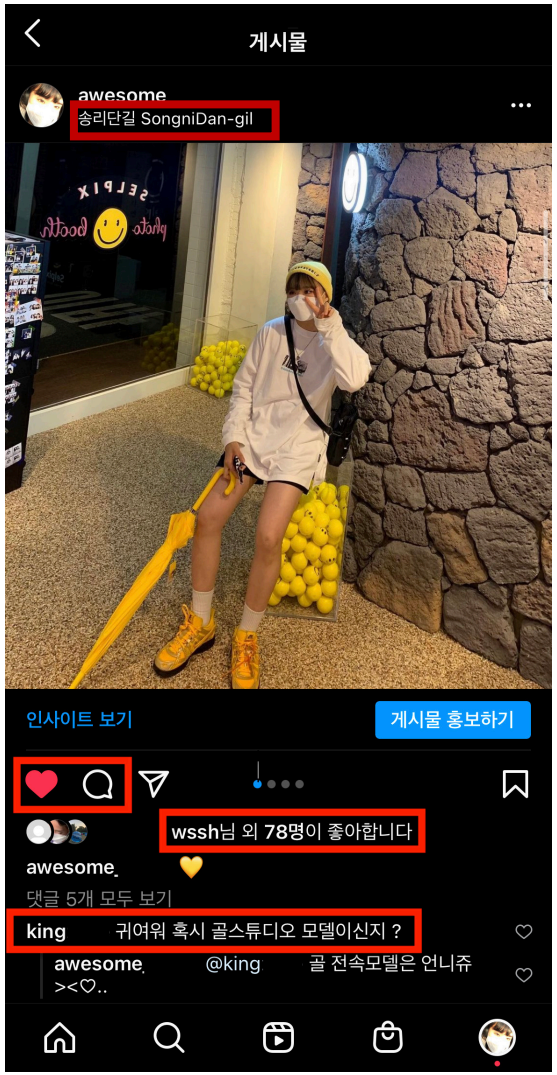
Intensity, Intimacy, Duration, Reciprocal services, Structural, emotional support, Social distance

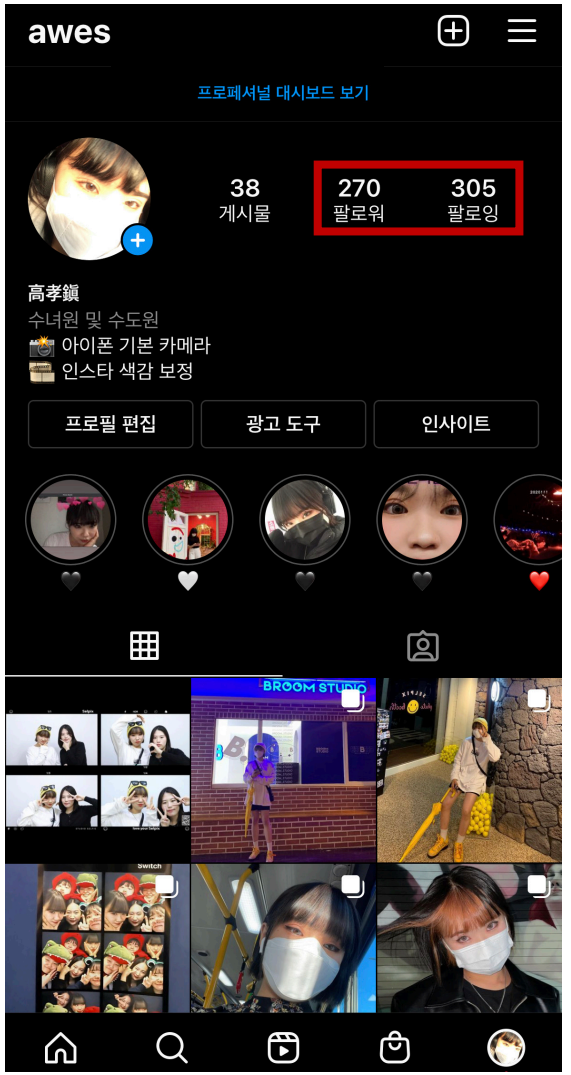
Ⅲ. Data Structure

:Tie strength

Similarity

: Location Preferences, hidden relationship between users

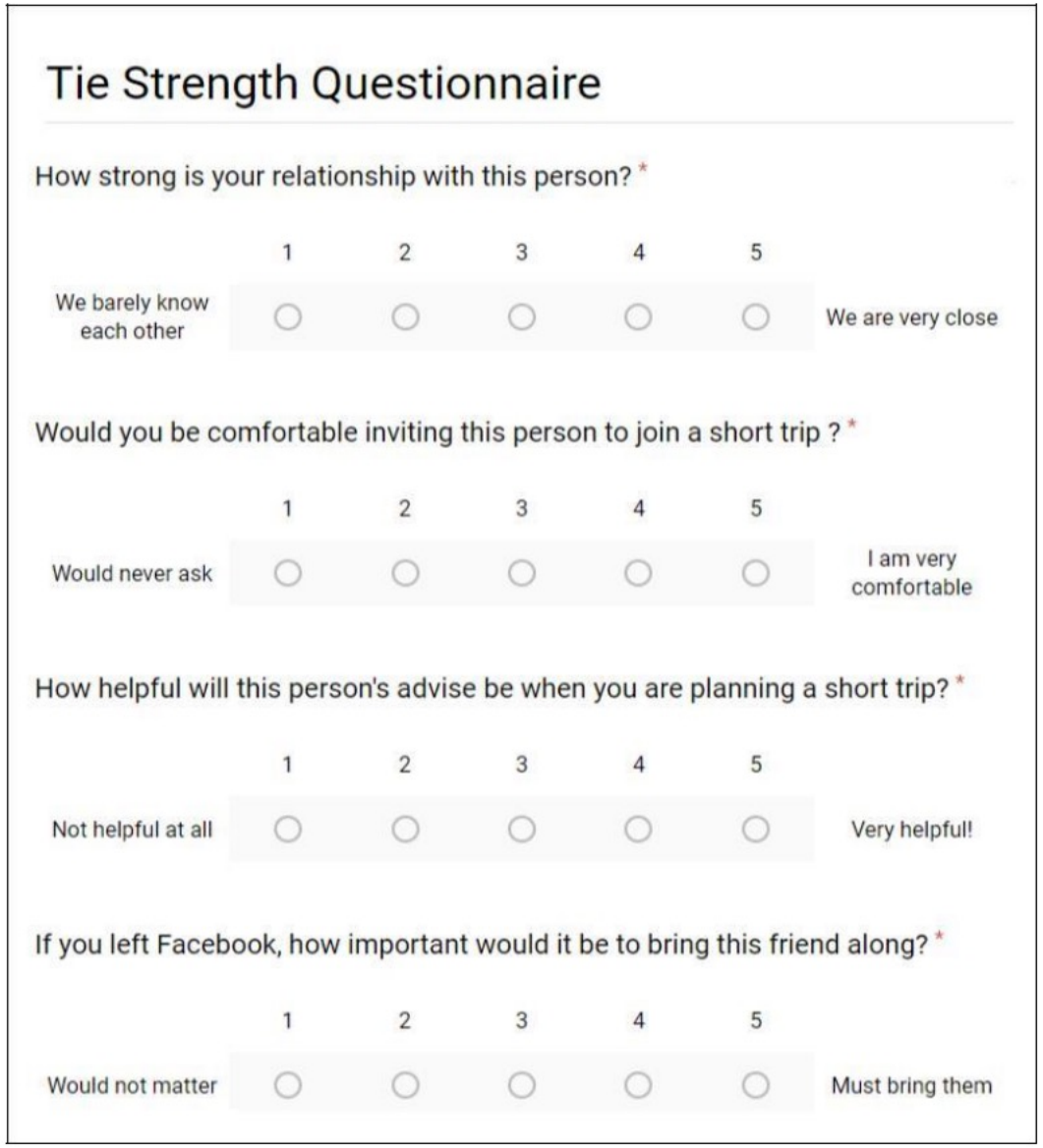




Ⅲ. Data Structure

:Tie strength

Capture the relationship strength between two users and its influence



Tie Strength Questionnaire

How strong is your relationship with this person? *

	1	2	3	4	5	
We barely know each other	○	○	○	○	○	We are very close

Would you be comfortable inviting this person to join a short trip ? *

	1	2	3	4	5	
Would never ask	○	○	○	○	○	I am very comfortable

How helpful will this person's advise be when you are planning a short trip? *

	1	2	3	4	5	
Not helpful at all	○	○	○	○	○	Very helpful!

If you left Facebook, how important would it be to bring this friend along? *

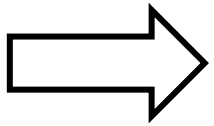
	1	2	3	4	5	
Would not matter	○	○	○	○	○	Must bring them

https://www.researchgate.net/publication/339198907

Ⅲ. Data Structure

:Tie strength

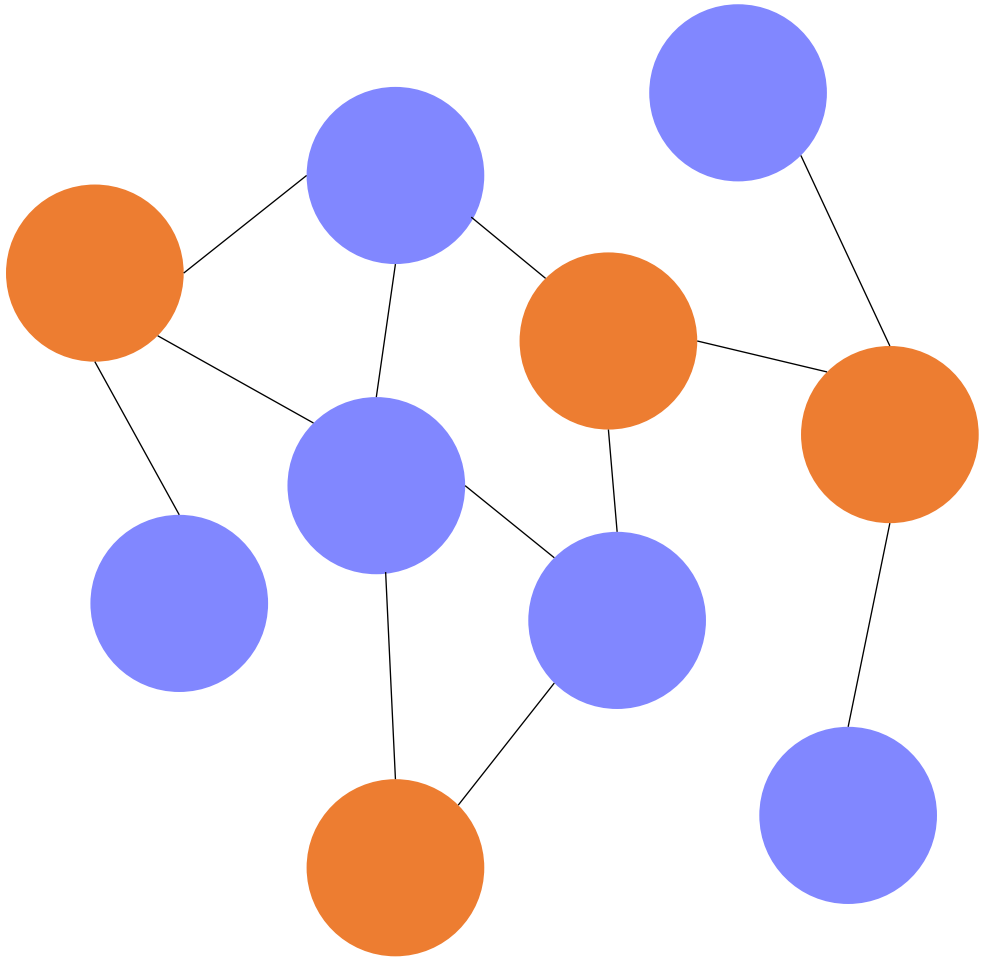
Predictive Variable	Tie Strength Dimension
Wall words	Intensity
Wall posts	Intensity
Number of likes	Intensity
Number of Comments	Intensity
Number of Friends	Intimacy
Locations visited together	Intimacy
Days since last communication	Duration
Number of mutual friends	Structural
Gender Similarity	Social Distance
Age gap (in days)	Social Distance



Variable	Coefficient	Standard Error	p-value	VIF
Gender Similarity	0.3710	0.123	0.003	1.096
Age Gap	0.0643	0.054	0.237	1.029
Wall Posts	0.2367	0.141	0.095	8.176
Mutual Strength	0.1303	0.073	0.076	1.950
Wall Words	0.0667	0.068	0.326	1.494
Locations Together	0.0943	0.102	0.354	4.335
Number of Friends	-0.1393	0.081	0.089	2.359
Number of Likes	0.1945	0.086	0.025	2.782
Days Since Last Communication	-0.1235	0.056	0.029	1.213
Number of Comments	0.2317	0.106	0.029	3.808
R-squared	**0.411**	**Adjusted R-squared**		**0.383**

https://www.researchgate.net/publication/339198907

Ⅲ. Data Structure (Graph)

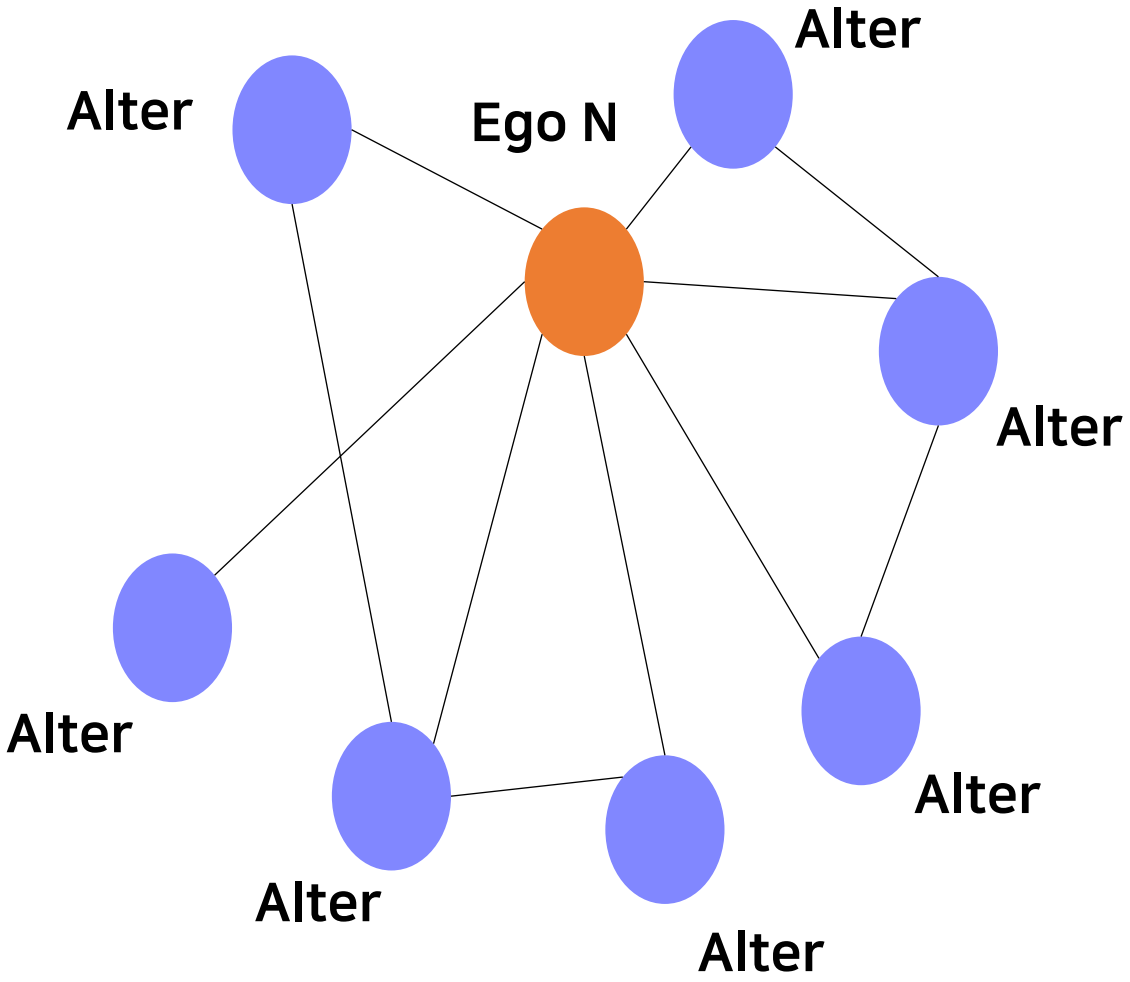


2-mod Network

: Extract 1-mod network from 2-mod network. Vice versa.

https://m.blog.naver.com/khj90733/221668777824

Ⅲ. Data Structure (Graph)



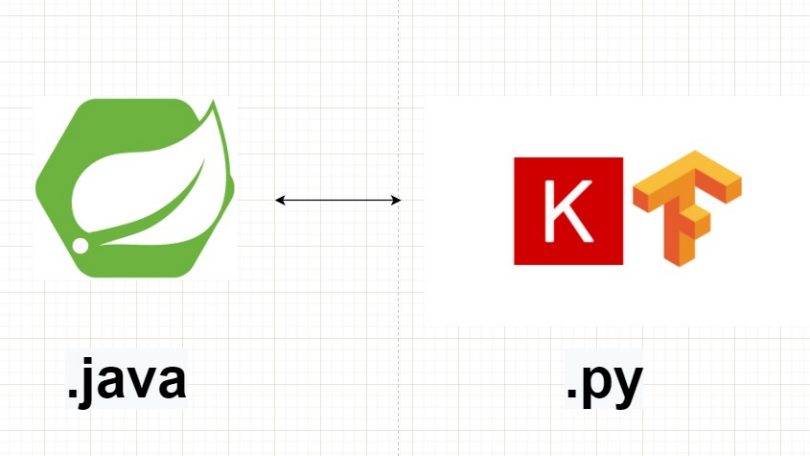
Ego Network

: Expending from Ego Network to full network.
Extract Ego network from Full network

https://m.blog.naver.com/khj90733/221668777824

Ⅲ. Data Structure

Appendix



1. ProcessBuilder API

```
@Test
public void givenPythonScript_whenPythonProcessInvoked_thenSuccess() throws Exception {
    ProcessBuilder processBuilder = new ProcessBuilder("python",
resolvePythonScriptPath("hello.py"));
    processBuilder.redirectErrorStream(true);

    Process process = processBuilder.start();
    List<String> results = readProcessOutput(process.getInputStream());

    assertThat("Results should not be empty", results, is(not(empty())));
    assertThat("Results should contain output of script: ", results, hasItem(
            containsString("Hello Baeldung Readers!!")));

    int exitCode = process.waitFor();
    assertEquals("No errors should be detected", 0, exitCode);
}
```

2. HTTP

> python -m http.server 9000

III. Collecting Data Through Crawling

twitter

트위터 팔로워 크롤링
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

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 크롤링 역시 팔로워-팔로잉 아이디 이상의 정보를 가져올 수 없을 것 같아서 사용하게 된다면 더미 데이터를 만들때 1차적으로 사용은 가능하겠지만 부가적인 데이터 추가가 필요함

Instagram

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


Python Module instaloader

Instaloader exposes its internally used methods and structures as a Python module, making it a powerful and intuitive Python API.

Start with getting an instance of Instaloader:

```
import instaloader

# Get instance
L = instaloader.Instaloader()

# Optionally, login or load session
L.login(USER, PASSWORD)      # (login)
L.interactive_login(USER)    # (ask password on terminal)
L.load_session_from_file(USER) # (load session created w/
                               #  `instaloader -l USERNAME`)
```

Instaloader provides the Post structure, which represents a picture, video or sidecar (set of multiple pictures/videos) posted in a

```
for post in instaloader.Hashtag.from_name(L.context, 'cat').get_posts():
    # post is an instance of instaloader.Post
    L.download_post(post, target='#cat')
```

Post instances can be created with:

- Post.from_shortcode()
  Use a Post shortcode (part of the Post URL, https://www.instagram.com/p/SHORTCODE/) to create a Post object:

```
post = Post.from_shortcode(L.context, SHORTCODE)
```

- Profile.get_posts()
  All media of a Profile.
- Profile.get_saved_posts()
  Media that the user marked as saved (Profile must be own profile for this to work)
- Instaloader.get_feed_posts()
  Media in the user's feed.
- Instaloader.get_explore_posts()
  Media that is suggested by Instagram to explore.
- Hashtag.get_posts()
  Media associated with given hashtag.

With the Profile class, instaloader also makes it easy to access metadata of a Profile. Profile instances can be created with:

- Profile.from_username():

```
profile = Profile.from_username(L.context, USERNAME)
```

- Profile.from_id()
  given its User ID. This allows to easily lookup a Profile's username given its ID:

```
Profile.from_id(L.context, USERID).username
```

- Profile.get_followees()
  Profiles that are followed by given user.
- Profile.get_followers()
  Profiles that follow given user.

- https://instaloader.github.io/as-module.html?highlight=follower
- 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)

근데 위에거는 제한적으로만 사용할수 있으니까 크롤링 해야 하는데,

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

상당히 어렵다고 함.

페이스북은 스크래핑에 대처하는 전담 팀이 아주 큰 규모로 있어서 다른 소셜네트워크에 비해 스크래핑이 어렵다함.

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

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.

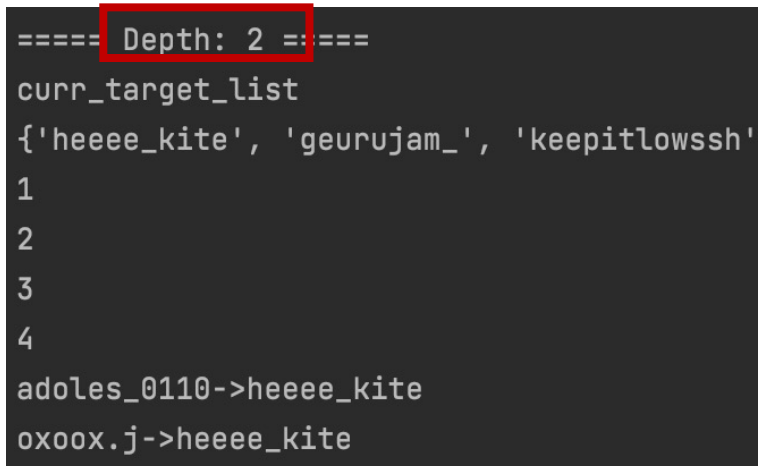
Ⅲ. Data Crawling

Instagram data crawling

- Follows

Crawler crawle Instagram user data with 60 seconds sampling through the temporary Instagram account what logged in.

As one cycle is performed, “depth” is increased.



Ⅲ. Data Crawling

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	it__dew	1
21	adoles_0110	鎭	비공계개정	adoles_0110	hoohoohoo__	1
22	adoles_0110	鎭	비공계개정	adoles_0110	jjuns.y__22	1

Ⅲ. Data Crawling

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	숖브누	supercalifr	lecinqmailuna	adoles_0110	0
124	__ss.ha			__ss.ha	adoles_0110	1
125	makesblaugh_	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

Ⅲ. Data Crawling

Usage of data

:Data collecting

1. Merge datasets that collected from each team member.
2. Add dummy data(gender, age...) to the merged dataset
3. Dataset training (Classification, Clustering)

Ⅳ. Current Progress

Previous plan

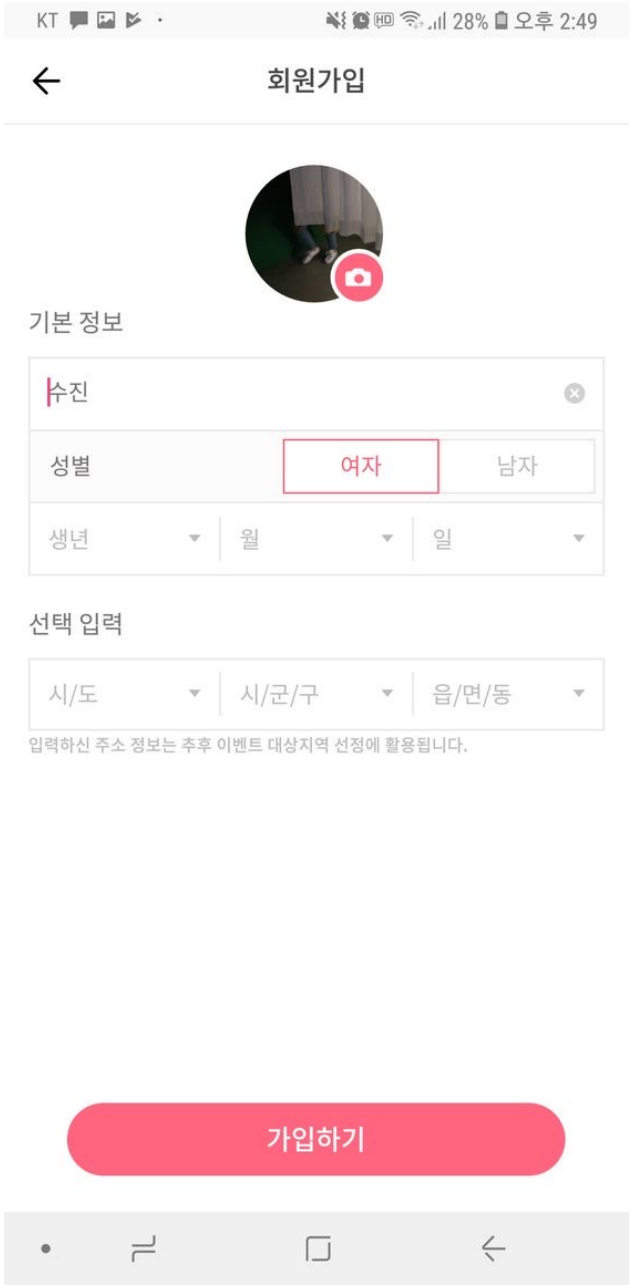
- Data had to be obtained for the development of recommend algorithms.
- Tried to crawl user information using SNS such as Facebook and Instagram.
- Failed to obtain meaningful data to derive features: company thoroughly prevented data crawling.

	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	it__dew	1
21	adoles_0110	鎭	비공계개정	adoles_0110	hoohoohoo__	1
22	adoles_0110	鎭	비공계개정	adoles_0110	jjuns.y__22	1

Meaningless data…

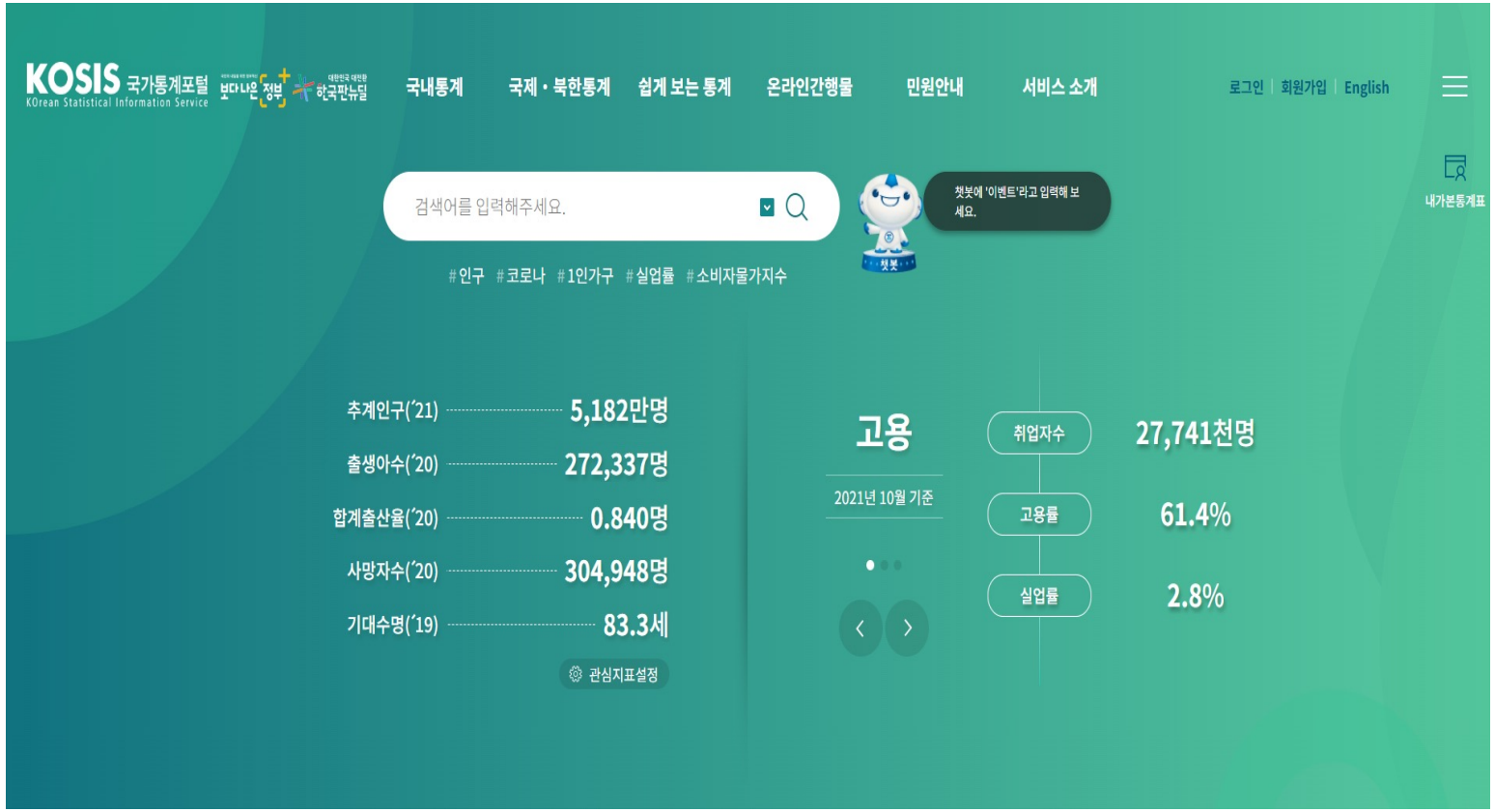
New plan

- Quickly launch applications and use data input from users for model learning.
- It will not be able to provide satisfactory services to early users.
- Later: Gradually improves.



Additional considerations

- Generate random data to build model.
- But literally ‘Random’ data is meaningless.
- Generate as appropriate data as possible using demographics.



Graduation Project

Group 5
2021.11.18

201935006 고효진
201533631 김도균
201935144 최윤찬