인스타그램 유저 프로필 예측

Instagram User Profile Prediction

1조 김도현, 김인용, 박진영, 조하담, 황지상

목차

1.데이터 수집

Data Crawling

데이터 수집

프로필 수집 방법

- 스크래피 패키지 이용
- 커뮤니티 분석을 위해 광고 및 비즈니스 계정 제거
  → 팔로잉, 팔로워가 700명 이하인 계정 수집
- 같은 카테고리 내 해시태그가 여러 개인 포스트 제거

Data Crawling

데이터 수집

프로필 수집 대상

① 인스타그램 내부 아이디
② 인스타그램 아이디
③ 프로필
④ 팔로어 수
⑤ 팔로잉 수

포스트 수집 대상

① 인스타그램 아이디
② 포스트 아이디
③ 포스트 내용
④ 해시태그
⑤ 장소 태그
⑥ 라벨링

Data Crawling

데이터 수집

프로필 검색 키워드

카테고리	카테고리 분류	카테고리 해시태그
연령별 카테고리	미성년자	#10대 #초1 #초딩스타그램 #초딩 #남고 #여고
	20대	#대학생 #대학생공스타그램 #학사모 #개강 #신병휴가 #숙명여대
	30대	#30대 #스물열살 #서른
	40대 이상	#40대 #마흔 #마흔틴 #꽃중년 #50대 #50대몸짱 #50대아줌마
성별 카테고리	남자	#남고 #신병휴가 #신병
	여자	#여고 #숙명여대

Data Crawling

데이터 수집

프로필 검색 결과

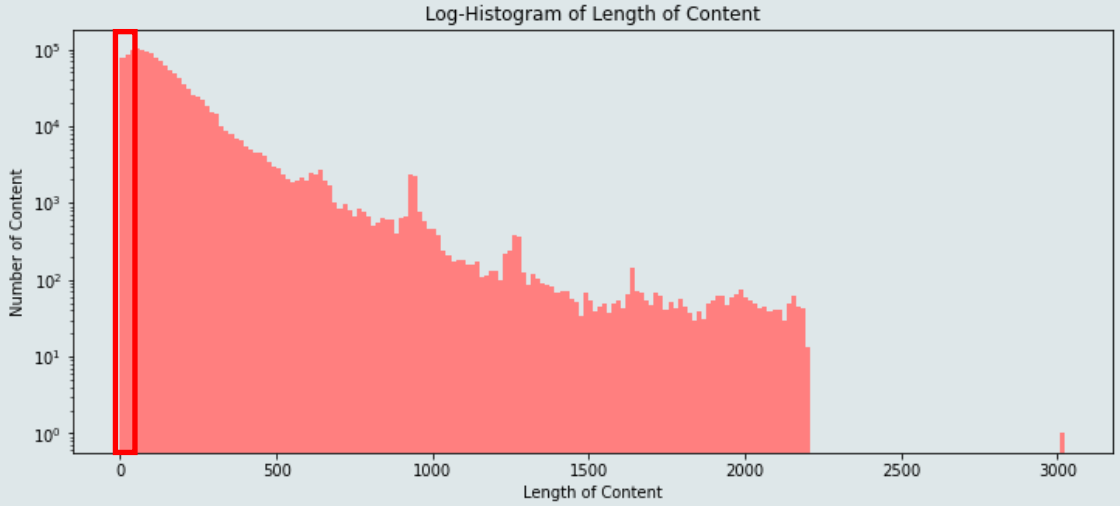
카테고리	카테고리 분류	유저 수	포스트 수
연령별 카테고리	미성년자	5,980명	849,903개
	20대	5,786명	233,156개
	30대	1,190명	125,342개
	40대 이상	8,44명	213,526개
성별 카테고리	남자	2,768명	69,314개
	여자	2,990명	128,592개

* 중복 유저 제거 전

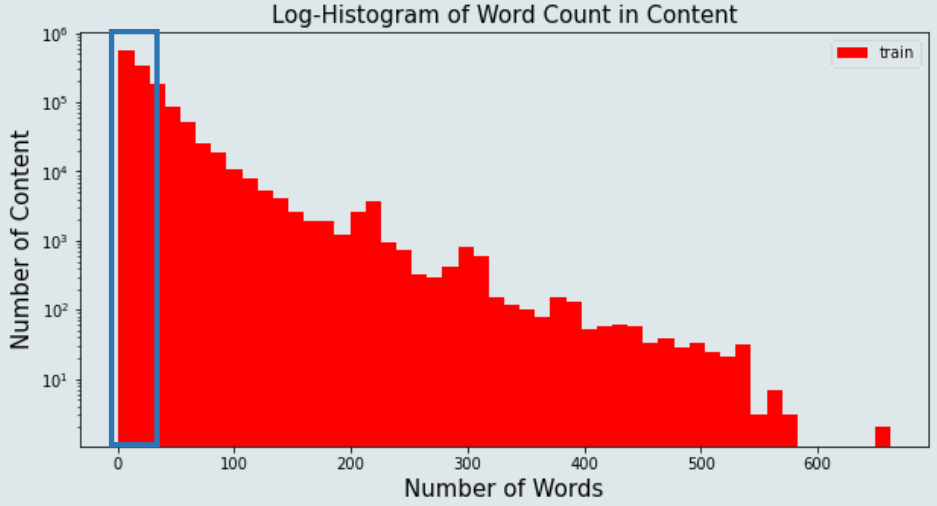
Data Crawling

데이터 시각화

포스트당 글의 길이



포스트당 단어의 개수



- 일정 글자수 미만의 경우 의미 없는 텍스트일 가능성 존재
- 인스타 사용자는 많은 글을 쓰지 않음을 확인

Data Crawling

데이터 시각화

포스트당 단어의 개수 박스 플롯


ngSwitchController) {
attr.ngSwitch || attr.on,
ngSwitchWatchAction(value)
previousElements.length; i < ii; ++i) {
previousElements[i].remove();
previousElements.length = 0;
selectedScopes.length; i < ii; ++i) {
selectedElements[i];
selectedScopes[i].$destroy();
previousElements[i] = selected;
$animate.leave(selected, function() {
previousElements.splice(i, 1);
selectedElements.length = 0;
selectedScopes.length = 0;
if ((selectedTranscludes = ngSwitchController.cases['!'
scope.$eval(attr.change);

2. 데이터 전처리

Data Preprocessing

데이터 전처리

데이터 전처리 전

	insta_id	post_id	content	hashtag	location	label
0	hahajuhye	CEoDHwChqOn	.₩n.₩n.₩n.₩n마이삭₩n조용히지나가길 기원하며₩nRain rain go away			0
1	hahajuhye	CEmLj59Brlx	엄마~₩n오늘₩n재하₩n생일이지?₩n엄마~₩n오늘₩n재야₩n생일이디~?₩n엄마~₩n...	그래, 너오늘생일이다		0
2	hahajuhye	CEmKotWBtxP	띠로리			0
3	hahajuhye	CEmKcZoB8Zn	.₩n.₩n.₩n.₩n조..좋은언니...2			0
4	hahajuhye	CEmJrSlhAqO	.₩n.₩n.₩n.₩n좋은언니♡₩n#흔들리는너의눈동자₩n#걱정마엄마여기있다	흔들리는너의눈동자, 걱정마엄마여기있다		0

데이터 전처리 후

	insta_id	post_id	content	hashtag	location	label
0	hahajuhye	CEoDHwChqOn	마이삭 조용히지나가길 기원하며 Rain rain go away			0
1	hahajuhye	CEmLj59Brlx	엄마~ 오늘 재하 생일이지? 엄마~ 오늘 재야 생일이디~? 엄마~ 엄ㅁㅏ~ #그래...	그래, 너오늘생일이다		0
2	hahajuhye	CEmKotWBtxP	NaN			0
3	hahajuhye	CEmKcZoB8Zn	조..좋은언니...2			0
4	hahajuhye	CEmJrSlhAqO	좋은언니♡ #흔들리는너의눈동자 #걱정마엄마여기있다	흔들리는너의눈동자, 걱정마엄마여기있다		0

→ 이후 content의 음절 수가 5개 이하인 문장 제거

Data Preprocessing

데이터 전처리

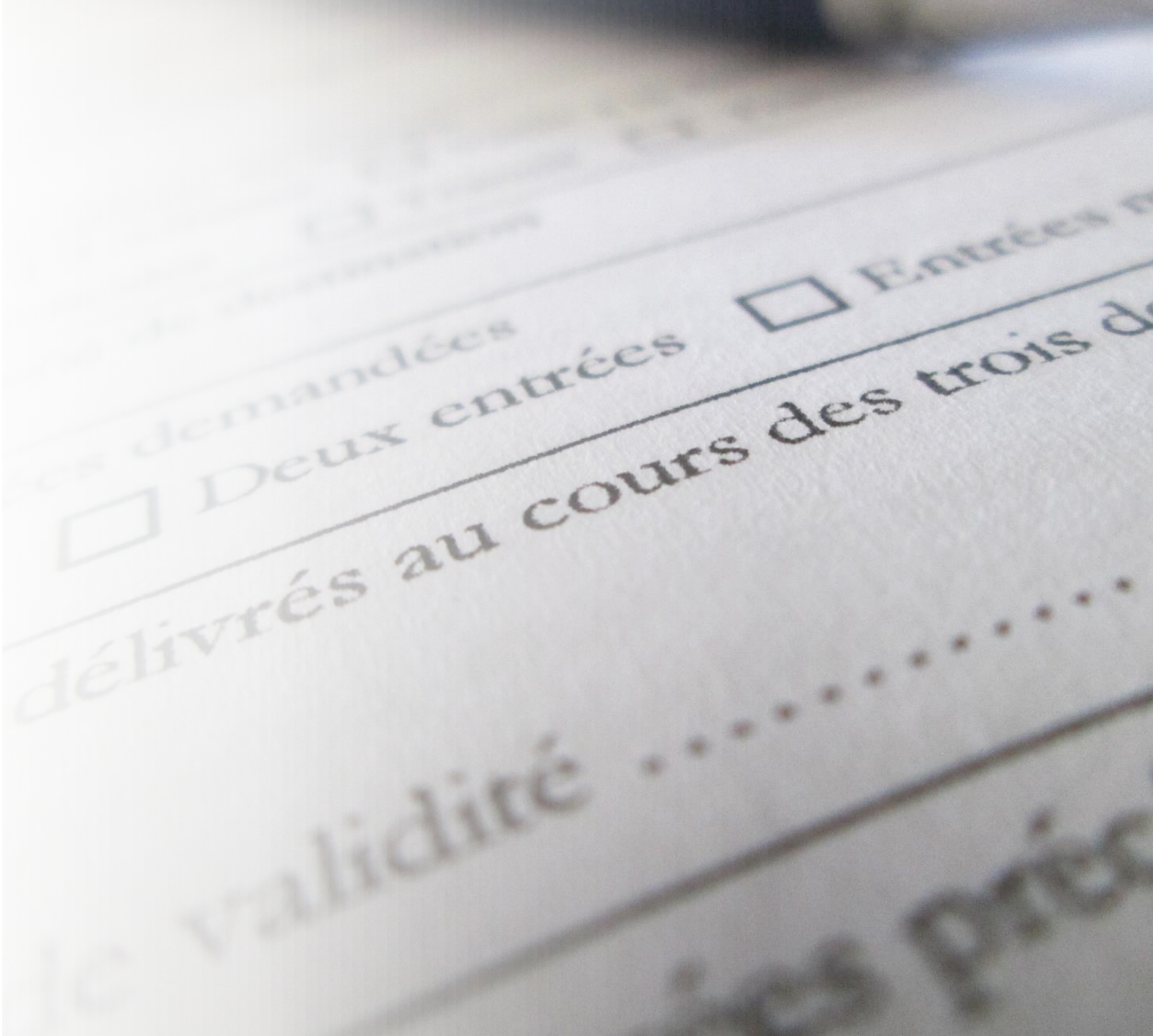
프로필 데이터 전처리 결과

	inner_id	insta_id	profile	follower_cnt	following_cnt
0	1383792898	soojin.0703	NaN	103	34
1	45919	shanmei511	2019.06.29 💑	95	261
2	11800229534	richjmin2019	This is Jamie.₩n먹고 즐기고 사랑하라.₩n방랑먹보의 특별한 매일	90	36
3	5573270284	seojin_hair	단 한명만을 위한 프라이빗 예약제 헤어샵.₩n초록창에서 "사하구서진헤어"를 검색하세...	57	57
4	26763196694	joayree	라떼토끼 스토리 와 자유그림, 나누고 싶은 책을 기록합니다.	117	123

포스트 데이터 전처리 결과

	insta_id	post_id	content	hashtag	location	label
0	rachel_izzy	CFKIupVFrv7	#2020_9_15 #뭇안 #밤라이딩🚲 ₩n아침에 작천정에서 걷기 운동 하고 ₩n...	2020_9_15, 뭇안, 밤라이딩, 절친그램, sean, jake, shule	None	40대
1	rachel_izzy	CFGF3RolZCY	#장하다 #그래비티 #백운산 ₩n#소나무릿지 실패 우벽 등반 ₩n첫 멀티등반 교육 ...	장하다, 그래비티, 백운산, 소나무릿지, 릿지, 피치, 슬랩	밀양 백운산	40대
2	rachel_izzy	CEqPaaglibK	#노안 이 왔습니다 ㅠㅠ ₩n지금도 글자가 잘 안보여요 안경 끼면 잘 보이고 ₩n멀...	노안	서울산두산위브	40대
3	rachel_izzy	CEnYdRElmau	#지난일상 ₩n더웠어도 너무나 재미났던 추억₩n체력만 좋아지면 한번 더 가장 ♡♡	지난일상	Ulsan, South Korea	40대
4	rachel_izzy	CEj9YQIl8hE	많이 묵고 아프지 마라₩n내 남표♡♡ ₩n내 새끼들♡♡ ₩n#쏘고기🐮	쏘고기	서울산두산위브	40대

3. 논문 구현



Thesis Implementation

논문 구현 과정

논문 구현

1. 전체 텍스트 중에서 명사를 추출 (soynlp 패키지 이용)

2. 라벨링 별로 Characteristic Term 추출

3. SVM으로 맞을 확률과 틀릴 확률 계산

4. 한 명의 Known User를 정해, 그 결과가 맞는지 틀린지 확인

Thesis Implementation

논문 구현 과정

텍스트 기반 방식 : 성별 👫

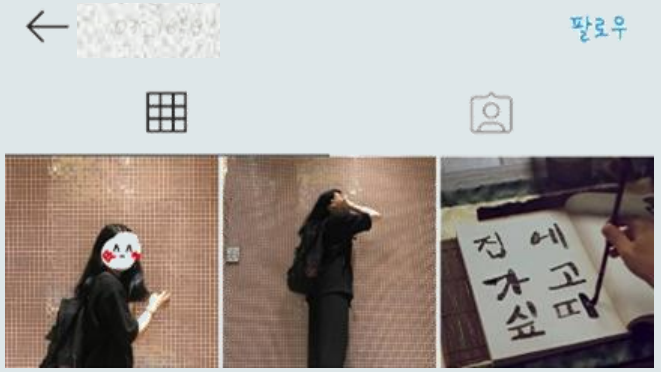
```
4: '마지막',
5: '엄마',
6: '사진',
7: '우리',
8: '운동',
9: '동작',
10: '시간',
11: '한잔',
12: '행복',
13: '함께',
14: '가격',
15: '내일',
```

<Characteristic Term 추출 결과>

- 성별에 따른 특징적인 단어를 추출
- 이후 SVM을 이용해 해당 단어가 성별을 맞출지에 대한 확률 계산

```
array([[0.45216849, 0.54783151]])
```

- 해당 단어를 쓸 유저가 여성일 확률 : 54.78%



<실제 유저 인스타 포스트>

Thesis Implementation

논문 구현 과정

커뮤니티 기반 방식 : 성별 👫

<남자 20대 타겟 유저>

- 수집한 프로필 가운데 유저 한 명을 예측 대상으로 선정
- 1(= 남자)로 라벨링 된 유저를 대상으로 선정

Thesis Implementation

논문 구현 과정

커뮤니티 기반 방식 : 성별 👫

```
from collections import defaultdict
graph=defaultdict(set)
for i,inner_id in enumerate(df["start"]):
    graph[inner_id].add(df["end"][i])
    graph[df["end"][i]].add(inner_id)
graph
```

<군집 형성 코드>

- 해당 유저를 중심으로 팔로워, 팔로잉 관계의 유저 군집 형성
- 하나의 군집은 최대 700명(Threshold)까지만 수집

Thesis Implementation

논문 구현 과정

커뮤니티 기반 방식 : 성별 👫

```
G1=snap.LoadEdgeList(snap.PUNGraph,"sample.txt",0,1," ")
CmtyV = snap.TCnComV()
modularity=snap.CommunityCNM(G1,CmtyV)
```

	insta_id	content	prediction
0	163kg__	내 주변엔 좋은사람들만 있어서 행복해막내동생이랑😜내 코 언제 나아,,,초딩아님‼1...	0
1	97s_ba_e	마니 컸다..세탁실 앞에서 자는 그는..하루종일 잔다#사기 #보이스피싱 조심하세요...	0
2	again._.ty	#다시 #짧은 #글귀 #글 #시 #데일리#사진 #빛바랜 #추억 #칠하기#팔로우좋아요...	0
3	anyeji620	성남시 청년지원센터해피봄스데이🥳🥳.청년이봄 1주년 축하드립니다><코로나19로 인해 ...	0
4	clothing.wholesaler	동대문~땡! • 도매언니/밤시장/ 밴드로 초대합니다.아래를 클릭하시면 밴드 가입 가능하...	1
5	damun_saju	🔎후기🔎마음이 편해지는 사주, 담은사주입니다!이 시국에 나가는 게 꺼려진다면 집...	1
6	dan_ggyu	김떡순 먹으러 가다가 먹고와서 일 할 생각하니 다리에 힘풀려서 길바닥 안부인사 묻고...	0
7	desenhos_da_rapaziada	Arrasta pro lado para ver o progresso do desen...	1
8	dgz_drawings	Izuku Midorya #desenhosanimes #anime #desenhos...	1
9	dongbo1608	대학수시(학교장추천) 가능여부로어제 오늘아침까지 전화기에 불났다불안한 맘은 알지만 ...	0
10	gangmyeongji5936	타로열강중 ㅋㅋㅋ#정신건강타로#심리타로지금은 타로를 배워 자신을 알기 프로젝트 ~•...	0
11	ha_rimi2	26번째 생일🎂#26살 #생일 #행복 #💔#오사카오랜만에 방문한 제주도✈ 숨겨진 예...	0
12	honey._.lim	독감주사맞고 열 39도까지 올라가며 3일동...	0
13	hyeoni_nyang_	#Repost @blanccat.co.kr (@get_repost) · · · 야옹야옹 블...	1
14	its.pusaaaaaa	Another ootd - ...#like4like #likeforlikes #li...	1
15	j_hye0920	사진공모전에 올라간 사진♡청주 성모의료기 사장님께서 직접 사진을 찍어서 올리신 사진...	0
16	je.kt.db	대존맛😳👍#살찌는소리#휘낭시에#마들렌#진짜넘맛있다#🥺#빵지순례#올드베이커리@kaen...	0
17	ju__hee2	👏👏👏👏👏👏👏👏👏👏👏👏👏👏👏👏👏👏👏👏👏👏👏👏🎉🎉🎉🎉🎉🎉🎉🎉🎉🎉🎉🎉🎉🎉🎉🎉🎉🎉🎉🎉🎉🎉...	1
18	jwithsecret	비동사몽에 포항가서크게 기대한건 아니었는데먹어보니 오잉?! 하는 맛😋1인2죽도 충...	0

<군집 내 유저 성별 예측 결과>

· 군집 내 유저들의 성별을 텍스트 기반으로 예측

· 0은 남성,1은 여성을 의미

```
(np.array(file).T[1]/np.array(file).T[1].sum(axis=1).reshape(-1,1))[33]
```

```
array([0.89849971, 0.10150029])
```

<군집의 성별 확률을 구하는 코드 및 결과>

· 해당 군집이 남성 군집일 확률 : 89.85%

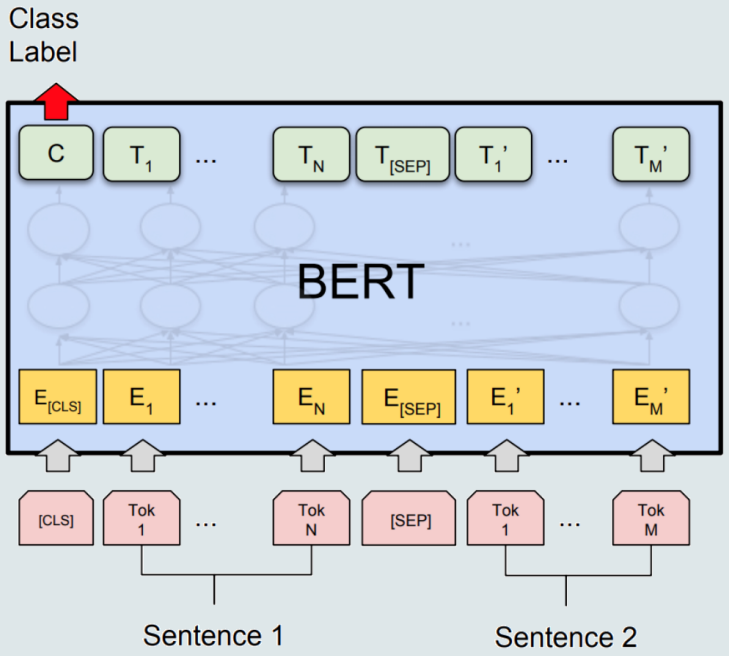
해당 군집이 여성 군집일 확률 : 10.15%

4. 딥 러닝 구현

Deep Learning

딥 러닝 구현 과정

BERT 모델



<BERT 모델의 형태>

- Transfomer를 활용한 언어 표현 모델인 BERT
- 대용량 데이터로 모델을 사전 학습시킨 후 목적에 맞게 Fine-Tuning을 진행
- 구글의 'BERT base Multilingual Cased'의 한국어 텍스트 분석에서 성능의 한계를 보임
- SKT가 제공하는 KoBERT의 사전학습양
  - 한국어 위키 : 59MB
  - 한국어 뉴스 : 290MB

Deep Learning

딥 러닝 구현 과정

텍스트 기반 딥러닝 방식 : 성별 👫

· 각 문장별로 라벨링이 포함된 리스트 형식의 데이터셋 형성

```
[['고맙습니다:) 열무엄마😘 #아이쿠뭘이런걸 #아까워서쓰겠나🙈 #생일선물 #발렌타인콜라보 #bottegaveneta #보테가 #홍연이생
일때어떡하지😔',
  0],
 ['제임스 열무 #007 #고양이 #cat #kitten', 0],
 ['양치기 소치기 심바', 0],
 ['여친이 찍어주는 사진 vs 남친이 찍어주는 사진... 휴....😒 #집들이 #사진 #차이보소 #잘놀다갑니다🙌', 0],
 ['자연광에서 사진찍고싶은데ㅋㅋㅋ요즘 낮에 화장하고 밖에 나갈일이 거의 없응..ㅋㅋㅋ 비타민D 챙겨먹어야지ㅋㅋㅋ', 0]]
```

· KoBERT에서 제공하는 토큰화 적용

```
(array([   2,  993, 6163, 6701,  629,  517,   40, 3358, 6228, 6872, 6141,
           0,  517,    9, 3121, 7550, 6249, 7096, 6040, 5390,  517,    9,
        3093, 5591, 7018, 6553, 6779, 5405, 5655,    0,  517,    9, 2704,
        7126, 6561,  517,    9, 2235, 6056, 7581, 7119, 7542, 6003, 6364,
         517,    9,  517,  380,  427,  442,  442,  389,  399,  367,  453,
         389,  426,  367,  517,    9, 2355, 7618, 5330,    3], dtype=int32),
 array(64, dtype=int32),
 array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       dtype=int32),
 0)
```

<첫 문장 토큰화 결과>

Deep Learning

딥 러닝 구현 과정

텍스트 기반 딥 러닝 방식 : 성별 👫

· 이진 분류 모델 Fine Tuning 진행

```
epoch 1 train acc 0.7170550533491405
epoch 1 test acc 0.770438407921308
```

· 첫 에폭에서 Train 정확도 71.71%, Test 정확도 77.04% 확인

· 성별 분류 모델의 Fine Tuning의 경우 에폭당 20분 소요

나이대별 분류 모델의 Fine Tuning의 경우 에폭당 3시간 소요

Deep Learning

딥 러닝 구현 과정

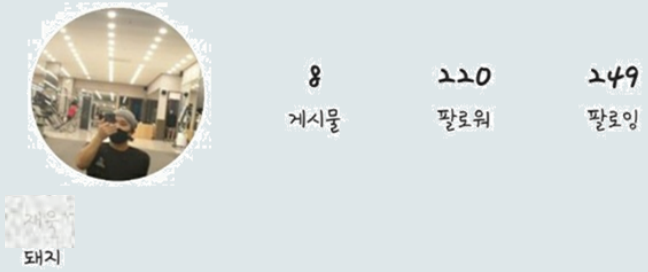
텍스트 기반 딥 러닝 방식 : 성별

· 각 유저가 작성한 전체 텍스트 대상 성별 예측

예측 결과 :

content 렉에 사람도 없고.. 반신욕기.. 아이좋아..#운동 #헬스 #헬린이앙 왕숙천..이제...
predict_label 0

실제 유저 : ← [redacted] : 20대 남성 유저



5. 결론

Project Result

프로젝트 결과 및 시사점

프로젝트 결과

· SVM과 딥러닝 모델링 시간이 오래 걸리는 관계로 현재 성별 분류만이 진행됨

→ 추후, 연령별 분류도 진행할 계획

· 해시태그 기반의 라벨링이기 때문에 정확하지 않은 경우 존재

ex) #10대남자 -> 10대 라벨링 -> 10대 남자의 엄마(실제)

#남자아이 -> 10대 라벨링 -> 남자 아이의 엄마(실제)

프로젝트 시사점

· 커뮤니티 기반 방식에서 군집 내 유저들을 예측할 때, Characteristic Term 기반으로 예측을 진행

→ 이 부분에서 딥 러닝을 이용하는 방안을 고려해볼만 함

Thank You