INPUT : 인스타ID,

OUTPUT :

대중성 3.5 <- 점수로 나타내기

영향력

광고효과

파급력

확장성

인플루언서 추천 <- 힘듦

카테고리 내 다른 인플루언서와 비교 -> 게시물 좋아요, 댓글 비율 / 언어(한국어 vs etc) 비율

키워드 추출(이 사람이 어떤 쪽으로 많이 언급하고 언급되나) 예를 들어 운동도 다이어트, 필라테스, 요가, 헬스, 보충제 등등등

**팔로워 1만~10만(micro) 정한 이유**

- 저비용/고효율을 위해서. 대형(?) 인플루언서는 비싸니까.

- mega에 쓸 돈으로 게시글 검색 시 다수의 게시글 노출을 위해(박리다매)

- mega보다 micro가 소통/참여수준이 높은 경향을 보여줌

- 특정 카테고리화가 되지 않은 사람들이기에 다양한 제품 홍보 가능

- mega인플루언서는 특정 카테고리로 크게 성장했기에 …

- 가짜 팔로워가 가장 많을 것이라 예상되어서.

- 가짜 댓글/팔로워의 형상이 가장 두드러질 것이라 예상되어서.

micro 선정이유? 진짜와 가짜를 확인해볼 수 있는 범위라고 생각되서? 효율이 조금 더 좋을 것 같은

효율성이 더 좋은 인플루언서를 찾기 위해?

- “진짜/가짜” 단어를 대체할 수 있는 표현 ->

**1. 이미지 및 본문 키워드 분석을 통해 카테고리 분류**

**2. 댓글 분석을 통해 제품과 관련된 댓글 비율 확인** -> 말뭉치 어케 만들건지? 애매하니까 최종때 빼도 됨

**3. 댓글 분석을 통해 광고주가 원하는 타겟인지 확인**

**4. 팔로워 리스트를 확인하여 자동생성 계정인지 분류**

**5. 팔로워 리스트를 확인하여 외국인 계정 비율 확인**

**6. 팔로워 대비 좋아요 비율을 확인하여 시각화 자료 제공**

**7. 팔로워 대비 댓글 비율을 확인하여 시각화 자료 제공**

**8. 게시글 수 대비**

**9. 포스팅 시 좋아요 수 예측**

- micro 인플루언서를 제공하고자 하는게 아니다.

- 게시글 수 대비 홍보글 수

- 대형인플루언서에서 데이터를 뽑아서 -> 정답값/기준이 됨.

- 피처링에서는 댓글에 직접 들어가서 연령/성별 데이터를 정리했음.

- “팔로워수가 1~10만 사이로 정해서 분석한 결과로, 모든 case에  
 적용시킬순 없다.” 로 아예 정해버리기.

- ‘야 여기 봐봐’ 등의 어그로성 댓글(작업 댓글)들을 모아서 ‘말뭉치’화  
 -> 데이터 수집 후 정리하기

- 댓글들의 연령/성별 추정하기 -> 이미지분석

- 사물 속성 분류 모델 -> 글에 포함된 사물 카테고리 분석

- 예측 ‘좋아요’ 수 -> 모르는 사람이 보면 혹 할 수 있는.

본문 포스팅 글, 댓글 수 -> 정답데이터: 실제 좋아요/댓글 수 -> 예측데이터 제공

**진짜 인플루언서 / 가짜 인플루언서 차이점은?**

- 진짜 : 팔로워수가 많은 만큼 게시글 수가 많다.(비율은 수집해봐야함)

- 가짜 : 의미 없는 댓글이 많다 (이모티콘 도배, @작업댓글)

- 팔로잉 하는 계정들 분석 -> isFake 700 / 700 -> 학습데이터

{"userFollowerCount": 420, 팔로워수 <- 따르는 사람은 없는데

"userFollowingCount": 4883, 팔로잉수 -> 팔로잉 하는사람이 너무많아

"userBiographyLength": 30, 프로필 밑에 글자 수 (이름 글자 수)

"userMediaCount": 8, 게시글수

"userHasProfilPic": 1, 프로필사진 있는지

"userIsPrivate": 0, 비공개계정인지

"usernameDigitCount": 0, 이름에 숫자개수?

"usernameLength": 12, 아이디길이

"isFake": 1}, 너 가짜

**도대체 댓글만 가지고 뭘 할 수 있는데!**

- 성별/연령 파악 -> 추가작업 필요 -> 결국 인스타는 10~30 or 40대까지만 분석

- 말뭉치화 대상: 이모티콘만 있는 댓글/실제 컨텍스트 없음, 짧고, 일반적이고, 반복적인 문구가 포함된 댓글 (e.g. “예쁘다”, “beautiful” 등), 관련이 없거나 맥락을 벗어난 댓글 (e.g. 음식 사진에 “이쁜 아이”와 같은 댓글 등)

- 좋아요 수 대비 댓글이 현저히 없는 경우

- K팝, K드라마와 같은 주제가 아닌 일상 한글 컨텐츠임에도 불구하고 좋아요를 누른 사람이 외국인(인도, 아랍 등)으로 편중된 경우

- 팔로워 분석(게시글이 없는데 팔로잉만 엄청 많은 경우 - 유령계정) -> 전체 팔로워는 본인만 알수있음

ㄴ> 팔로워를 계속 더 보기가 안된다고 하셨음. 그래도 노가다로 이상한 팔로워를 많이 갖고 있는 인플루언서를 찾아낸다면…

아니면 1차로 필터링 작업할 때 팔로워 수 대비 댓글이나 좋아요 수가 작은 인플루언서 거르고 하는건 어떨까요?(하나의 방법론)

-> 이 사람들이 있어야 가짜인플루언서 데이터 분석 가능하지 않나용?

-> 데이터를 버린다는게 아니라 그.. 결과를 뽑는 과정에서는 저런 방법으로 걸러보는게 좋지 않을까라는 생각인데

**목표 : 광고주에게 제대로 된 정보를 제공해 주겠따!**

- 헤시태그 -> 키워드

- 멀티레이블 -> 회귀.예측

fine-tuning 말고 허깅페이스에서 -> 인물만 추론해주는?

허깅페이스에서 이미지 -> 나이, 성별 분류 추론

팔로워 리스트를 확인하여 외국인 계정 비율 확인 -> 댓글중에서 한국어가 아닌거 찾기

~~좋아요 수 예측 -> 제품별로 얼마나 차이날지도~~ -> 팔로워 많으면 좋아요 수 많을테니까 의미없음

아래 가짜 인플루언서를 만드는 업체들이 주로 하는 작업 리스트를 통해 역으로 방법 고려해볼 수 있을까요?

인스타 조작 블랙마켓

http://www.wizsns.com/guide/price

[인스타그램]

1. 외국인 사진 좋아요

2. 외국인 팔로워

3. 한국인 사진 좋아요

4. 실제 한국인 자동 좋아요

5. 유령 한국인 팔로워

6. 리얼 한국인 팔로워

3. 댓글 분석을 통해 제품과 관련된 댓글 비율 확인

1. 언어 분류를 통해 한국인 댓글만 확인 → 외국인 비율 확인
2. 한국인 댓글 중에서 제품에 흥미를 느끼는 댓글인지 아니면 일반적인 댓글인지 판단 → 제품과 관련되어있는지!!!

→ 이걸 어떻게 할 것인가

1. 긍정, 부정 감성 분석을 통해 긍정인 비율만 확인하는 방법(이모티콘 제거) → 대부분 긍정으로 나눌 것 같음. → 만약 한다면 멀티클래스 문제로 다양한 감성 분석
2. 이모티콘과 인사말만 분류를 하는 방법 → 이건 너무 단순(딥러닝 사용 X)
3. 댓글에서 키워드만 추출하는 방법(카테고리가 다양해서 정하는데 어려움이 있을 수 있음, 또는 너무 많은 키워드가 필요할 수 있음) → 키워드를 추출하는 작업이 필요.
4. 비율

텍스트는 없고 이모티콘만 있는 댓글 → 비율로 시각화하여 보여주면 좋음.

반복적 또는 스팸성 댓글(예: 나이스, 굿, 댓달, 댓글반사 등) → 분류 모델

게시물과 관련 없는 댓글(예: 음식 사진에 “이뻐요”) → 분류 모델

부자연스러운 문법 또는 철자 오류 댓글

1. 본문과 유사도 분석????? → 잘 될지 모르겠음.
2. 관계없는 말뭉치를 노가다를 통해 파악 -> 이모티콘만 붙이는 사람 많음  
   -> 타고 들어가보면 분명 가짜계정으로 보이는 계정도 있으나, 눈팅만 하는 사람들 생각해보면 파악이 쉽지 않음.(가짜라고 단정짓기 애매함)  
   \* 말뭉치 카테고리도 설정해야함

댓글 아이디 패턴 분석 -> muk\_ooo\_ooo, ID유사도분석,  
 -> 직접 그 아이디로 타고 들어가서 isFake판별 후 정답데이터 작성 -> 회사단위로 운영되는, 조작(?)해주는 계정이라면 활동하는 것 처럼 보이기에 판단이 쉽지 않음

fasttext(준단어사용) 사용하여 임베딩을 하고 코사인 유사도를 활용하여 아이디 유사도 확인 → fasttext가 아이디까지 임베딩이 될지는 해봐야 알 것 같습니다.

자동생성된 느낌의 댓글들 분석

**발표 시**

- 인플루언서 2-3명 예시로 들기 -> 진짜/가짜 대조하기

**정답 레이블을 어떻게 구성할 것인가.**

- 총 댓글 중 - - 한 댓글의 비율을 시각화

- 댓글 중 외국인 비율

- 이모티콘만 있는 댓글 비율

- 추천지수

- 활동지수

- ‘~ 가 평균보다 높아요!’ 같은 결과 띄우기

**결과값**

- 점수 : 리스크가 크다

- 시각화 자료만 제공하는걸로 : 결과를 보고 애매한 기분. (피처링 : 시각화 자료+부문마다의 점수로 표기)

빼자! → 나중에 시간이 되면 한번 더 고민) 댓글 분류 말고 > 성의없는 댓글 등의 비율만 확인하자

ㅇㅇㅇㅇㅇ

말자! 말뭉치 구축 등 댓글 분석해서 하자.

ㅇ

fake id 찾는 방법으로는 ML 사용

kaggle 에서 사용했던 피쳐들을 사용하여서 분류 모델 만들기( 케글에 있는 데이터셋을 사용하거나 우리가 직접 데이터셋을 만들어서 사용) → 좋은 생각입니다!!

아니면 직접 auto id로 판별된 계정을 수집하여 피쳐로 넣는 방법도 있음.

번역투 댓글도 구분 필요

가짜는 아니지만 팔로워들끼리 서로 댓글 품앗이 (주로 팔로잉 n천명대의 인플루언서가 함) -> 실제 광고제품 매출에는 영향 없을듯… → 댓글만 보고 판단 할 수 있다면 좋은 말뭉치가 될 듯합니다.

1. 본문 키워드 분석을 통해 카테고리 분류

2. 포스팅 시 좋아요 수 예측

3. 댓글 분석을 통해 제품과 관련된 댓글 비율 확인

4. 댓글 분석을 통해 광고주가 원하는 타겟인지 확인

5. 팔로워 리스트를 확인하여 자동생성 계정인지 분류

6. 팔로워 리스트를 확인하여 외국인 계정 비율 확인

7. 팔로워 대비 좋아요 비율을 확인하여 시각화 자료 제공

8. 팔로워 대비 댓글 비율을 확인하여 시각화 자료 제공

9. 게시글 수 대비 댓글 비율을 확인하여 시각화 자료 제공

10. 좋아요 수 대비 댓글 비율을 확인하여 시각화 자료 제공

댓글수 -> len 함수

게시글,팔로워,팔로우 -> 수집중

나중에 댓글들의 연령 및 성별 확인을 위한 크롤링 필요 -> 댓글 단 사람들의 계정을 분석 → 나중에 몇명 보여줄 때 필요한 부분 이라서 2~3명만 게시글 들어가서 찾아서 모으면 됩니다.

이걸 일찍 할려면 이 2~3명을 빨리 선택해 놓는게 좋을 수도 있습니다. → 사전 학습된 모델에 돌릴 거기 때문에 시간이 많이 필요하진 않습니다.

댓글 있는 게시글만 수집한다. -> 광고이니 좋아요/댓글 등의 반응이 있는게 더 좋아보이고…

못먹어도 고!!

