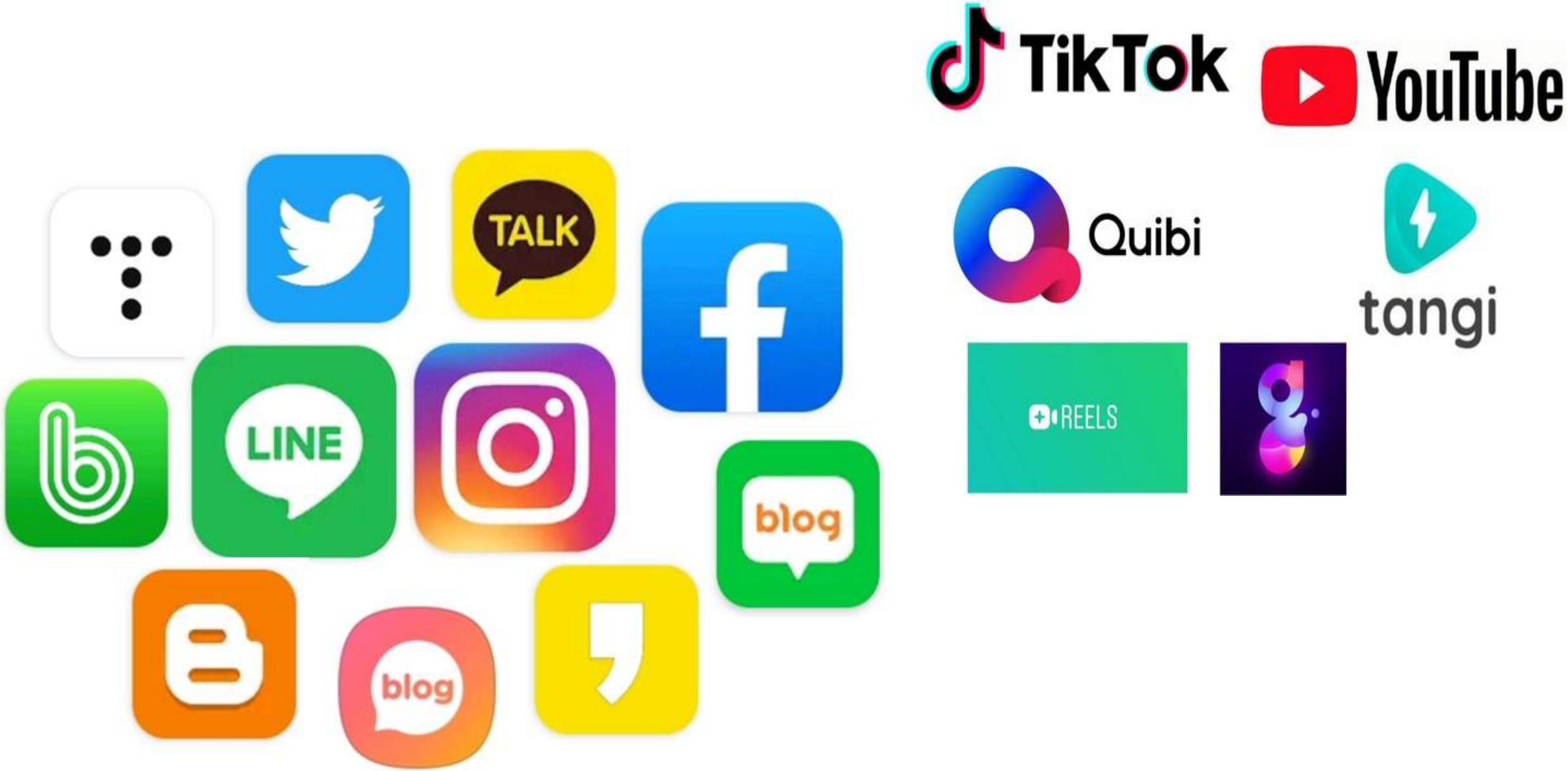
자연어 처리를 활용한 한국어 유튜브 댓글 종류 분류

목차

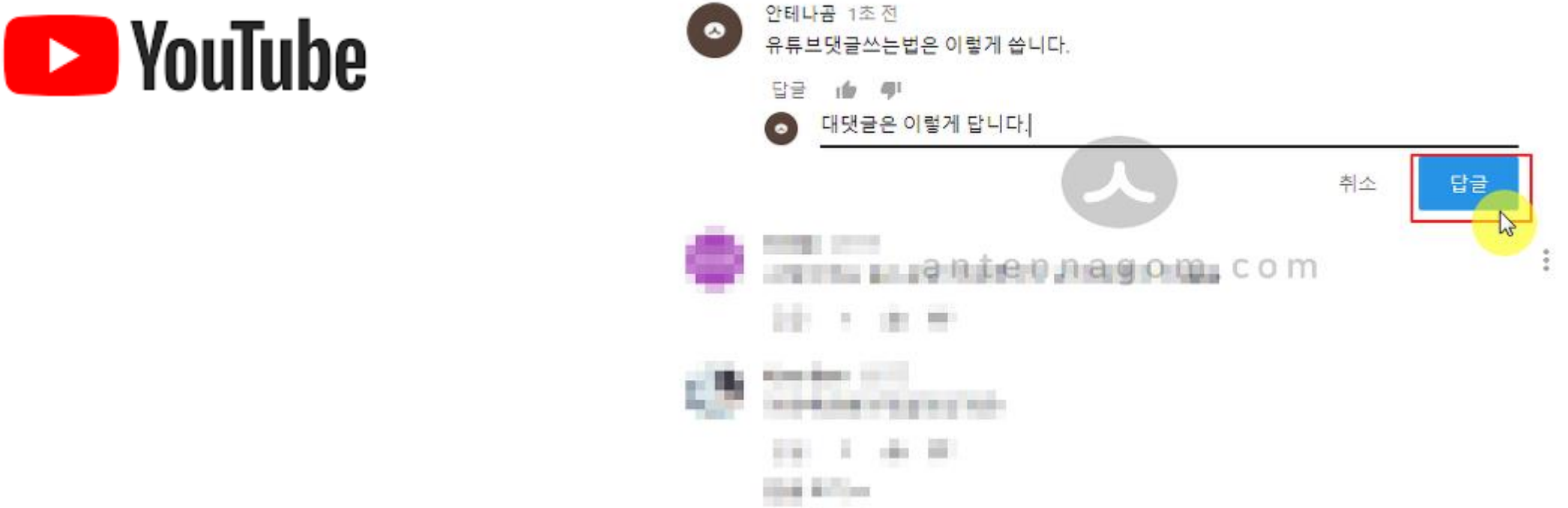
Part 1

프로젝트 소개

1.1 개요 및 목적


TikTok
YouTube
Quibi
tangi
REELS
TALK
LINE
blog
blog

Part 1.1 개요 및 목적



Part 2

진행 상황

2.1 데이터 셋
2.2 실행 결과
2.3 한계 및 문제점

Part 2.1 데이터 셋

	아이디	
0	ㅏ)ㅏ	방장이 얼마나 열심히 일해놨으면 아직까지 빈자리를 채우는 침튜브.. 항상 감사하다
1	박상헌	방장이 없어도 돌아가는 침튜브, 이게 바로 제로 아닐까?
2	쓸데없지만재밌어	침착맨마저 제로가 된 세상.
3	덕키	쉴때 든든한 고봉밥 만들어 놓아서 이렇게 침순실없이 영상보네 침순이는 행복합니다
4	플레르드뽀	침착맨 쉰다니까 침착맨원본박물관,침착맨플러스 정독하고 있음 ... 오히려 좋아 ..밀린 거 많았는데 따라가라고 시
5	부장아재해돌쿤	아니 얼마나 남겨 놨길래 유투부쟁이들은 개방장 방송 쉬는걸 느낄수가 없네 침수자들 넘모 고마워요 요로분♡
6	xnbdjsak	뇌절과 호들갑에 누구보다 심술을 잘 내면서 관련 컨텐츠는 꼭 찍어 내보내는 역설적인 유튜버
7	오현우	나의 저녁을 책임져준 침착맨 제로 감사하다
8	Nateee	이 날 침하하에서 침착맨 사랑해로 방송시작하자해서방송 시작 할 때 다 같이 침착맨 사랑해 채팅 쳤는데컨텐츠하.
9	우주고양이	이렇게 쉬더라도 끝없이 편집본이 나오는걸보니 이제 격달제로 라이브를 할지도
10	Play- maker	침착맨없이도 굴러가는 침튜브를 봤으니 이제 안심하고 원펀데이를 맞이할수있겠다
11	지니	방송을 쉬면서도 계속 올라오는 침튜부 감사합니다
12	콩	펩시제로 라임맛 뒤로 빼는 거 보고 이 사람은 진짜 우리가 어느 포인트에서 열받는지 알고 있다는 생각이 들었음
13	검은냥이 코코& 잡덕집사	ㅋㅋㅋ 맛나게 드시지만 점점 뒤에서 배불러하는 맥콜병건 너무 잼나요
14	쮸스	방장 없이 돌아가는 침튜브... 방장 얼마나 일을 하고 간고야
15	김도지	지속가능한 침튜브를 위해 지금같이 열흘하고 품앗이 다니고 휴식 취하는건 어떨까
16	seunghun	제로가 되어버린 방장의 제로 음료 리뷰이게 진짜 제로 아닐까?
17	ㅅㅇ	환타제로 진짜 맛있지. 먹고 놀라서 한박스 쟁여놨었음
18	널디언니사랑해	"제로"침튜브에 올라온 "춘추제로시대".....이건 귀하다...
19	주마등	침착맨은 이제 하나의 기업이다. 그가 없어도 톱니바퀴는 돌아간다...
20	이윤경	폭력적인 귀여움에 갸날픈 손꾸락까지 완벽하다

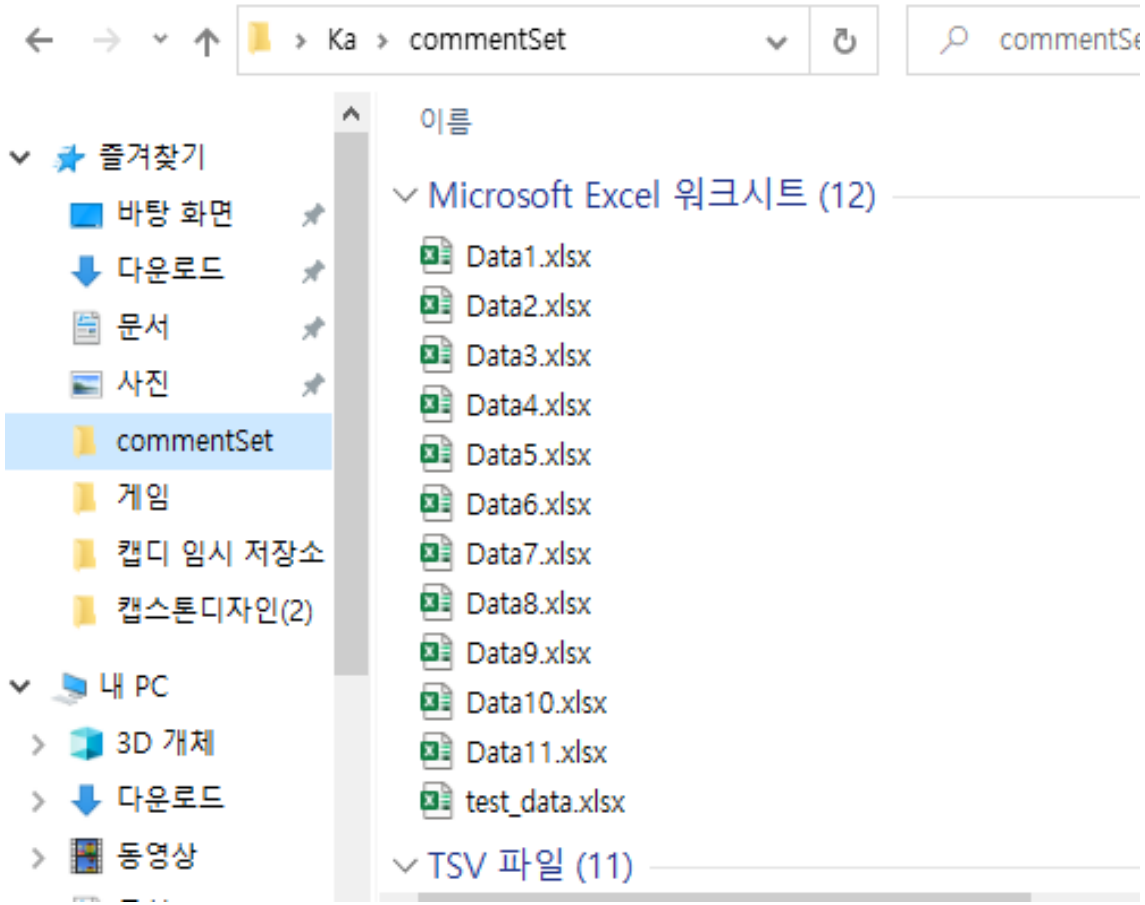
수집 데이터

- 유튜브 영상 하나를 기준으로 약 1100~1200개의 댓글 수집가능

- 작성자 아이디 정보를 제외하고 순수 댓글 내용만 수집

- 같은 주제 영상의 댓글을 수집하여 한번에 분류할 예정

Part 2.1 데이터 셋



수집 데이터

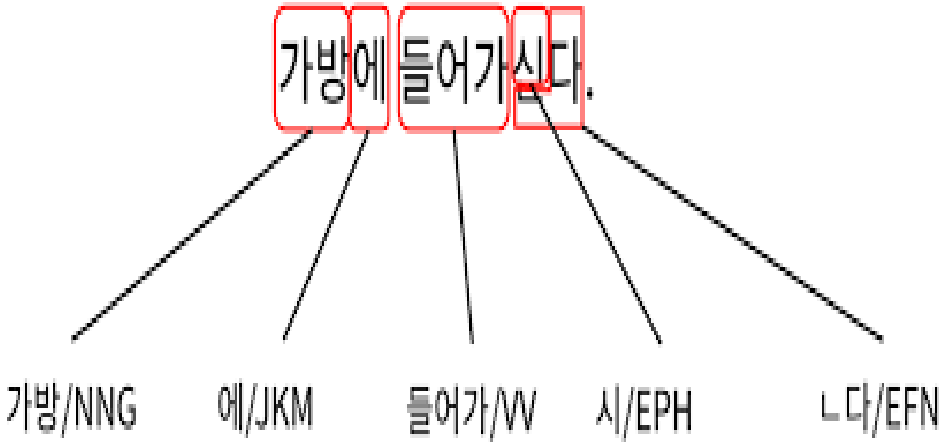
– 유튜브 영상 하나를 기준으로
약 1100~1200개의 댓글 수집가능

– 작성자 아이디 정보를 제외하고
순수 댓글 내용만 수집

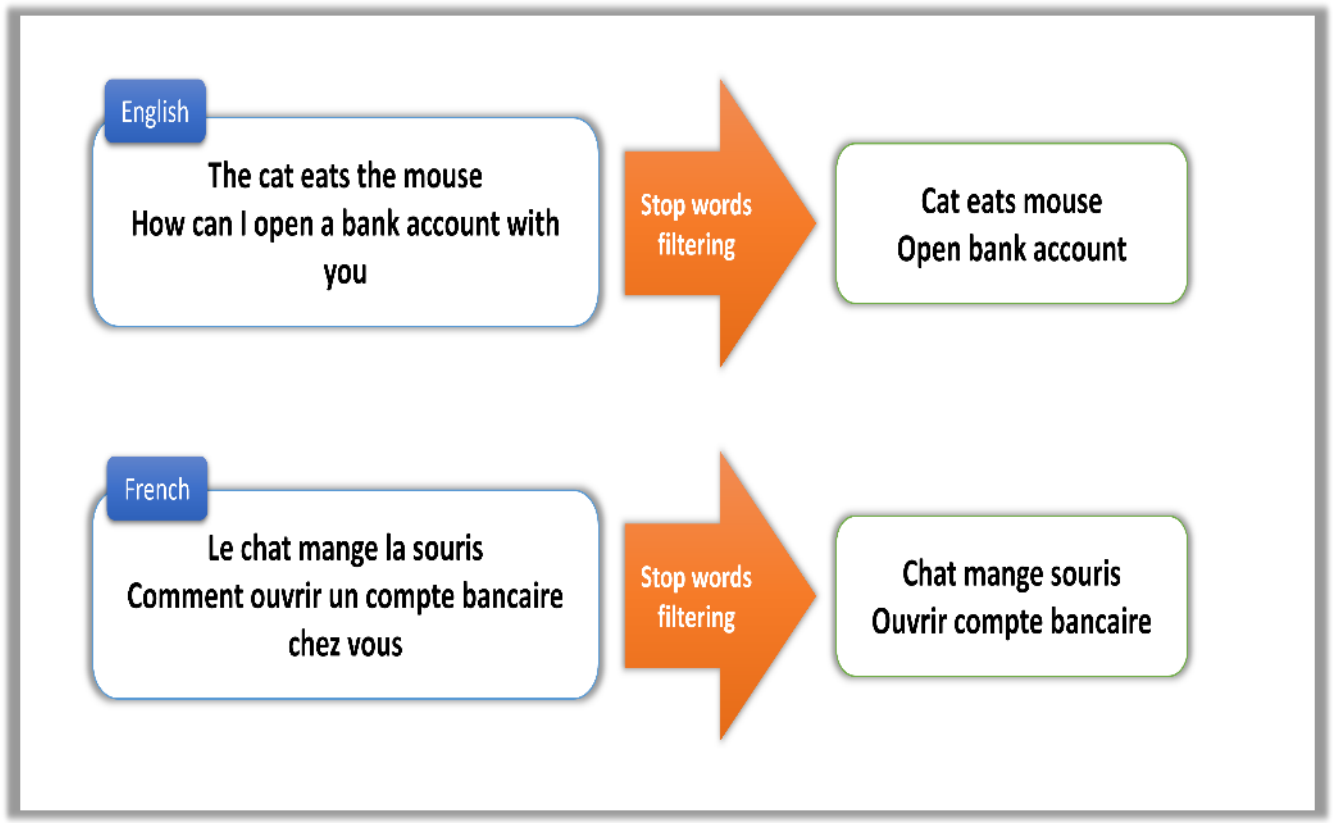
– ~~같은 주제 영상의 댓글을 수집하여
한번에 분류할 예정~~
– 학습을 위해 다양하게 수집중

KoNLPy

한국어 자연어 처리 패키지



형태소 분석
(문장 토큰화)



불용어 (stop_word) 처리

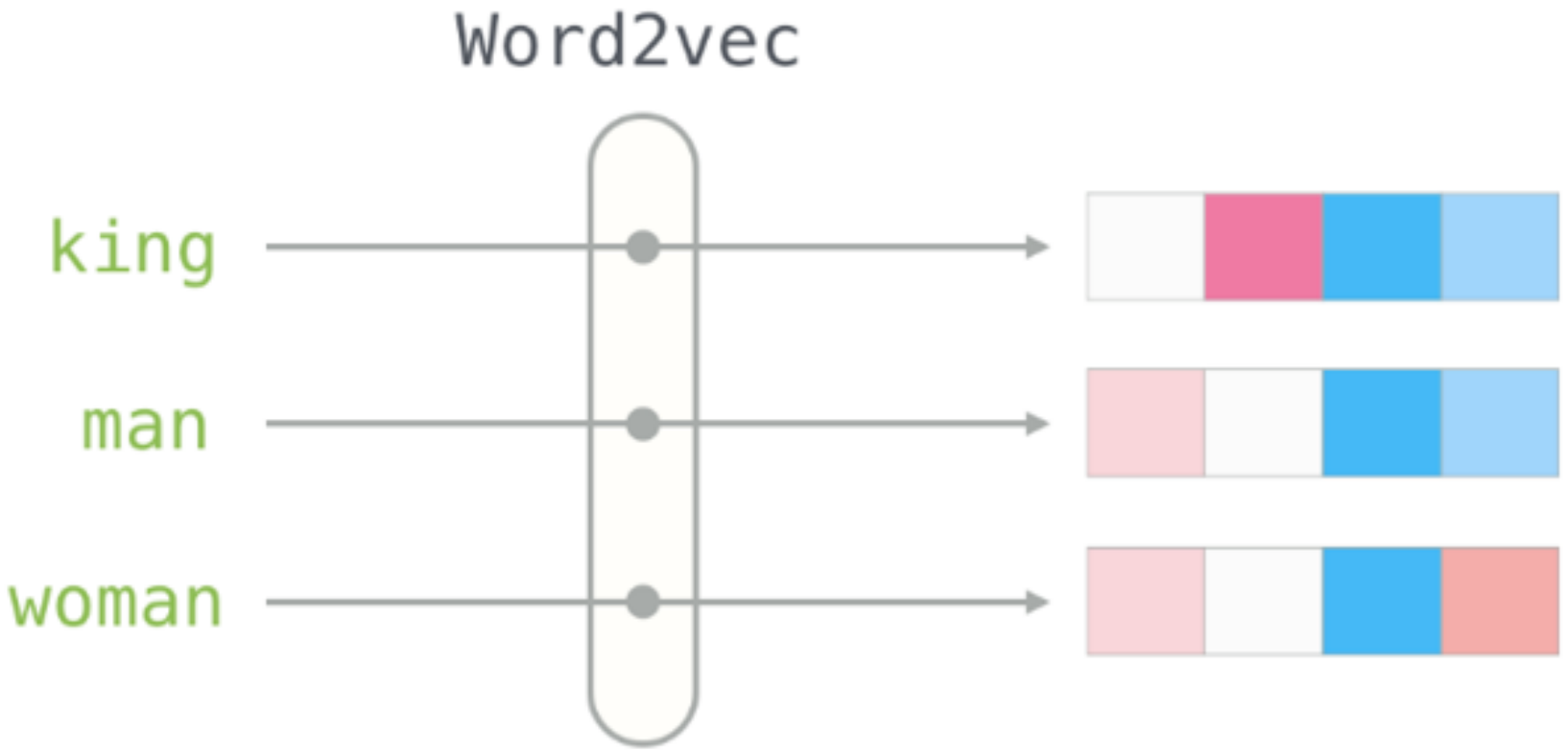
현재 konlpy 형태소 분석을 통해 조사, 접속사를 제거하는 방식으로 적용했다.

인터넷 댓글의 경우 구어체이고, 특유의 은어 등의 문제가 있으므로 추후 성능 개선을 위해 추가적으로 불용어 사전을 만들어 보려 함.

Part 2.1 데이터 셋


대통령
말
진짜
주
일
시간
생각
사람
하루
근무
때
더
노동
저녁
좀
정책
윤석열
정말
대한민국
공무원
인간
과로사
일해
바짝
회사
노동시간
국회의원
문제
노동자
제대로
정부
지금
하나
누가
노동
업무
이상
왜
임
찍
먼저
제발
정치
야근
소리
출퇴근
근로자
고용노동부
한국
도대체
무슨
퇴근
저런
거
뭐
우리
정도
추진
평균매일
우리나라
자기
세상
일요일
생활
시작
돈
애
답
해
수
개
일단
걸
게
안
요
힘
함
밥
쉬
무리
공장
바보
월급
임금
달
보고
근로
얼마나
연차
살
여기
모두
최대
아무
투표
다시
직
언제
모든
게임
과로
직원
이건
정신
댓글
해도
욕
개발
이제
반발
판
세대
휴식
장관
포함
뉴스
무조건
역행
일단
자고
연장
위
한번
휴가
식사
시행
수당
자체
건
탄핵
건가
불만
다음
잔업
가지
노예
알
제
집
때문
웃음
이해
네
월
열
식
중
시대
역대
젠
뭘
모습
번
줄
분
날
도

Part 2.2 실행 결과 : word2vec model



Part 2.2 실행 결과 : pre_trained_word2vec

```
In [7]: kovec.wv.most_similar(positive=['일본', '서울'], negative=['한국'])

Out[7]: [('도쿄', 0.49620240926742554),
         ('영등포', 0.4607112407684326),
         ('서울특별시', 0.45662832260131836),
         ('경성', 0.44781729578971863),
         ('아현동', 0.4475313723087311),
         ('경성부', 0.447209239006042 5),
         ('세종로', 0.44181060791015625),
         ('혜화동', 0.44022461771965027),
         ('원효로', 0.4394114017486572),
         ('상도동', 0.4373798370361328)]
```

```
In [173]: kovec.wv.vocab

Out[173]: {'관위': <gensim.models.deprecated.keyedvectors.Vocab at 0x1bd037754f0>,
           '정어리': <gensim.models.deprecated.keyedvectors.Vocab at 0x1bd037754c0>,
           '유식론': <gensim.models.deprecated.keyedvectors.Vocab at 0x1bd03775580>,
           '장로회': <gensim.models.deprecated.keyedvectors.Vocab at 0x1bd037755e0>,
           '춘추관': <gensim.models.deprecated.keyedvectors.Vocab at 0x1bd03775640>,
           '도입부': <gensim.models.deprecated.keyedvectors.Vocab at 0x1bd037756a0>,
           '민병': <gensim.models.deprecated.keyedvectors.Vocab at 0x1bd03775700>,
```

pre-trained model 사용

Part 2.2 실행 결과 : 문장의 유사도 측정 방식

```
print('문서 1과 문서2의 유사도 :',cos_sim(sentence2vec[0], sentence2vec[1]))
print('문서 1과 문서3의 유사도 :',cos_sim(sentence2vec[0], sentence2vec[2]))
print('문서 1와 문서4의 유사도 :',cos_sim(sentence2vec[0], sentence2vec[3]))

similarity = []

input_sentence_vec = sentence2vec[0]
temp = 0
for sentence in sentence2vec:
    similarity.append(cos_sim(input_sentence_vec, sentence))
```

```
문서 1과 문서2의 유사도 : 0.29738772
문서 1과 문서3의 유사도 : 0.4482292
문서 2와 문서3의 유사도 : 0.5610777
```

1. 광고성 악성댓글의 수가 매우 적거나 없다.
=> 해당 댓글 분류는 따로 수집하여 라벨링하고 테스트에 활용 계획.

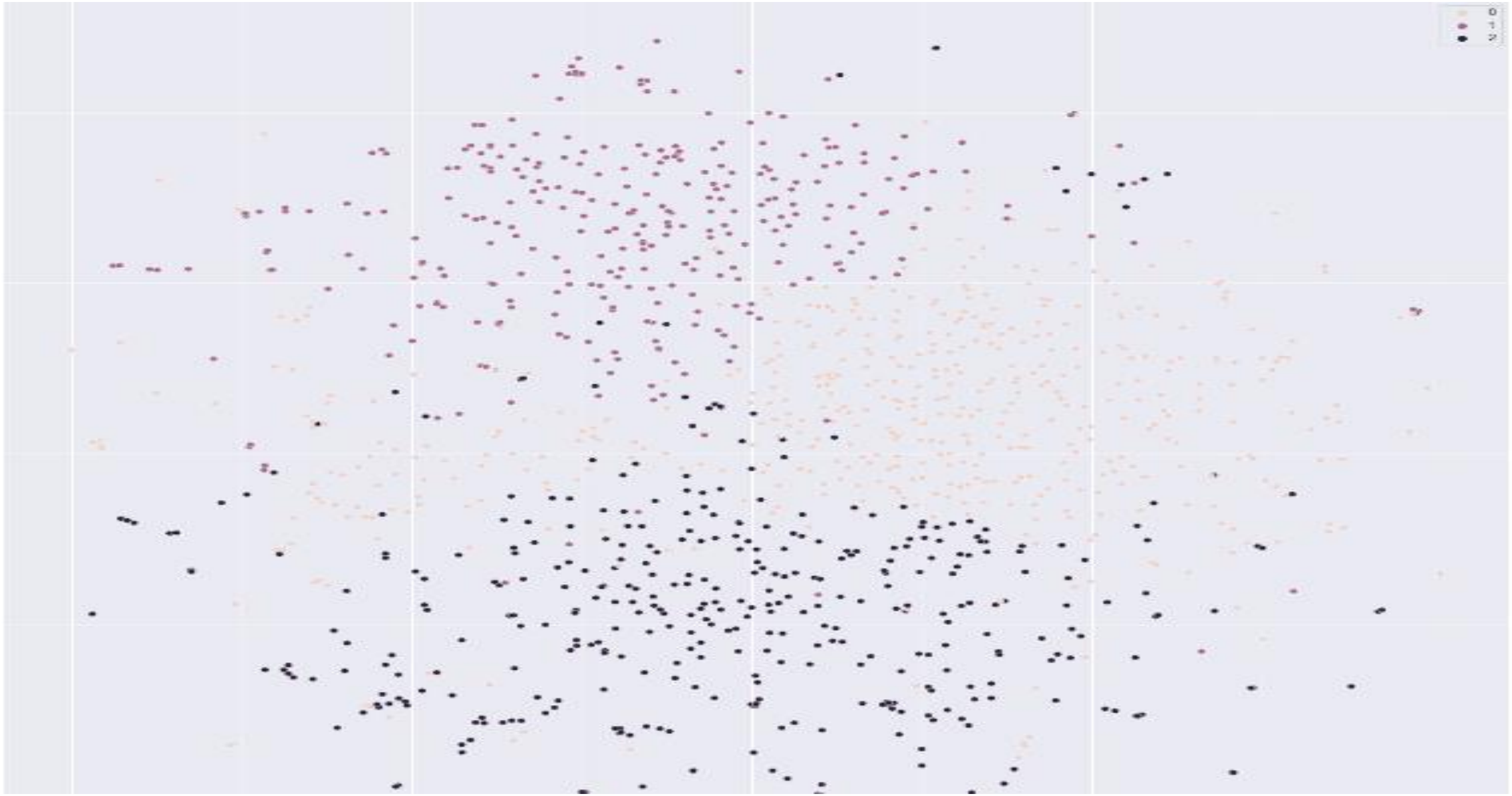
2. 수집한 데이터(근무시간 69시간 연장 관련 뉴스 댓글)에서 주제에 대 한 찬성 반대 데이터가 고르게 분포되어 있지 않다.
(2~3천개의 댓글중 10개 미만)
=>최대한 다양한 의견이 존재하는 영상을 테스트에 사용할 예정
(ex. 토론, 뉴스, 논란이 많은 주제, 안티가 많은 유튜버)

Part 2.2 실행 결과 : 유사문장 출력 결과(1)

```
printResult("어떤 혐의도 적용되지 않는다",W2Vmodel ,"commentSet₩₩Data12.xlsx")
```

```
inputComment : 어떤 혐의도 적용되지 않는다
----------------------------------
유사한 댓글
먼저 한대 맞고 대응으로 두대를 때리면 정당방위가 아니라 과잉대응으로 오히려 먼저 맞은 사람이 피의자가 되버리는 우리나라와는 완전 다르네
여..
이 상황 자체가 너무 무섭네요 ㅎㄷㄷ
사람을 죽였다는 사실 자체가 스트레스로 다가올텐데 아무쪼록 정신건강 잘 챙기셨으면 좋겠네요
----------------------------------
반대 댓글
스폿츠도해보고 ㅋr지노도 해보고 다 찍어먹어보자 ~비제이벳
훌륭한 일 하셨습니다
오우 객사 축하한다
----------------------------------
```

Part 2.2 실행 결과 : 유사문장 출력 결과(1)

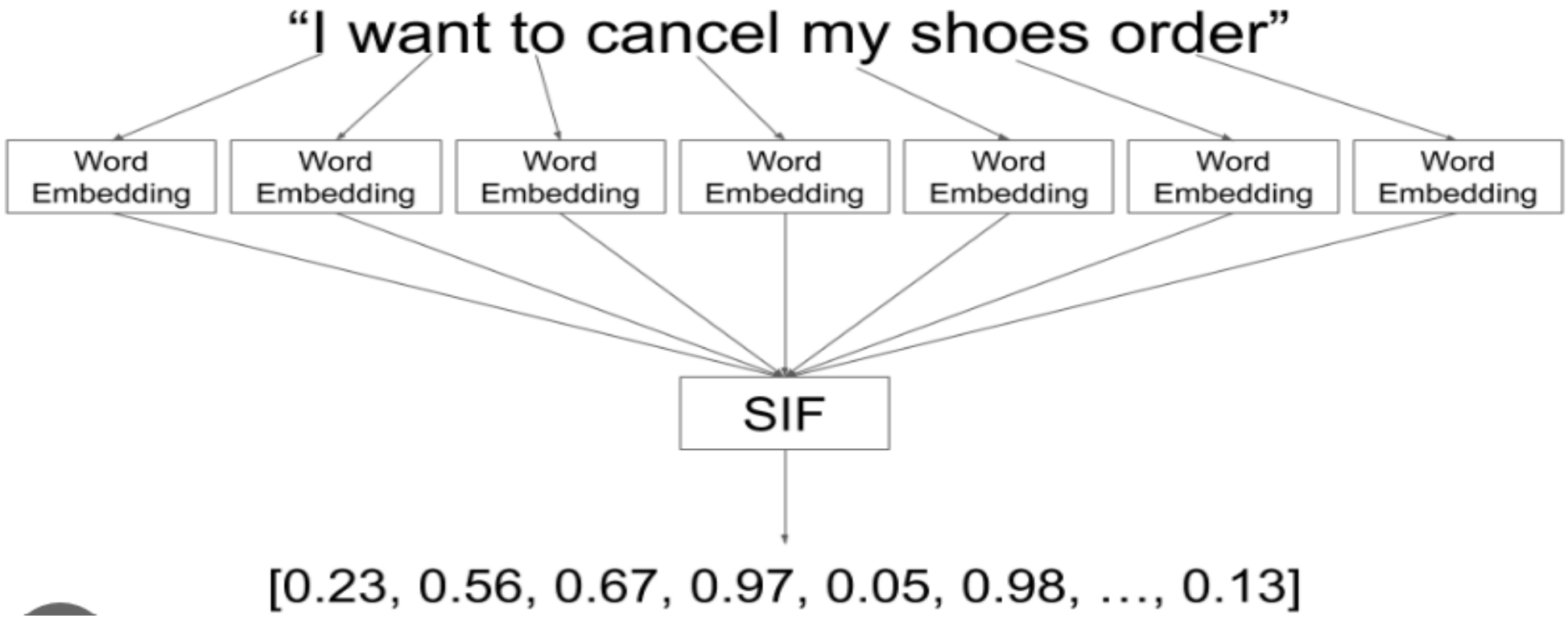


1. pre_trained_word2vec 모델 vocab에 없는 단어가 많아서 분류하려는 문장을 충분히 표현해 주지 못했다.
   =>Fasttext 모델을 사용하여 개선 예정

2. Sentence embedding값을 문장을 구성하는 단어의 vector값의 평균으로 계산했는데 이러한 방식으로 구한 값이 문장을 충분히 표현하지 못하는 것 같다. =>SLF 가중치를 적용하여 개선 예정

3. 분류 결과를 확실하게 구분할 수 있을 정도의 데이터 셋을 구해야 모델이 유용하게 작동하는 지 확인할 수 있을 것 같다.
   =>최대한 다양한 의견이 존재하는 영상을 테스트에 사용할 예정

Part 2.3 한계 및 문제점 : sentence embedding



Part 3

향후 계획

3.1 개선 필요 사항
3.2 향 후 일정

1. Word2vec 모델을 통한 sentence embedding은 효과적이지 않다.
word embedding
모델 교체 => Fasttext, Glove

2. Sentence embedding을 구현하는 더 좋은 방법이 있는지 공부.
=>SLF 가중치를 적용하는 방법

3. 성공적으로 댓글 분류가 이루어 졌을 때 이를 적용할 방법 구상.
ex) ~~유튜브 채널 관리 보조도구, 브라우저 확장프로그램~~
=> 분류 결과 시각화, 스팸 댓글 분류 시도 예정.

Part 3.2 향후 일정

001

개선 필요 사항 정리

모델 학습을 위한 데이터 추가 수집

여름방학~2023.10

002

최소 10만개 댓글 데이터 수집

FastText, Glove Word embeddding모델 학습

SIF(smooth inverse frequency) 문장 임베딩 구현

2023.10~11

003

분류 결과 정리

활용 방안 정리

2023.11~12

004

프로젝트 마무리 보고서 작성 및 최종 발표 준비

2023.12~2학기 종강