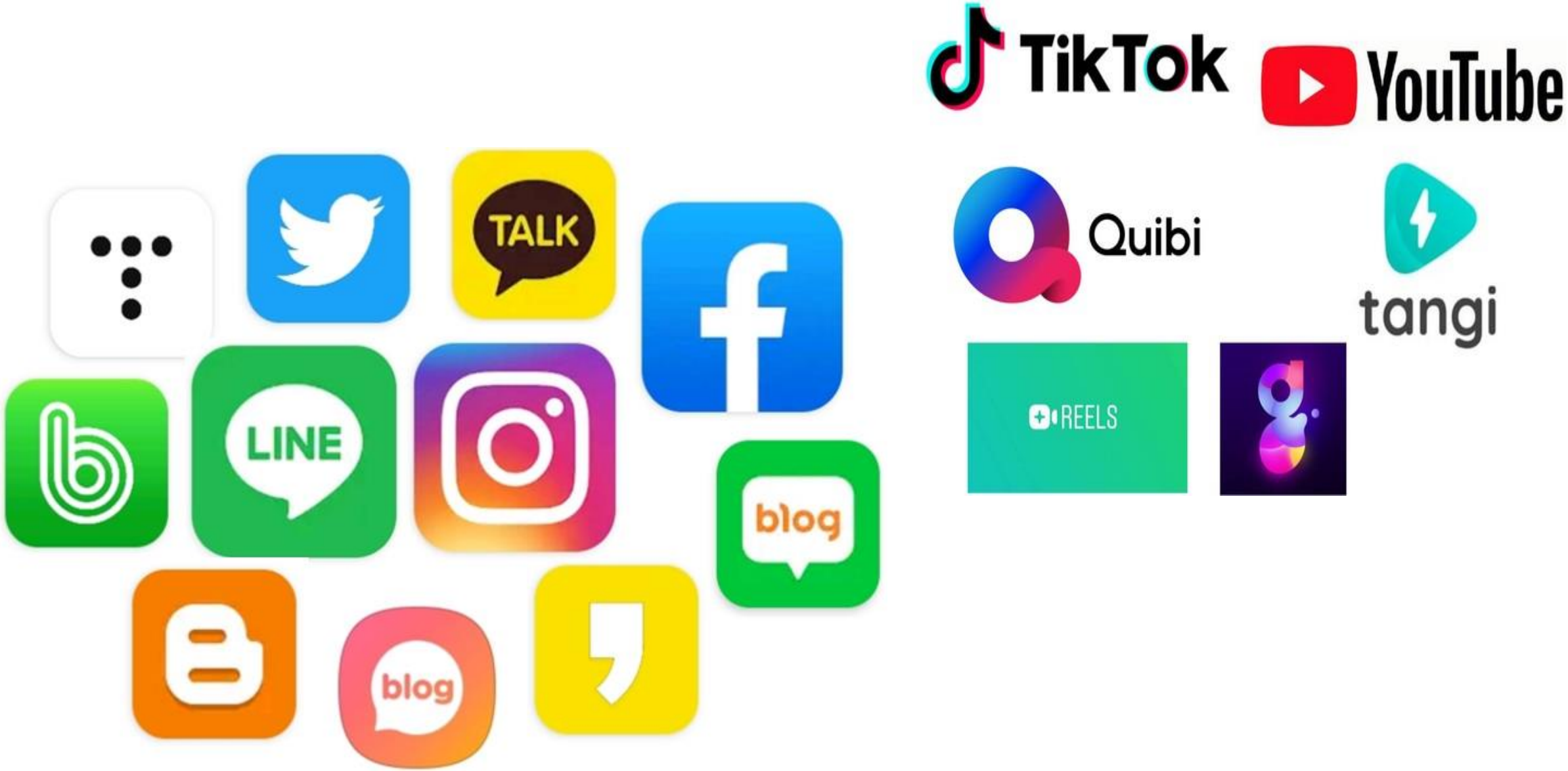
자연어 처리를 활용한 한국어 유튜브 댓글 종류 분류

목차

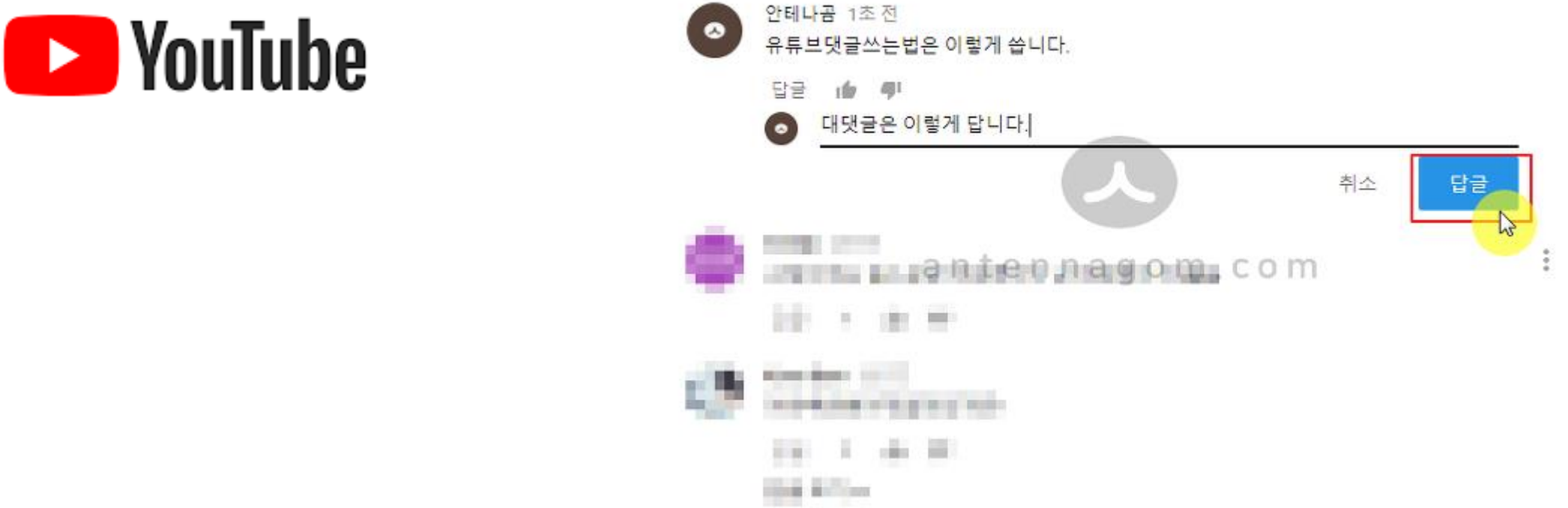
Part 1

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

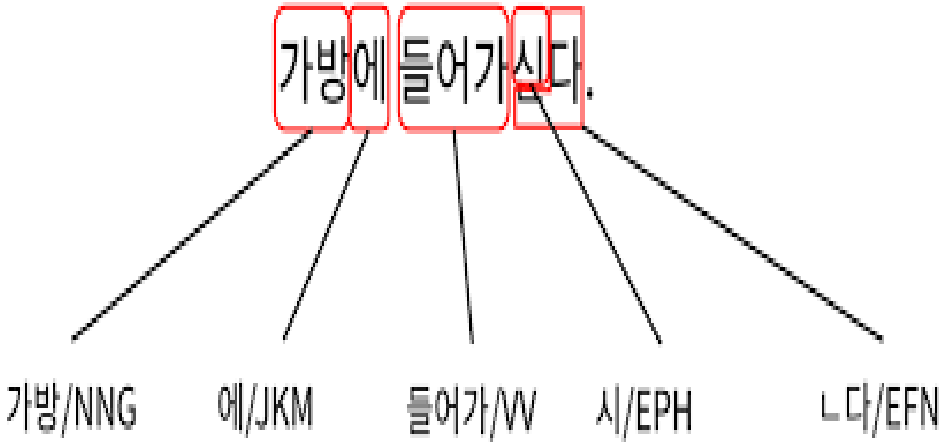
- 유튜브 영상 하나를 기준으로 약 1100~1200개의 댓글 수집가능

- 작성자 아이디 정보를 제외하고 순수 댓글 내용만 수집

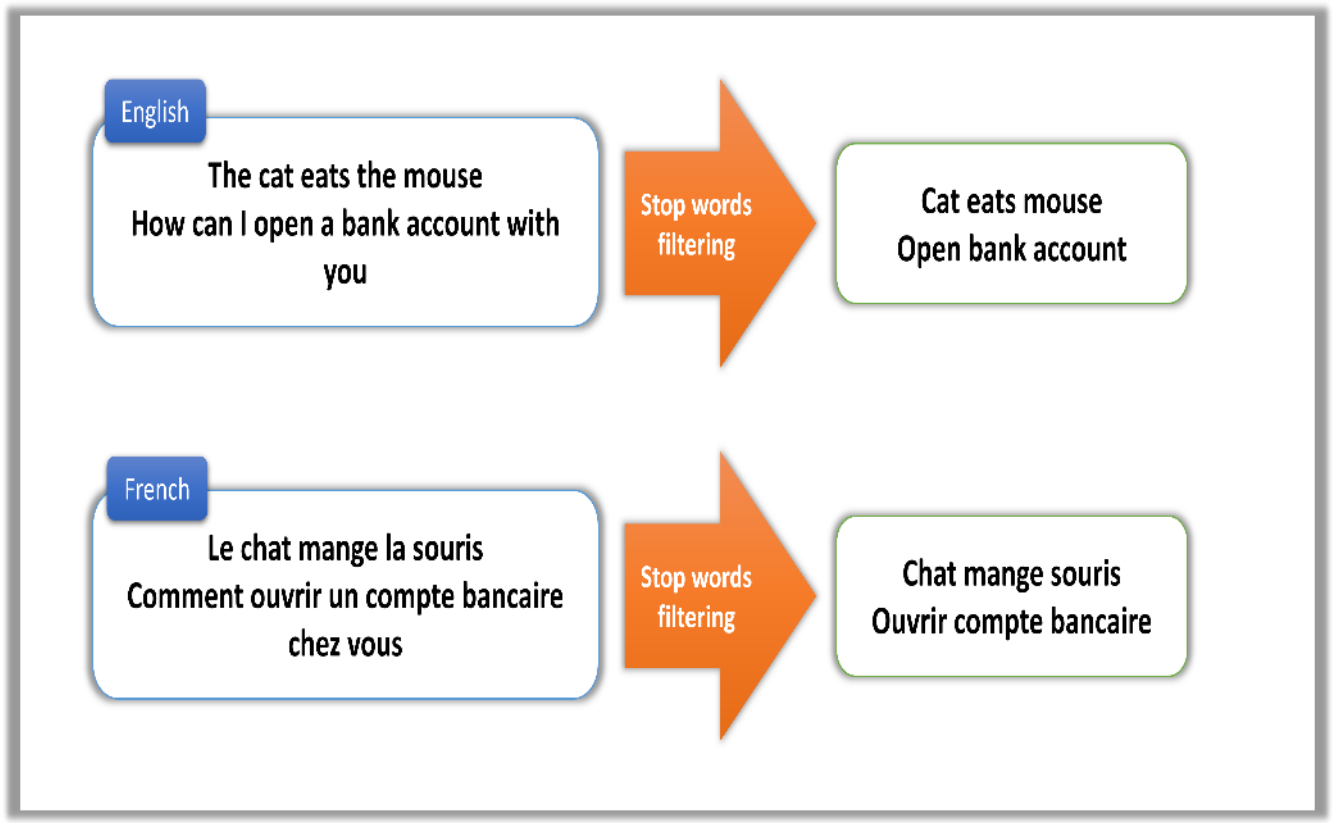
- 같은 주제 영상의 댓글을 수집하여 한번에 분류할 예정

KoNLPy

한국어 자연어 처리 패키지



형태소 분석
(문장 토큰화)



불용어 (stop_word) 처리

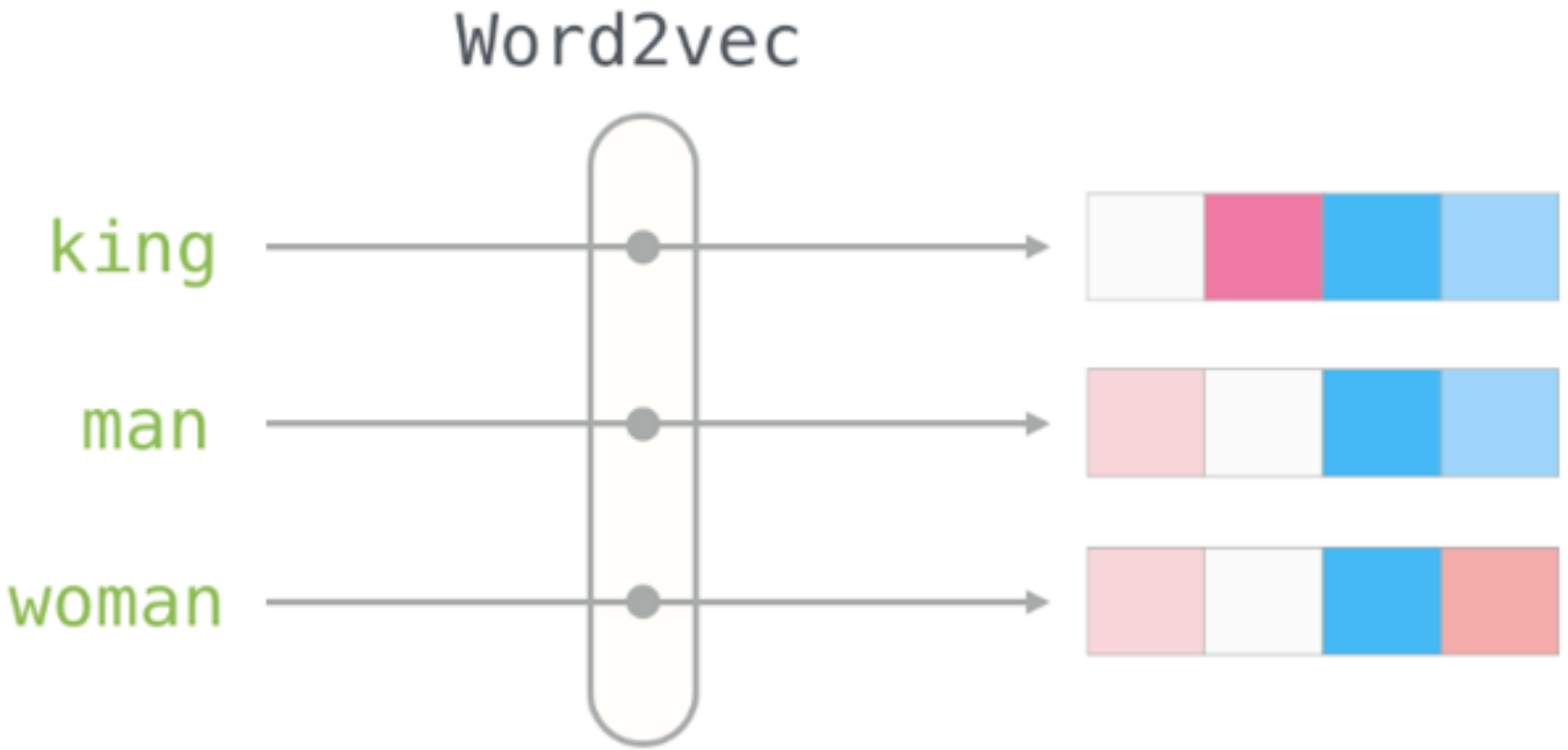
현재 konlpy 형태소 분석을 통해 조사, 접속사를 제거하는 방식으로 적용했다.

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
일
주
시간
생각
하루
사람
정책
윤석열
대한민국
공무원
노동
근무
노동시간
회사
문제
노동자
바짝
제발
지금
더
때
일해
정말
좀
과로사
인간

Part 2.2 실행 결과 : word2vec model



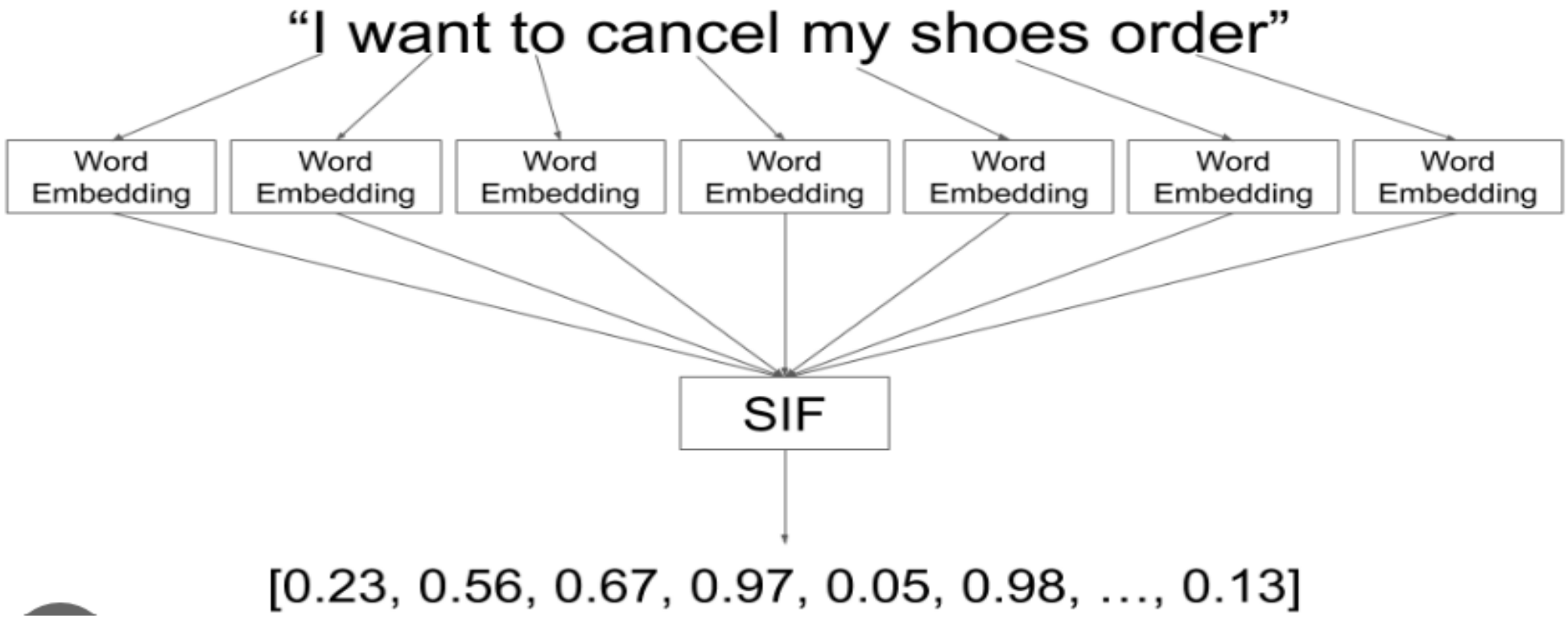
Part 2.2 실행 결과 : pre_trained_word2vec

```
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 : sentence embedding



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.

2. 수집한 데이터(근무시간 69시간 연장 관련 뉴스 댓글)에서 주제에 대한 찬성 반대 데이터가 고르게 분포되어 있지 않다.
(2~3천개의 댓글중 10개 미만)

Part 2.2 실행 결과 : 유사문장 출력 결과(1)

input sentence
대통령님이 솔선수범하셔서 주 170시간 업무를 연중무휴 휴가없이 임기동안 하시고 쉬어주는 모습을 보인다면 국민들도 큰 불만 없을것 같습니다.
similarity top_5 sentences
대통령도 주60시간일해보자. 먼져 솔선수범을해야 국민이따르지 ㅋ
국민을 사랑하시고 국민보다 한 발짝 앞서가는 윤대통령님부터 본보기로 120시간 근무 하시는거죵?
120시간 일하고 2주 놀아? 저런게 대통령이라고...윤석열 뽑은 인간들은 평생 반성하며 살아라.
과거에 삶의 체험 현장 tv가 있는데삶의 체험현장을 다시 부활해서..첫번째 출연자를 대통령으로 모셔서.주 120시간. 근로자들의 삶이 얼마나 힘든지 몸소 체험해봤으면.....

Part 2.2 실행 결과 : 유사문장 출력 결과(2)

input sentence : 서민들이 부자와 기득권을 위해 투표잘했습니다.
top_5 similarity sentences
국민 여러분들 투표 제대로 합시다 좀
이래서 투표를 잘해야한다
절대로 친일당 다신 뽑지않으리 이렇게 투표가 중요합니다
불만가지는놈들 누구 투표햇는지 알고싶네 ㅎ
투표의 중요성!!!

Part 2.2 실행 결과 : 유사문장 출력 결과(3)

input sentence : 교대근무의 활성화로 과로사망을 예방해야 한다 !
top_5 similarity sentences
노동자들 평균수명 줄여서 국민연금 건강보험 개혁을 도모 하는건가...
현재 꾸준히 문제가 되고있는 과로사 문제… 업무상재해로서 주52시간 이상 근무하면 과로사로 인정하는데, 주 60시간을 근무하면 과로사 인정기준이 높아지는 걸까요, 더 많은 사람들이 과로사로 떠나게 될까요? 어느 쪽으로든 염려스럽습니다.
120시간 근무에 따른 최저시급 이상의 계약근로급여 + 야간근무 급여 + 식사비 + 휴식비 등을 정부에서 모두 지급하되, 그 시간 근무에 따른 산재피해 및 과로사 등등 산재피해자와 사망자들의 가족/유족들에게 모든 피해보상금을 다 합하여서 100% 정부에서 지급하십시오.
알콜성 치매 아닌지 시급히 검사 합시다.
짠한 공무원들... 수습하느라고과로사 걱정되네요

Part 2.2 실행 결과 : 유사문장 출력 결과(4)

input sentence : 윤석열이 말한 게임산업 120시간 근무 맞는 말이긴함. 세기의 게임 명작 "위쳐3" 개발사가 실제 이 근로시간 제한 때문에 차기작 폭망했음. 게임개발 하려면 제일 중요한게 팀원들끼리 스프린트 해서 밤새서 만들어 내는게 중요한데, 근로 시간 법으로 지정해 놔서 팀원들끼리 그냥 퇴근하고 각자 플레이하다 차기작 폭망해서 게임사 거의 망했음. 진짜임. 그리고 이 근로시간 제한때문에 게임 개발 비용이 3배 이상 올라서, 고비용 저품질 현상이 고착화 됨. 나도 개인적으로 주 50시간 이렇게 법으로 근로시간 정해놓으면, 직원들 스프린트해서 일해야 할 시기에는 어떻게 대처할건지 의문. 회사는 당장 스프린트해서 결과물 내놔야하는데, 법때문에 직원들 집에 보내야하면 회사 망하는거지.

80-90년대만해도 야근에 회식123차에 토일 주말 근무까지 다하면 주80시간은 기본이었는데....뭐 그래 열심히들 일한다고 희생양인것처럼 아득바득 몇시간에 열을 올리나... 업무 효율 역량은 떨어지면서 보상만 오지게 따지지...회사 경영진들이 불쌍하다....

input sentence : 80-90년대만해도 야근에 회식123차에 토일 주말 근무까지 다하면 주80시간은 기본이었는데....뭐 그래 열심히들 일한다고 희생양인것처럼 아득바득 몇시간에 열을 올리나...업무 효율 역량은 떨어지면서 보상만 오지게 따지지...회사 경영진들이 불쌍하다....

top_5 similarity sentences

1 연장근무하게해줘요~~교대근무하는사람은연장을해야돈이되는데...업종별로달리해서근무기준표만들면좋겠어요~~~~~특히제조업은주52시간너무애매해요..주60시간이상해야월급다운돈을받죠.

1. pre_trained_word2vec 모델 vocab에 없는 단어가 많아서 분류하려는 문장을 충분히 표현해 주지 못했다.

2. Sentence embedding값을 문장을 구성하는 단어의 vector값의 평균으로 계산했는데 이러한 방식으로 구한 값이 문장을 충분히 표현하지 못하는 것 같다.

3. 분류 결과를 확실하게 구분할 수 있을 정도의 데이터 셋을 구해야 모델이 유용하게 작동하는 지 확인할 수 있을 것 같다.

Part 3

향후 계획

3.1 개선 필요 사항
3.2 향 후 일정

1. 분류하려는 데이터 word를 모두 표현해 주기 위해 직접 word2vec 모델을 디자인 해야 한다.

2. Sentence embedding을 구현하는 더 좋은 방법이 있는지 공부.

3. 유의미한 분류가 가능한 다양한 의견의 데이터 셋이 필요하다.

4. 성공적으로 댓글 분류가 이루어 졌을 때 이를 적용할 방법 구상.
ex) 유튜브 채널 관리 보조도구, 브라우저 확장프로그램

Part 3.2 향후 일정

001

개선 필요 사항 정리
Word, Sentence Embedding을 위한 사전 지식 학습

2023.05~07

002

Embedding 모델
분류 모델 구현
데이터 분류

2023.07~09

003

분류 결과 활용
모델 구현
(채널관리 보조도구)

2023.09~10

004

프로젝트 마무리
최종 발표 준비

2023.10~2학기 종강