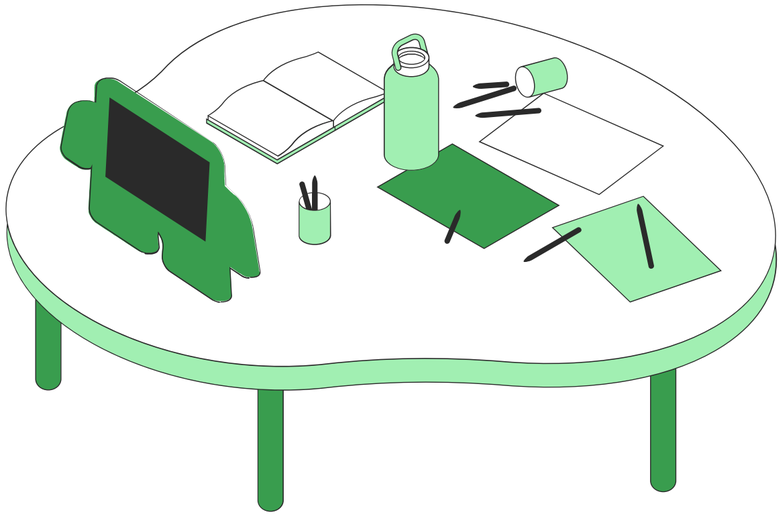
제 3 회 KUBIG Conference

고파스 게시물 분류/예측

분류/예측 2팀
15기 이병주 16기 최규빈 이은찬 이영노

Index

고파스 게시물 분류/예측 프로젝트



00. 과제 소개

01. Concept

02. 데이터

03. EDA

04-1. 분류 : BiLSTM을 이용한
인기게시물 분류

04-2. 예측 : BiLSTM을 이용한 조회수 예측

04-3. 분류 : 감성분석

04-4. 분류 : 클러스터링

00. 과제 소개

자유게시판 1

다양한 주제를 공유하는 호랭이들의 광장입니다. (존댓말)

처음으로 | 내가 쓴 글 | 스크랩 | 알림 1 | 쪽지 19 | 배심원

댓글	제목	읽음	날짜
6	남자 매직 한 직모 바가지머리		12:54
5	여자 방광염 진료 검사 비용		12:52
5	의사님들 질문있습니다. 눈 관련		12:47
20	역시 탈고대가 답인가요		12:44
12	투표 할아버지 핸드폰 s22 vs a33		12:44
4	눈물샘 눈물길 막힘? 눈물이 계속 나요.....		12:21
44	지금 계량기 보고 있는데, 70만원 실화일까요..?		12:07
4	상담 어플 추천 부탁드립니다		11:54
11	LG 유플러스 개인정보 유출 확인해보세요		11:42
17	아옳이 하트시그널 서주원 상간녀 소송제기		11:41
7	직장 병행 가능한 경영대학원 있나요?		11:19
11	장학금 이중수혜 관련 질문		11:18
23	2주년을 준비 못했습니다..		11:10
13	(신규가입) 교재 PDF 사기당했어요...		11:08
4	솔로지옥 후기		11:04

00. 과제 소개

자유게시판 1

다양한 주제를 공유하는 호랭이들의 광장입니다. (존댓말)

처음으로 | 내가 쓴 글 | 스크랩 | 알림 1 | 쪽지 19 | 배심원

댓글	제목	읽음	날짜
6	남자 매직 한 직모 바가지머리		12:54
5	여자 방광염 진료 검사 비용		12:52
5	의사님들 질문있습니다. 눈 관련		12:47
20	역시 탈고대가 답인가요		12:44
12	투표 할아버지 핸드폰 s22 vs a33		12:44
4	눈물샘 눈물길 막힘? 눈물이 계속 나요.....		12:21
44	지금 계량기 보고 있는데, 70만원 실화일까요..?		12:07
4	상담 어플 추천 부탁드립니다		11:54
11	LG 유플러스 개인정보 유출 확인해보세요		11:42
17	아옳이 하트시그널 서주원 상간녀 소송제기		11:41
7	직장 병행 가능한 경영대학원 있나요?		11:19
11	장학금 이중수혜 관련 질문		11:18
23	2주년을 준비 못했습니다..		11:10
13	(신규가입) 교재 PDF 사기당했어요...		11:08
4	솔로지옥 후기		11:04

00. 과제 소개

자유게시판 1

다양한 주제를 공유하는 호랭이들의 광장입니다. (존댓말)

처음으로 | 내가 쓴 글 | 스크랩 | 알림 1 | 쪽지 19 | 배심원

댓글	제목	읽음	날짜
6	남자 매직 한 직모 바가지머리	324	12:54
5	여자 방광염 진료 검사 비용	253	12:52
5	의사님들 질문있습니다. 눈 관련	203	12:47
20	역시 탈고대가 답인가요	1,431	12:44
12	투표 할아버지 핸드폰 s22 vs a33	183	12:44
4	눈물샘 눈물길 막힘? 눈물이 계속 나요.....	376	12:21
44	지금 계량기 보고 있는데, 70만원 실화일까요..?	2,565	12:07
4	상담 어플 추천 부탁드립니다	174	11:54
11	LG 유플러스 개인정보 유출 확인해보세요	1,729	11:42
17	아옳이 하트시그널 서주원 상간녀 소송제기	3,562	11:41
7	직장 병행 가능한 경영대학원 있나요?	359	11:19
11	장학금 이중수혜 관련 질문	579	11:18
23	2주년을 준비 못했습니다..	2,038	11:10
13	(신규가입) 교재 PDF 사기당했어요...	1,662	11:08
4	솔로지옥 후기	746	11:04

01. Concept

제목/본문으로 게시글 분류

- [클러스터링, 감성분석, 게시글 조회수] 기준으로 분류

- 클러스터링 : DBSCAN, Kmeans 군집화 (Word2Vec 임베딩)
- 감성분석 : 사전방식 빈도수 기준
- 조회수 분류 : BiLSTM을 이용하여 22년 게시글 조회수 분류 제목을 작성하면 5분위수 기준 분류값을 내놓음 (무작위 임베딩)

제목으로 게시글 조회수 예측

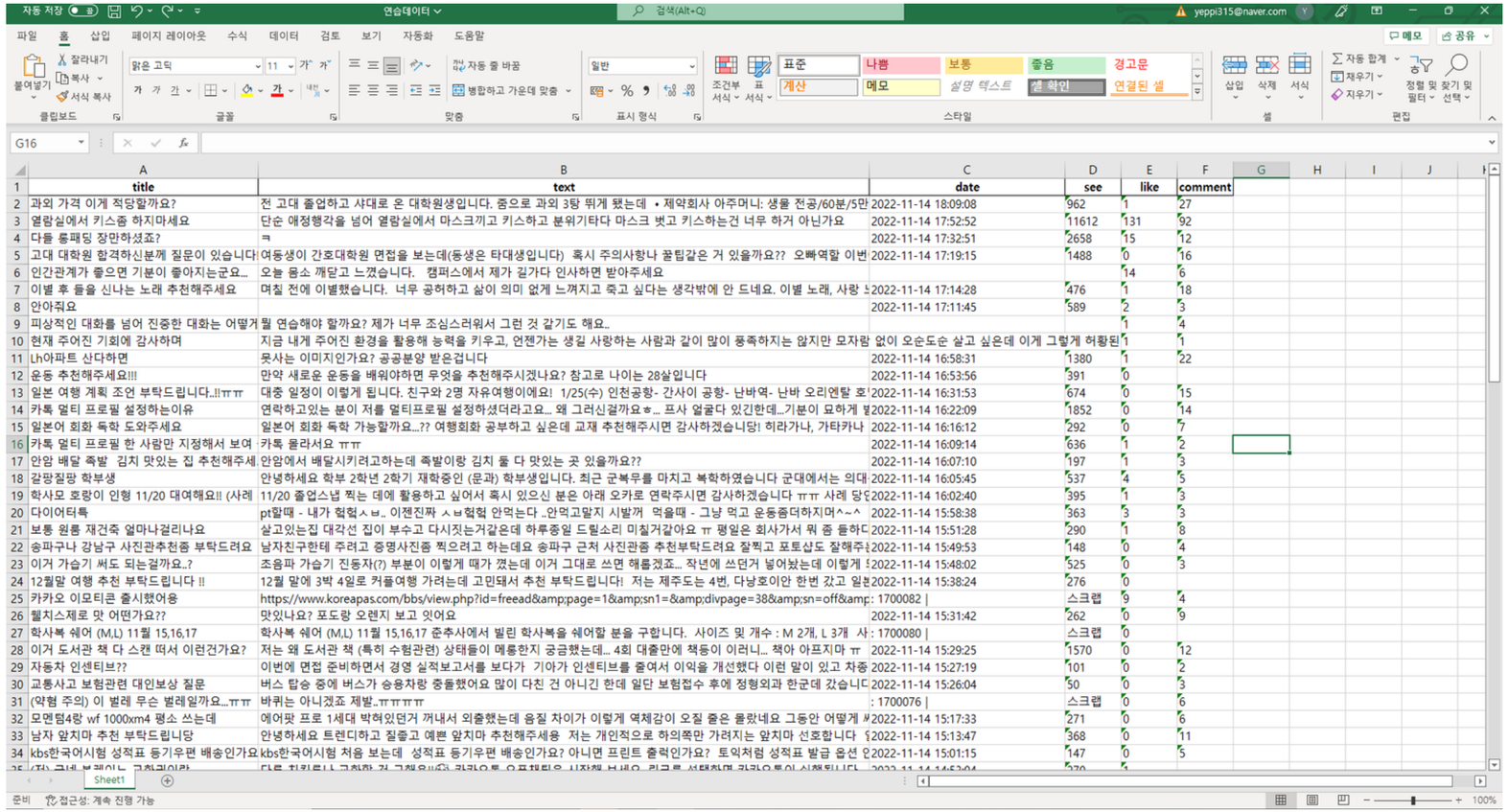
- BiLSTM을 이용하여 22년 게시글 조회수 예측
- 제목을 작성하면 조회수 예측값을 내놓음

02. 데이터

고파스 자유게시판에서 크롤링

beautifulsoup, selenium 통해서 웹크롤링
2022년 게시글 60000개 중 NA제외한 48096개의 제목/본문/업로드시각/조회수/좋아요수/댓글수 (변수 6개)

```python
from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.service import Service

import re
import time

from webdriver_manager.chrome import ChromeDriverManager

driver = webdriver.Chrome(service=Service(ChromeDriverManager().install()))

def remove_html_tags(data):
    p = re.compile(r'<.*?>')
    return p.sub(' ', str(data))

def get_crawl(URL):
    response = driver.get(URL)
    html = driver.page_source
    soup7 = BeautifulSoup(html, 'html.parser')
    ex_id_divs = soup7.select('#revolution_main_table > tbody > tr:nth-child(7) > td:nth-child(3) > a > span')
    crawl_data = remove_html_tags(ex_id_divs)
    return crawl_data

driver.implicitly_wait(3)
driver.get("https://www.koreapas.com/")
print(driver.page_source)

time.sleep(1)

# 로그인 버튼 XPath
login_x_path = '/html/body/div/div[5]/div[2]/div/table[2]/tbody/tr/td[2]/form/div/div/input'
# ID 칸 css selector
driver.find_element(By.CSS_SELECTOR, 'body > div > div:nth-child(7) > div:nth-child(2) > div > table:nth-child(2) > \
tbody > tr > td:nth-child(2) > form > table > tbody > tr:nth-child(1) > td:nth-child(1) > input').send_keys("자기 아이디")
time.sleep(1)
# PW 칸 css selector
driver.find_element(By.CSS_SELECTOR, 'body > div > div:nth-child(7) > div:nth-child(2) > div > table:nth-child(2) > \
tbody > tr > td:nth-child(2) > form > table > tbody > tr:nth-child(2) > td > input').send_keys("자기 비밀번호")
time.sleep(1)
driver.find_element(By.XPATH, login_x_path).click()
time.sleep(3)

main_url = "https://www.koreapas.com/bbs/zboard.php?id=memto"
data = get_crawl(main_url)

print(data)
quit()
```



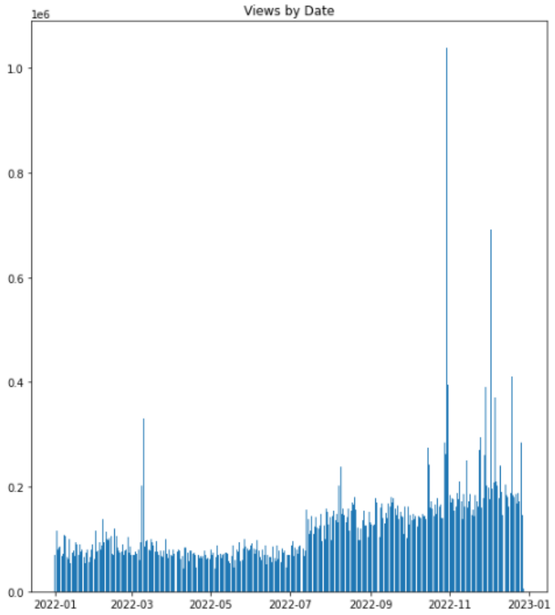
	title	text	date	see	like	comment
2	과외 가격 이게 적당할까요?	전 고대 졸업하고 샤대로 온 대학원생입니다. 줌으로 과외 3탕 뛰게 됐는데 • 제약회사 아주머니: 생물 전공/60분/5만	2022-11-14 18:09:08	962	1	27
3	열람실에서 키스좀 하지마세요	단순 애정행각을 넘어 열람실에서 마스크끼고 키스하고 분위기타다 마스크 벗고 키스하는건 너무 하거 아닌가요	2022-11-14 17:52:52	11612	131	92
4	다들 롱패딩 장만하셨죠?	ㅋ	2022-11-14 17:32:51	2658	15	12
5	고대 대학원 합격하신분께 질문이 있습니다	여동생이 간호대학원 면접을 보는데(동생은 타대생입니다) 혹시 주의사항나 꿀팁같은 거 있을까요?? 오빠역할 이번	2022-11-14 17:19:15	1488	0	16
6	인간관계가 좋으면 기분이 좋아지는군요...	오늘 용소 깨닫고 느꼈습니다. 캠퍼스에서 제가 길가다 인사하면 받아주세요			14	6
7	이별 후 들을 신나는 노래 추천해주세요	며칠 전에 이별했습니다. 너무 공허하고 삶이 의미 없게 느껴지고 죽고 싶다는 생각밖에 안 드네요. 이별 노래, 사랑	2022-11-14 17:14:28	476	1	18
8	안아줘요		2022-11-14 17:11:45	589	2	3
9	피상적인 대화를 넘어 진중한 대화는 어떻게	뭘 연습해야 할까요? 제가 너무 조심스러워서 그런 것 같기도 해요..			1	4
10	현재 주어진 기회에 감사하며	지금 내게 주어진 환경을 활용해 능력을 키우고, 언젠가는 생길 사랑하는 사람과 같이 많이 풍족하지는 않지만 모자람	없이 오순도순 살고 싶은데 이게 그렇게 허황된	1	1	
11	Lh아파트 산다하면	못사는 이미지인가요? 공공분양 받은겁니다	2022-11-14 16:58:31	1380	1	22
12	운동 추천해주세요!!!	만약 새로운 운동을 배워야하면 무엇을 추천해주시겠나요? 참고로 나이는 28살입니다	2022-11-14 16:53:56	391	0	
13	일본 여행 계획 조언 부탁드립니다..!!ㅠㅠ	대충 일정이 이렇게 됩니다. 친구와 2명 자유여행이에요! 1/25(수) 인천공항- 간사이 공항- 난바역- 난바 오리엔탈 호	2022-11-14 16:31:53	674	0	15
14	카톡 멀티 프로필 설정하는이유	연락하고있는 분이 저를 멀티프로필 설정하셨더라고요... 왜 그러신걸까요ㅎ... 프사 얼굴다 있긴한데...기분이 묘하게	2022-11-14 16:22:09	1852	0	14
15	일본어 회화 독학 도와주세요	일본어 회화 독학 가능할까요...?? 여행회화 공부하고 싶은데 교재 추천해주시면 감사하겠습니당! 히라가나, 가타카나	2022-11-14 16:16:12	292	0	7
16	카톡 멀티 프로필 한 사람만 지정해서 보여	카톡 몰라서요 ㅠㅠ	2022-11-14 16:09:14	636	1	2
17	안암 배달 족발 김치 맛있는 집 추천해주세	안암에서 배달시키려고하는데 족발이랑 김치 둘 다 맛있는 곳 있을까요??	2022-11-14 16:07:10	197	1	3
18	갈팡질팡 학부생	안녕하세요 학부 2학년 2학기 재학중인 (문과) 학부생입니다. 최근 군복무를 마치고 복학하였습니다 군대에서는 의대	2022-11-14 16:05:45	537	4	5
19	학사모 호랑이 인형 11/20 대여해요!! (사례	11/20 졸업스냅 찍는 데에 활용하고 싶어서 혹시 있으신 분은 아래 오카로 연락주시면 감사하겠습니다 ㅠㅠ 사례 당	2022-11-14 16:02:40	395	1	3
20	다이어터특	pt할때 - 내가 헉헉ㅅㅂ.. 이젠진짜 ㅅㅂ헉헉 안먹는다 ..안먹고말지 시발꺼 먹을때 - 그냥 먹고 운동좀더하지며^~^	2022-11-14 15:58:38	363	3	3
21	보통 원룸 재건축 얼마나걸리나요	살고있는집 대각선 집이 부수고 다시짓는거같은데 하루종일 드릴소리 미칠거같아요 ㅠ 평일은 회사가서 뭐 좀 들하다	2022-11-14 15:51:28	290	1	8
22	송파구나 강남구 사진관추천좀 부탁드려요	남자친구한테 주려고 증명사진좀 찍으려고 하는데요 송파구 근처 사진관좀 추천부탁드려요 잘찍고 포토샵도 잘해주	2022-11-14 15:49:53	148	0	4
23	이거 가습기 써도 되는걸까요..?	초음파 가습기 진동자(?) 부분이 이렇게 때가 꼈는데 이거 그대로 쓰면 해롭겠죠... 작년에 쓰던거 넣어놨는데 이렇게	2022-11-14 15:48:02	525	0	3
24	12월말 여행 추천 부탁드립니다 !!	12월 말에 3박 4일로 커플여행 가려는데 고민돼서 추천 부탁드립니다! 저는 제주도는 4번, 다낭호이안 한번 갔고 일본	2022-11-14 15:38:24	276	0	
25	카카오 이모티콘 출시했어용	https://www.koreapas.com/bbs/view.php?id=freead&page=1&sn1=&divpage=38&sn=off&	: 1700082 |	스크랩	9	4
26	웰치스제로 맛 어떤가요??	맛있나요? 포도랑 오렌지 보고 잇어요	2022-11-14 15:31:42	262	0	9
27	학사복 쉐어 (M,L) 11월 15,16,17	학사복 쉐어 (M,L) 11월 15,16,17 준추사에서 빌린 학사복을 쉐어할 분을 구합니다. 사이즈 및 개수 : M 2개, L 3개 사	: 1700080 |	스크랩	0	
28	이거 도서관 책 다 스캔 떠서 이런건가요?	저는 왜 도서관 책 (특히 수험관련) 상태들이 메롱한지 궁금했는데... 4회 대출만에 책등이 이러니... 책아 아프지마 ㅠ	2022-11-14 15:29:25	1570	0	12
29	자동차 인센티브??	이번에 면접 준비하면서 경영 실적보고서를 보다가 기아가 인센티브를 줄여서 이익을 개선했다 이런 말이 있고 차종	2022-11-14 15:27:19	101	0	2
30	교통사고 보험관련 대인보상 질문	버스 탑승 중에 버스가 승용차랑 충돌했어요 많이 다친 건 아니긴 한데 일단 보험접수 후에 정형외과 한군데 갔습니다	2022-11-14 15:26:04	50	0	3
31	(약혐 주의) 이 벌레 무슨 벌레일까요...ㅠㅠ	바퀴는 아니겠죠 제발..ㅠㅠㅠㅠ	: 1700076 |	스크랩	0	6
32	모멘텀4랑 wf 1000xm4 평소 쓰는데	에어팟 프로 1세대 박혀있던거 꺼내서 외출했는데 음질 차이가 이렇게 역체감이 오질 줄은 몰랐네요 그동안 어떻게	2022-11-14 15:17:33	271	0	6
33	남자 앞치마 추천 부탁드립니당	안녕하세요 트렌디하고 질좋고 예쁜 앞치마 추천해주세용 저는 개인적으로 하의쪽만 가려지는 앞치마 선호합니다	2022-11-14 15:13:47	368	0	11
34	kbs한국어시험 성적표 등기우편 배송인가요	kbs한국어시험 처음 보는데 성적표 등기우편 배송인가요? 아니면 프린트 출력인가요? 토익처럼 성적표 발급 옵션	2022-11-14 15:01:15	147	0	5

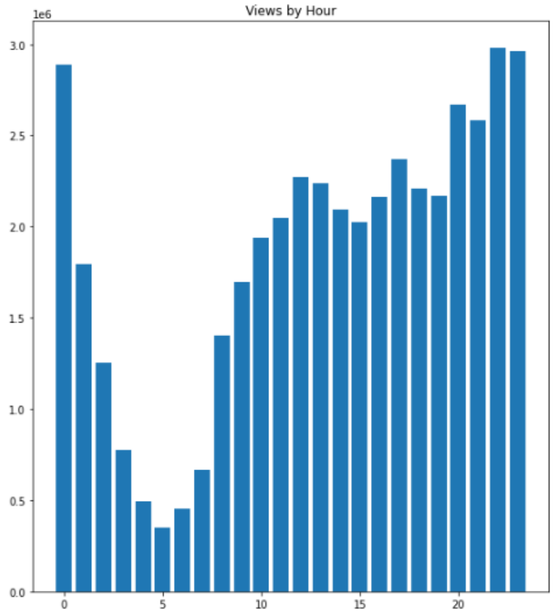
03. EDA

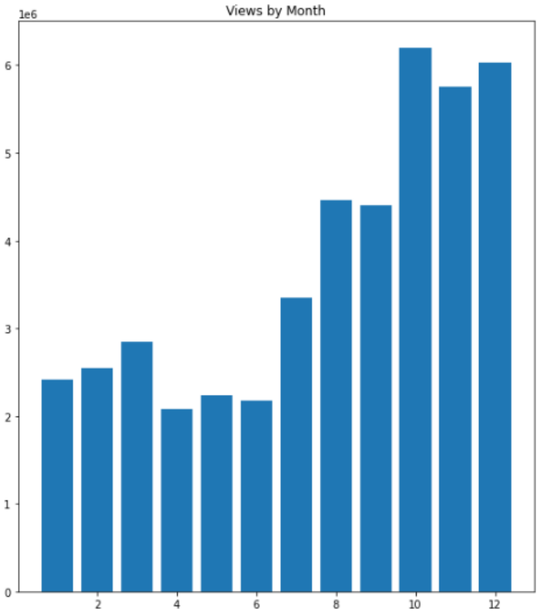
1. 일별/시간대별/월별 조회수

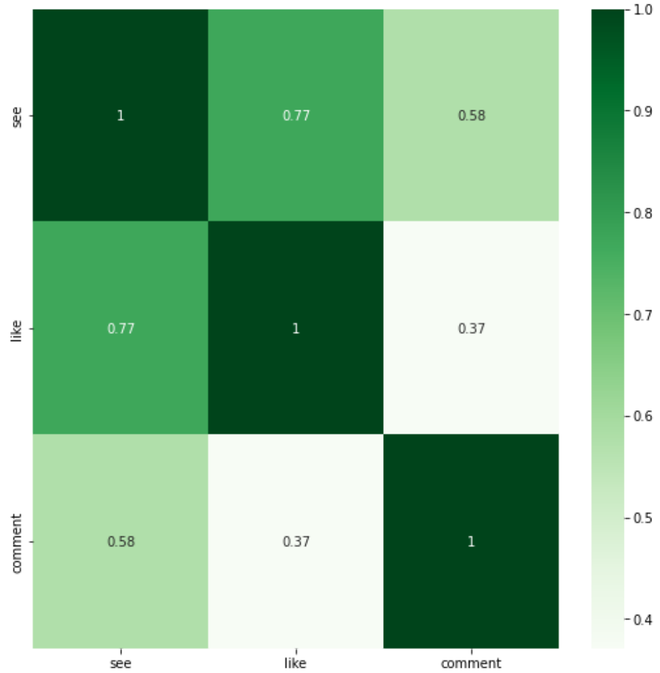
7월 중순 기점으로 불연속 증가세를 유지 -> selection bias 의심
월별 게시글수, 월별 NA비율이 동일한 수준 -> 표본추출 과정에서 생긴 오류 X, 데이터 내부의 요인으로 추정

조회수,추천수,댓글수 간 high correlation -> multi-target 아닌, '조회수' 변수만 only target









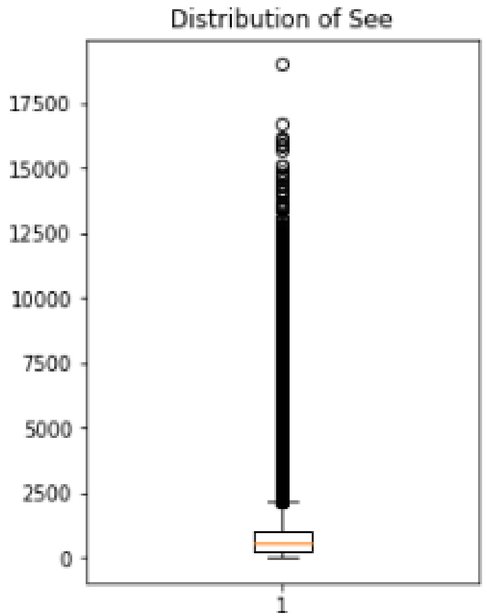
03. EDA

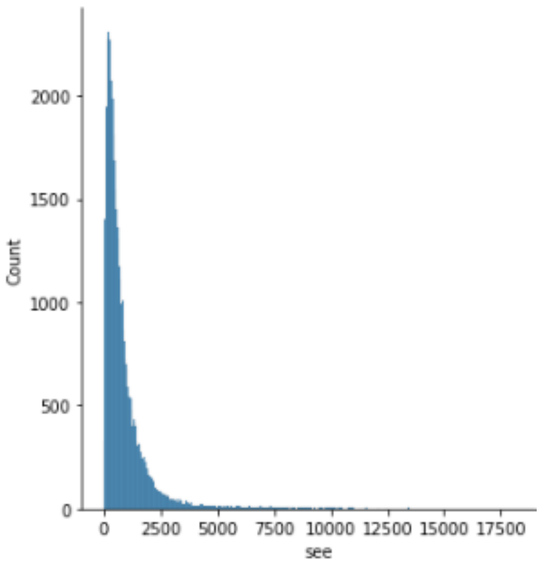
2. 형태소 기준 빈도수 분석 : 상하위 조회수 기준

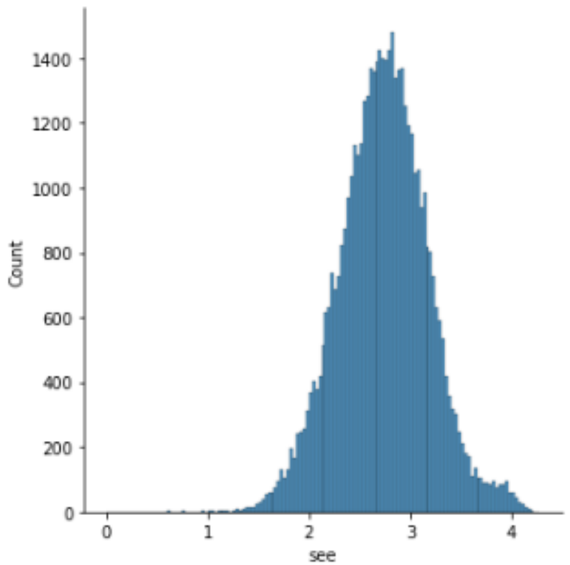
boxplot 결과 outlier가 많지만, 분석목적은 '상위 조회수 게시글의 분석'이므로 outlier유지
outlier의 영향력을 살펴보고자 기존분포를 로그화 해주었더니 정규분포와 유사

형태소 단위 토큰화(Okt) -> 불용어 제거 -> 상위 1%, 5%, 하위 10% 기준 최빈 형태소 분석

```
불용어 제거 전 : ['한국', '교회', '는', '남', '들', '에게', '민폐', '를', '끼쳐서', '조롱', '받는', '거']
불용어 제거 후 : ['한국', '교회', '남', '민폐', '끼쳐서', '조롱', '받는']
```





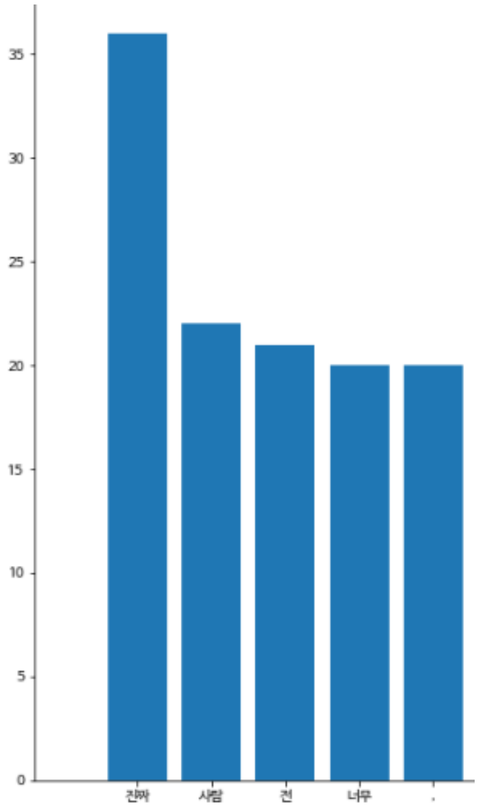


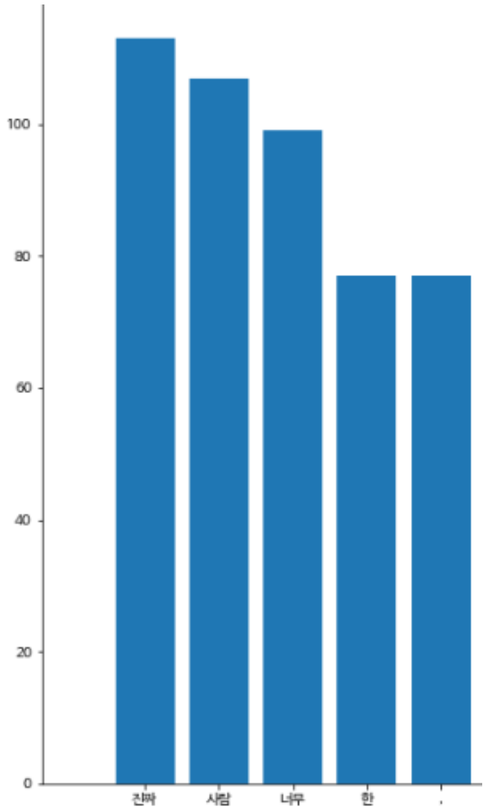
03. EDA

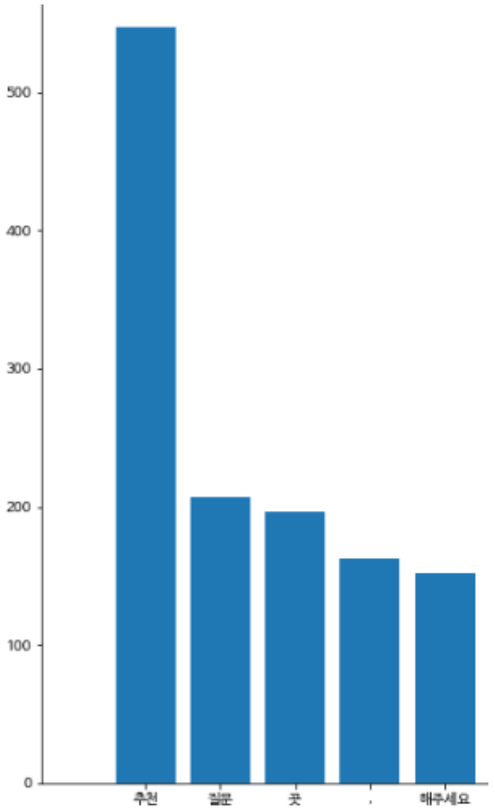
2. 형태소 기준 빈도수 분석 : 상하위 조회수 기준

상위 1%, 5% : '진짜', '너무' 의 빈도가 높음. 과장하는 수사어구를 넣은 제목이 조회수가 높게 나옴.

하위 10% : '추천', '질문', '곳', '해주세요' 의 빈도가 높음.
해당 목적으로 사용자에게 적극적인 행동을 요구하는 제목은 조회수가 적게 나옴.







04-1. 분류 : BiLSTM을 이용한 '인기게시물' 분류

1. RNN(BiLSTM)을 이용한 다중클래스 분류

조회수 분포를 로그화한 분포의 z-score값을 예측변수로 정의함.
로그화한 분포의 z-score 값에서 20% 간격으로 class 지정. (5분위수)

Vectorization : 각 단어에 해당하는 인덱스를 할당해줌
Embedding : random 초기값에서 학습값에 따라 역전파되는 기울기를 바탕으로 조정
Model : BiLSTM (dropout = 0.2)

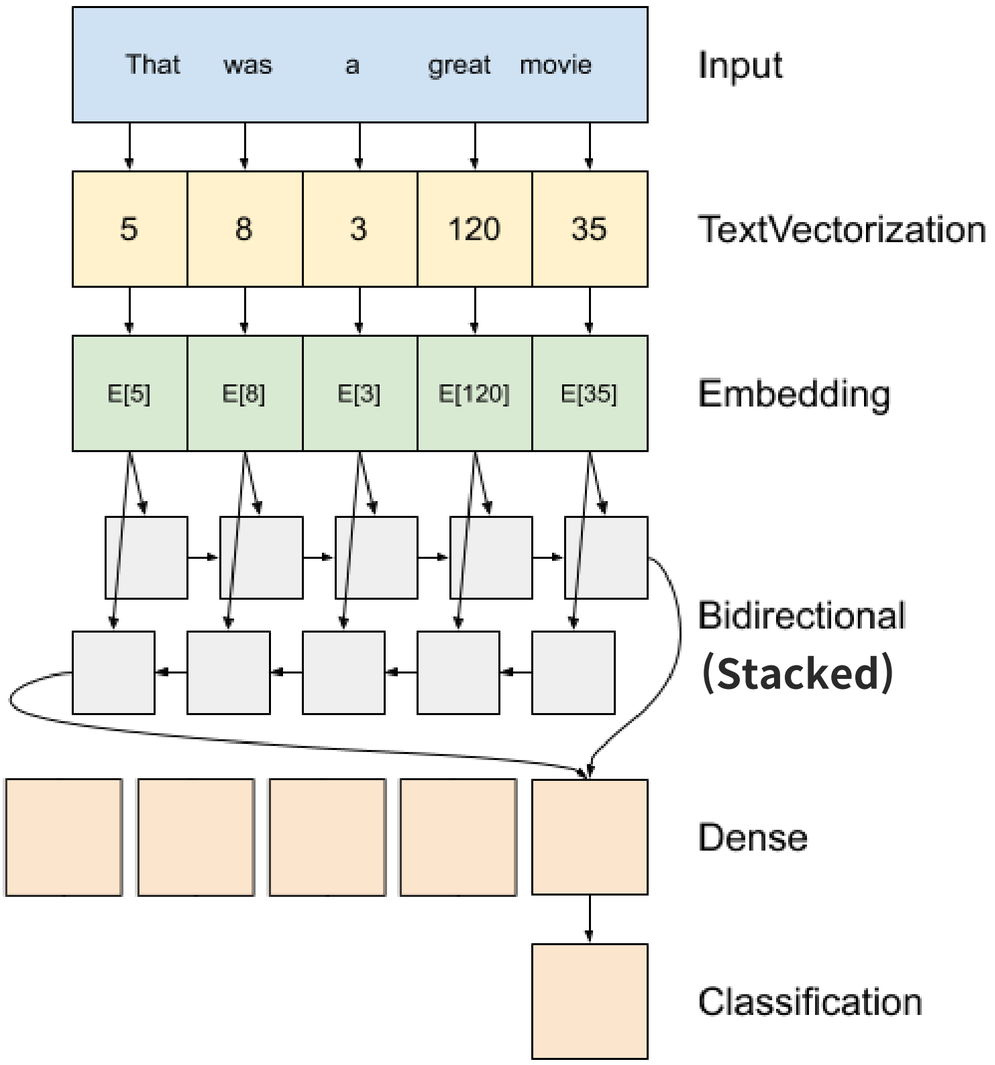
Optimizer function : Adam
Loss function : Categorical CrossEntropy

```
[ ] X_test_raw.iloc[10]
```

```
title    [11, 학번, 인데]
Name: 43163, dtype: object
```

```
[ ] X_test[10]
```

```
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
         0,   0,   0,   0, 963, 273,  16], dtype=int32)
```



04-1. 분류 : BiLSTM을 이용한 '인기게시물' 분류

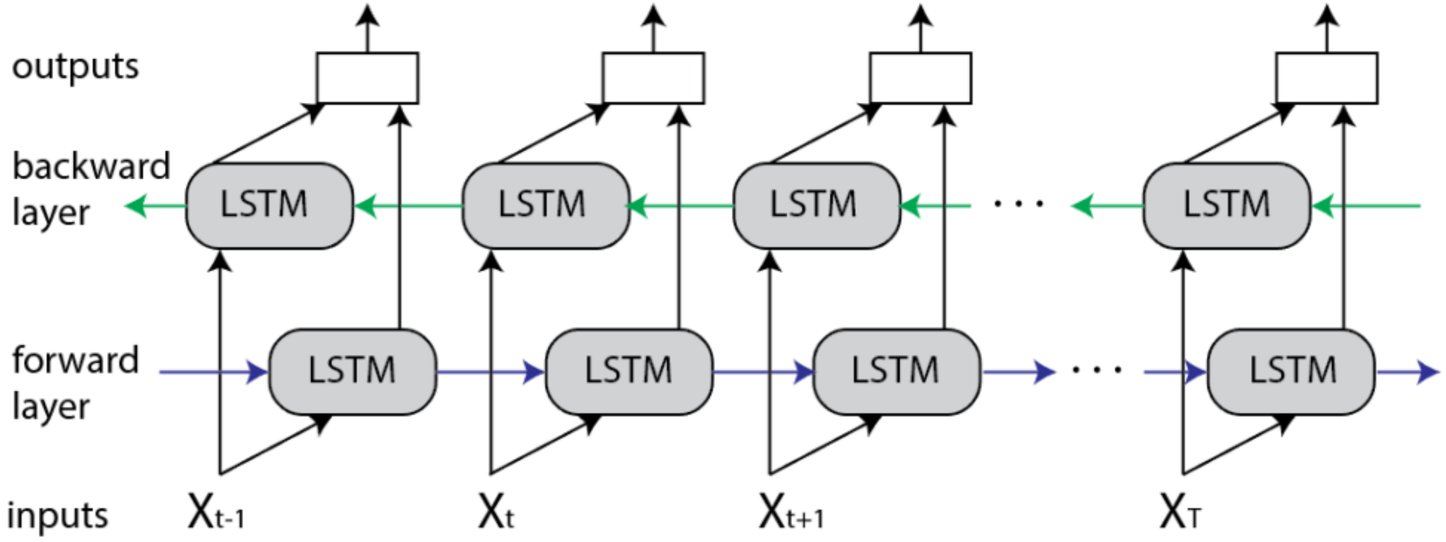
2. simple RNN의 문제점과, BiLSTM, Stacked RNN

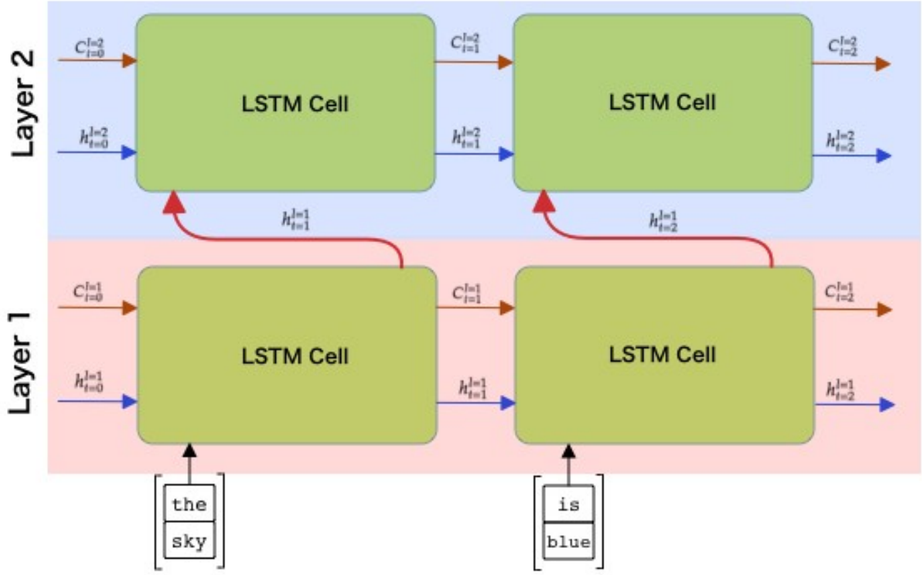
- simple RNN의 문제점 :

1. 입력순서가 순차적이기 때문에, output이 직전 input에 영향을 크게 받음
   이것은 데이터 길이가 길고, 층이 깊어질수록 과거의 정보가 손실되는 문제를 초래.
2. Vanishing Gradient problem : 하이퍼볼릭 탄젠트가 -1과 1사이이기 때문에,
   시퀀스가 길고 층이 깊어질수록 gradient가 소멸되는 현상.

$$
\begin{aligned}
h_t &= \tanh\left(W_{hh}h_{t-1} + W_{xh}x_t + b_h\right) \\
&= \tanh(z_t)
\end{aligned}
$$

- BiLSTM : 기존 LSTM layer에 역방향으로 처리하는 LSTM layer를 추가하여, 정보의 손실을 방지.
- Stacked RNN : 이전 레이어의 hidden state vector가 다음 레이어의 input이 됨.





04-1. 분류 : BiLSTM을 이용한 '인기게시물' 분류

3. 결과값 비교와 시사점

- 임베딩 벡터의 차원의 크기변화는 성능에 유의미한 영향을 주지 못했음.
- hidden layer 에 ReLU 함수 적용시 accuracy 하락.
- Stacked RNN 적용시 accuracy 유의미한 상승.

	index	title	0	1	2	3	4
0	15796	[콜라보, 레이, 트는, 이어폰, 사용, 불가, 인가요]	0.424726	0.277771	0.173817	0.082397	0.041288
1	34668	[회장, 님, 선물]	0.543072	0.279368	0.124852	0.038139	0.014569
2	33562	[타투, 스티커, 어디서, 사나]	0.221236	0.333329	0.254268	0.142479	0.048688
3	36799	[뒷북, 일본, 옛날, 여배우, 중, 누가, 제일, 이쁘다고, 생각, 하나요]	0.027047	0.093041	0.153218	0.335933	0.390761
4	22782	[이사, 톤, 트럭, 불렀고, 용달, 기사, 님, 짐, 옮기는거, 해주셨는데, 비용...	0.243707	0.336876	0.252116	0.122336	0.044964

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_9 (Embedding)     (None, None, 256)         6956544

 bidirectional_10 (Bidirecti  (None, None, 256)        394240
 onal)

 bidirectional_11 (Bidirecti  (None, 256)              394240
 onal)

 dense_12 (Dense)            (None, 5)                 1285

=================================================================
```

```
ttl4='입실렌티 티켓 양도합니다'
ttl5='주식 관련 세금 납부 질문이요!'
ttl6='담배 필거면 맛있게 펴라...'
ttl7='남친이 연애 초반보다 확실히 변한 것 같아서..'
```

```
1/1 [==============================] - 0s 39ms/step
[[0.08205146 0.13956986 0.23512551 0.29031023 0.25294286]]
4  번째 Qunatile에 게시글이 분류되었습니다!

1/1 [==============================] - 0s 39ms/step
[[0.51611507 0.31497014 0.11869733 0.03871565 0.01150176]]
1  번째 Qunatile에 게시글이 분류되었습니다!

1/1 [==============================] - 0s 41ms/step
[[0.1825896  0.22214699 0.24951018 0.20524837 0.1405048 ]]
3  번째 Qunatile에 게시글이 분류되었습니다!

1/1 [==============================] - 0s 53ms/step
[[0.00418337 0.01003826 0.04767931 0.20764607 0.73045295]]
5  번째 Qunatile에 게시글이 분류되었습니다!
```

04-2. 예측 : BiLSTM을 이용한 '조회수' 예측

1. RNN(BiLSTM)을 이용한 예측

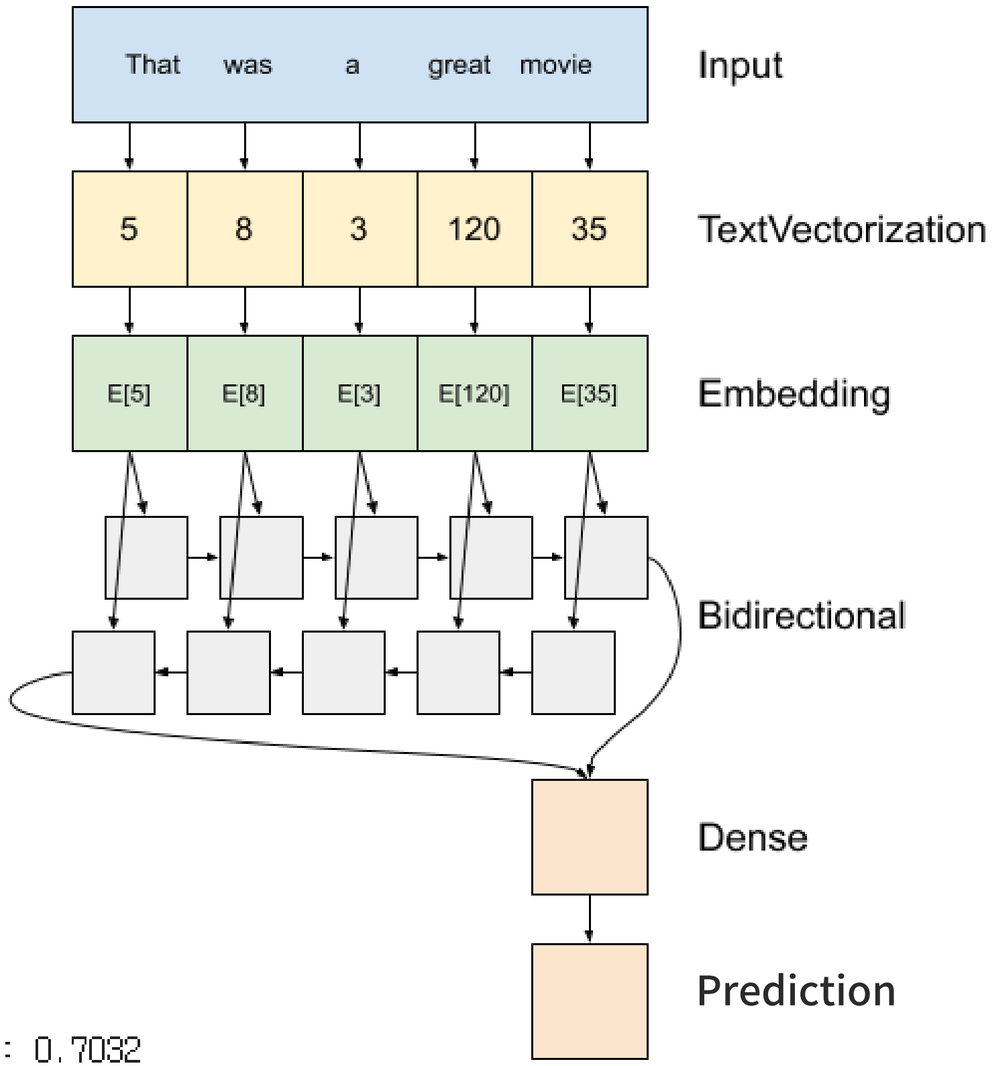
조회수 분포를 로그화한 분포의 z-score값을 예측변수로 정의함.

Vectorization : 각 단어에 할당하는 인덱스를 할당해줌
Embedding : random 초기값에서 학습값에 따라 역전파되는 기울기를 바탕으로 조정
Model : BiLSTM (dropout = 0.2)

Optimizer function : Adam
Loss function : MSE

- 분류 : 5개의 레이블 (softmax를 통해 0~1 사이 확률값을 반환)
- 예측 : 1개의 레이블

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, None, 256)         7680256

 bidirectional (Bidirectiona  (None, 256)              394240
 l)

 dense (Dense)               (None, 1)                 257

=================================================================
```



테스트 오차(MSE): 0.7032

04-2. 예측 : BiLSTM을 이용한 '조회수' 예측

2. 결과값 비교와 시사점

- Stacked RNN 유의미한 성능변화를 가져오지 않았음.
- 분포가 정확한 정규분포와 차이가 있어 오차 발생. 조회수가 높은 게시글이 예측성능 우수.
- 통계기반 임베딩(TF-IDF), 혹은 문맥 고려 임베딩(Word2Vec) 없이도 준수한 성능 구현.

10	2분 전 읽음 바람난 행정고시랑 싸우고 남친 합격했습니다	1,931	2380.47
2	졸리당	113	461.02
1	안암 헬스장 추천해주세요	248	306.18
36	코로나 아직 한번도 안걸린 사람 있나요?	1,529	518.80
11	CPA 준비하는데 연애 다들 어떻게 하세요?	1,037	1269.95
9	사당역 오랜만에 가봤는데 많이 바뀌었더군요	1,084	519.99
	신년맞이 책 추천 부탁드립니다!!!	78	177.94

```
1/1 [==============================] - 0s 23ms/step
바람난 행정고시랑 싸우고 남친 합격했습니다
Log-Z Score : 1.452291488647461, 조회 수 : 2380.4765144505486

1/1 [==============================] - 0s 23ms/step
졸리당
Log-Z Score : -0.16481103003025055, 조회 수 : 461.0233669308691

1/1 [==============================] - 0s 33ms/step
안암 헬스장 추천해주세요
Log-Z Score : -0.5679559111595154, 조회 수 : 306.1866257464352

1/1 [==============================] - 0s 22ms/step
코로나 아직 한번도 안걸린 사람 있나요?
Log-Z Score : -0.04850289225578308, 조회 수 : 518.800459882825

1/1 [==============================] - 0s 22ms/step
CPA 준비하는데 연애 다들 어떻게 하세요?
Log-Z Score : 0.833354651927948, 조회 수 : 1269.9592408520261

1/1 [==============================] - 0s 23ms/step
사당역 오랜만에 가봤는데 많이 바뀌었더군요
Log-Z Score : -0.04623974859714508, 조회 수 : 519.9937420668863

1/1 [==============================] - 0s 24ms/step
신년맞이 책 추천 부탁드립니다!!!
Log-Z Score : -1.1025607585906982, 조회 수 : 177.94754894585202
```

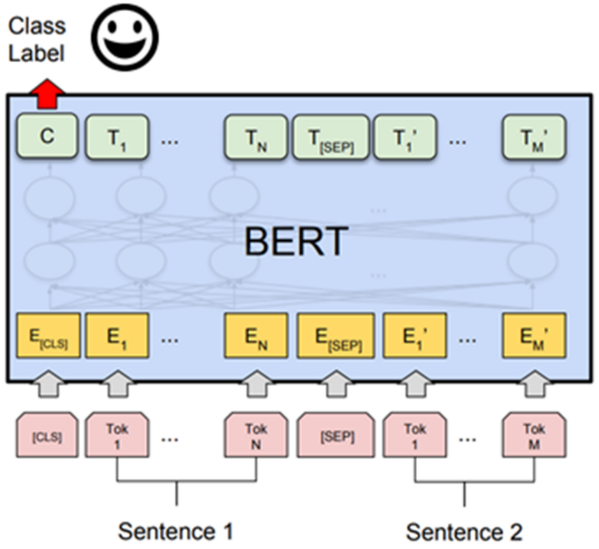
04-3. 분류 : 감성분석

1. 감성분석의 종류와 선정이유

모델기반 감성분석은 일일히 긍정/부정 어휘를 일일히 라벨링 해줘야 하기에, 분석의 용이성을 위해 사전기반 분석 채택

모델기반 감성분석

labeling 되어있는 데이터를
기반으로 모델을 구축해 감성을 예측함.



규칙/어휘 기반 감성분석

감성 사전을 기반으로 데이터의
감성점수를 매겨 긍정/부정 문장을 분류함.

단어	감성점수	빈도수	합계
Happy	1	3	+3
Sad	-1	2	-2
Great	1	2	2
Ugly	-1	1	-1
합계			+2

04-3. 분류 : 감성분석

2. 빈도수 기준 감성분석

고파스 게시글의 특성을 고려하여 이모티콘, 축약어 등을 포함한 SNS 실험에서 높은 성능을 보여준 KNU 감성사전을 채택
형태소 단위 토큰화(Okt) -> 불용어 제거 -> 감성사전에 따른 점수 부여

	sent	point
0	(-;	1.0
1	(;_;)	-1.0
2	(^^)	1.0
3	(^-^)	1.0
4	(^^*	1.0
...	...	...
14850	갈등	-1.0
14851	의혹	-1.0
14852	내팽개치다	-2.0
14853	횡령	-2.0
14854	불안증	-2.0

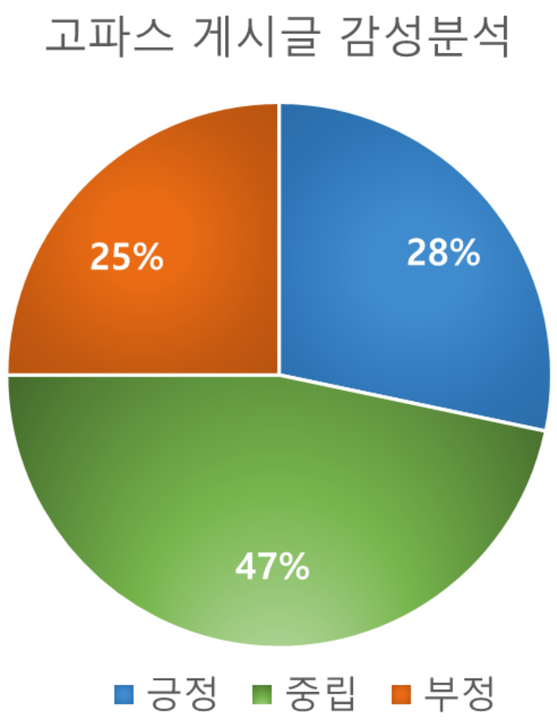
14855 rows × 2 columns

```
Input : 문장 S

if 긍정어휘 수 > 부정어휘 수
        S ← 긍정문장

else 긍정어휘 수 < 부정어휘 수
        S ← 부정문장
```

그림 1 감성분류 기준

```
if 긍정문장
    감성점수 = 긍정어휘 수 - 부정어휘 수

else 부정문장
    감성점수 = 부정어휘 수 - 긍정어휘 수
```

그림 2 감성점수 식



04-4. 군집화 : DBSCAN & K-means를 이용한 클러스터링

1. 워드임베딩

단어 간 유사도 및 중요도 파악을 위해 단어를 저차원의 실수 벡터로 맵핑하여 의미적으로 비슷한 단어를 가깝게 배치하는 자연어 처리 모델링 기술

TF-IDF

– 횟수기반 임베딩으로 핵심어 추출 위해 단어의 특정 문서 내 중요도 산출
– TF: 단어의 문서내 빈도, IDF: 문서 빈도 수(DF)의 역수

Word2vec

– 단어를 벡터 평면에 배치하여 문맥적 의미 보존
– 한국어로 사전 학습된 모델 사용
– 토큰들의 벡터를 평균내어 문장의 벡터를 표현

단어의 종류가 매우 많았기 때문에 TF-IDF는 결과가 좋지 않았음

04-4. 군집화 : DBSCAN & K-means를 이용한 클러스터링

2. DBSCAN을 이용한 클러스터링

- DBSCAN의 장점 :

심각한 outlier가 존재하는 경우, K-Means 의 경우 centroid 를 심각하게 변경하여, 클러스터의 품질 저하 risk 존재
K를 지정할 필요가 없으며, 클러스터의 형태에 영향을 받지 않음. Noise를 따로 분류하여 명확한 분류가 가능.

생성된 군집 별로 군집 크기의 내림차순을 표시

-1	1384
0	1177
4	23
16	21
14	19
3	18
13	15
8	10
7	9
12	7
23	6
22	6
21	6
33	6
6	6

-1의 경우 노이즈로 판별 난 것인데 대부분의 데이터에서 군집화가 이루어지지 않음

클러스터 4, 16, 14, 3 확인

04-4. 군집화 : DBSCAN & K-means를 이용한 클러스터링

2. DBSCAN을 이용한 클러스터링

```
df1[df1['dbscan cat']==4]['all']
```

```
45337                                열정후 해보신 선배님 계신가요??어떤가요??
45391    후참 맛없더라구요후라이드 잘한다구 해서 한번 사먹어봤는데 \n \n그 가격에 맛인는...
45589    변호사님 계신가요 수임료 입금후 현금영수증제가 학생인데 소송하게 돼서요 \n돈입금해...
45607                     열정후 휴면버튼이 어디잇죠?ㅠㅠㅠ못찾겟어요 ㅠㅠ 알려주실분 ㅠㅠ
45715    카톡대화 캡쳐 후 한글파일에카톡대화 캡처한거 한글파일 밑에 첨부하고 싶은데요 \n ...
45946    열정후 등록은 해놓고 한번도 안했는데열어보니까 여자분 무슨 카드가 와있더라구요. \...
46191    열정후로 만나서 젤 오래 만난 커플아마 저희이지 않을까 싶은데 \n곧 300일이네요 캬캬
46285    사래 걸린 후 인후통어제 점심 먹다가 사래가 걸렸는데 \n그 이후로 계속 감기 걸렸...
46444    열정후 현타가 이런건가요 ㅠ에프터 약속 정하고 어제까지 잘 대화하다가 갑자기 끊겨서...
46537    열정후 대화신청수여자분들은 보통 대화신청이 하루에 최소 5-10개 온다는데 \n전 ...
46561    열정후 엠비티아이 자기걸로 나오시나용??전 신기하게 똑같이 나오더라구요ㅋㅋㅋㅋㅋㅋㅋ...
46616    열정후가 만남앱이었어요???  저는 학교에 맛집 앱 있다해서 열정후가 그건줄... ...
46668    탈색후 물빠짐  탈색하고 애쉬바이올렛으로 염색했는데 \n왜 아랫부분이 초록색이 된걸...
46685    열정후로 만나시는 분들 주변에 뭐라고 얘기하나요....?  어플로 사귀는걸 주변에 ...
46971    사랑니 발치후 식사  사랑니 이쁘게 난거 하나 뽑는데 \n아침에 뽑고 저녁에 친구만...
47158                       열정후오예  설레네요  \n기분좋군용귀여우시던데      출
47489    열정후 매칭됐던 사람이랑 또 되기도 하나요?  전에 분명히 본 사진같은데.. \n또...
47519    구형 아이패드 반납후 신형을 저렴하게 살수있나요?  프리스비 매장에서 해주나요?  ...
47559    열정후 매칭이 너무 안되네요  학교 다열고 \n연령대 연하 동갑 연상 다 열어도 \...
47594                       열정후  열정후 하는법이 있나요? 어플인건지..      출
47710    열정후 다들 대화시작하자마자 만남잡으시나요?  저는 어느정도 서로 관심사를 알고  ...
47724    열정후 다들 진지한 관계 찾기위해 하시나요?  아니면 다른 어플만남들처럼  가벼운 ...
48018    열정후 메세지 중복  저만 열정후 메세지가 두 번씩 보내지고 그러나요...ㅜㅜ 그것...
Name: all, dtype: object
```

클러스터 4: 열정후 관련 글

```
df1[df1['dbscan cat']==16]['all']
```

```
45907    아이패드 프로 11 vs 12.9안녕하세요, \n아이패드 프로 구매를 고민 하고 있...
46228    애플워치 어디서 사는게 제일 쌀까요??졸업생이라.. aoc 할인을 못받는다면 그냥 ...
46231    몽블랑 제품 정품 여부안녕하세요,  \n \n이 제품 중고나라에서 사려고 하는데 정...
46480                     중소기업 브랜드 노트북과 티비추천해주실  수 있을까요?
46528    버즈2 vs 버즈프로버즈프로가 훨씬 좋나요? \n버즈라이브 쓰고잇는데 \n귀도 아프...
46549        맥북프로 애플케어플러스 필수인가요이번에 샀는데 애케플이 32만원이라 고민되더라고요
46624    아이패드 애플케어 필수인가요?  엄마가 쓰실 건데 물건 험하게 쓰시는 분이 아니라 ...
46684                     에어팟프로 애플케어살까요?  쓸일 있을까요?      출
46798                          12년만에 공개된 애플코리아 연매출      출
46975    지하철에서 브랜드쇼핑백 여러개 들고 다니시는   할아버지들이 주로 하시던데 무슨 알...
47204    패드나 폰에서 엑셀 볼려고 폴라리스 오피스 깔았는데..  광고가 넘 많고 길어서 빡...
47237                 혹시 유니스토어에 폰케이스.. 파나요??  궁금하네여..      출
47253    맥북 아이패드 애플케어플러스 필요한가요?  기존에 노트북 사용할 때 딱히 3년 안에...
47280    아이패드 에어4사려고하는데  기본옵션으로 하고 애플 펜슬이랑 애플케어만 추가해도 9...
47330         노트북 충전기로 폰 충전해도 되나요?  폰에 무리가 가진 않을까요??      출
47529    맥북 에어 vs 프로 vs 안 산다  현재  \n가벼운 문서작성 \n3분짜리 동영상...
47568    버즈2 갤캠스에서  지금 노티드케이스랑 119000에 팔더라구요 \n케이스는 필요없...
47677    애플워치 셀룰러(sk)  sk데이터 무제한 요금제 쓰는데요 \n(89000원짜리) ...
47800    이마트 매장내 노브랜드와 노브랜드전문점에 가격차이가 있나요?  아니면 단순히 노브랜...
47966    혹시 애플 에어팟 맥스(헤드셋) 아시는분 있을까요?  쿠팡 이런데가 제일 싸게 사는...
48027    맥푹 m1 프로 vs 에어  소규모로 온라인 창업/사업? 을 해보려고 하는데 노트북...
```

클러스터 16: 애플, 맥북, 버즈, 노트북, 폰 등 전자기기 관련 글

04-4. 군집화 : DBSCAN & K-means를 이용한 클러스터링

2. DBSCAN을 이용한 클러스터링

```
df1[df1['dbscan cat']==14]['all']
```

```
45867                                 술마실 때 여자가 손대보자고 하는 거의미있는 행동인가요?
46015    오늘 하루도 보람된 시간을 보내셨나요?하루하루 힘내세요!!! \n \n저스스로한테 ...
46167       오늘 산책하기 좋을까요?마실거 간단히 가져가서 동네천에 걸으려는데 많이 추울까요?
46461                                 아이패드 병 치유했습니다사과뭐시기가 좋긴 좋네요
46620            오늘 자전거나 동네천 산책가기 괜찮은 날씨인가요  날씨요정 없나요      출
46778    항우울제 하루 안먹고 술마셔도 되나요???  맥주 마실듯한데 하루정도는 괜찮을까요 ...
46905                                 오늘 sk미래관 안여나요?  아시는분      출
46927    하루 일과 루틴 알려주실수 있나요  방학 알차게 보내고 싶어요. \n \n예) \n...
47097    아 백신 어제 맞고 하루종일 컨디션 엉망이네요ㅠㅠ  어제는 괜찮았는데 오늘은 하루종...
47290    오늘 자전거  타러가려고 그저께부터 맘먹었었는데..... \n하ㅜㅜㅜㅜㅜㅜㅜㅜㅜ  ...
47409                              오늘 토익 보신 분!!  후기를 들려주세요      출
47417    술 잘드시는 분들(= 주량 3병이상) 궁금한거 있는데요  3병 이상 정도 드시는 분...
47700                   스걸파 오늘 생방이었나요???  무대 전부 다 생방이었나요?      출
47735    어금니 레진이 하루만에 떨어질수 있나요..?  오늘 낮에 레진으로 어금니 충치 떼웠...
47794    남친 술먹고  남친이 이성+동성들과 술자리를 가졌습니다 \n9시에 술집에서 나온것 ...
47893    술한잔 하고 잘까요 그냥 잘까요  본인 수험생 \n오늘 공부 끝남 \n \n내일 오...
47911    술 무엇이든 물어보세요  현직 술관련 업자입니다.  \n \n술에 관한 질문만 부탁...
48041                        혹시 오늘 안암 올리브영 문 열었나요??  ??????      출
48066    영철버거 오늘도 하나요?  경기도인데 갑자기 생각나서.. \n먼길이지만 열었으면 가...
```

클러스터 14: 질문 관련 글

```
df1[df1['dbscan cat']==3]['all']
```

```
45336                          93~96년생 여성분들!혹시 현재 남친이나 배우자가 있으신가요?
45858    변시 끝나고 연애 많이 하시나요?연애는 하고 싶은데 막상 소개팅 해준다고 하면 부담...
46047    잠수이별잠수이별 해보신 분 계세요? \n그러는 심리가 뭔지 이해가 안되서 왜 그러는...
46123    여성분들은 왜 덴탈마스크를 선호하시나요?덴탈 쓰시는분들 중 압도적으로 여성분들 비율...
46443    연애가 너무 하고 싶어요매일 일상 공유하고  \n티키타카 장난도 치고 \n손 잡고 ...
46458    그해우리는 같은 학창시절 연애를 했어요ㅎㅎ제가 딱 고2~3때 연애한거랑 너무 비슷하...
46590    여성분들 군생활 간접경험하고 싶으시면뷰티풀군바리 한 번 봐보세요 \n남녀모두 군대가...
46714    여자 연애? 관련  이쁘고 성격 좋으면 주변에서 가만 안 둔다고 하잖아요 \n솔직히...
46964    이재명이 과연 여성들을 대변해줄까요?  해당 게시물은 고파스 배심원 평결로 제재받았...
47300                   고학번인데 연애하고 싶어요  학교에 남아있으니 쉽지 않네요      출
47371    연애하냐고 물어보는 거  별로 친하지 않은 사이에서도 이 질문 많이 하시나요? \n...
47461    비대면 연애   얼굴을 보지 않은 상태에서 하는 사랑이 가능하다 생각하시나요? 오로...
47620    연애 경험 적은 분들  손 들어 보세요… \n분명 계시겠죠 위안 좀 얻으렵미다 \n...
47621    장거리연애 이별 어떻게 하나요?  전화로 말하는건 예의가 아닌 것 같고 \n주말에 ...
47641    (연애관련) 남자들은 부익부빈익빈이 심한 것 같아요  잘생(상위 15퍼) 키 177...
47803    삼수생이랑 연애  동갑이고 올해 삼수하는 친구랑 연애… \n안하는게 맞죠? \n저는...
47884                             다들 연애하시나요?  저는 솔로입니다 호호      출
48042    연애!! 하고!! 싶어요!!!!!!  다음주부터 소개팅 3개 잡혀있긴 한데 \n경험...
```

클러스터 3: 연애 관련 글

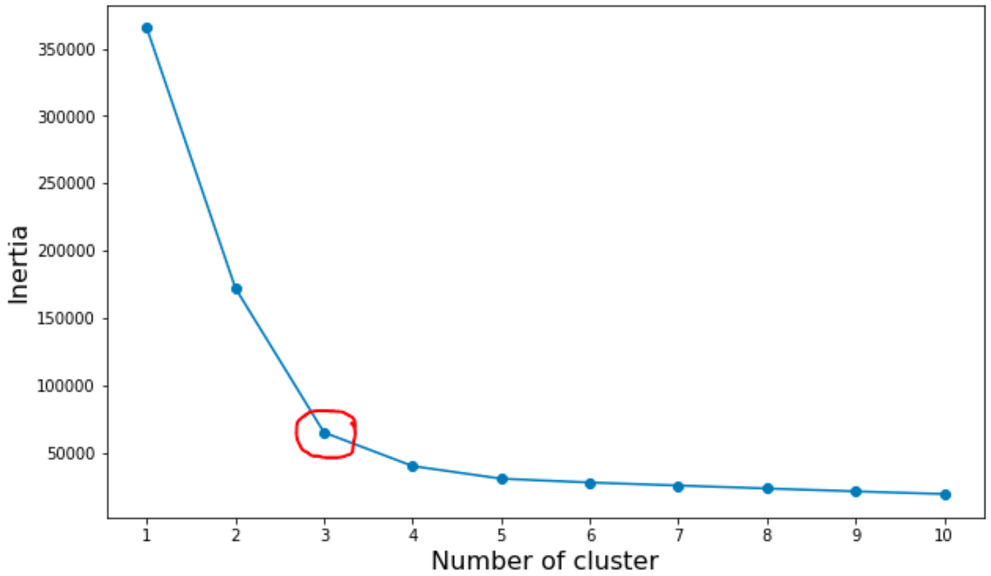
대부분은 군집화 되지 않은 데이터들이지만
생성된 군집은 동일한 주제의 글임을 확인

04-4. 군집화 : DBSCAN & K-means를 이용한 클러스터링

3. Kmeans를 이용한 클러스터링

군집의 수 k를 지정해줘야하는데 최적의 k를 찾는 두가지 방법 시도

Elbow method



Cluster 간의 거리의 합을 나타내는 inertia가 급격히 떨어지는
구간이 생기는데 이 지점의 K 값을 군집의 개수로 사용

04-4. 군집화 : DBSCAN & K-means를 이용한 클러스터링

3. Kmeans를 이용한 클러스터링

군집의 수 k를 지정해줘야하는데 최적의 k를 찾는 두가지 방법 시도

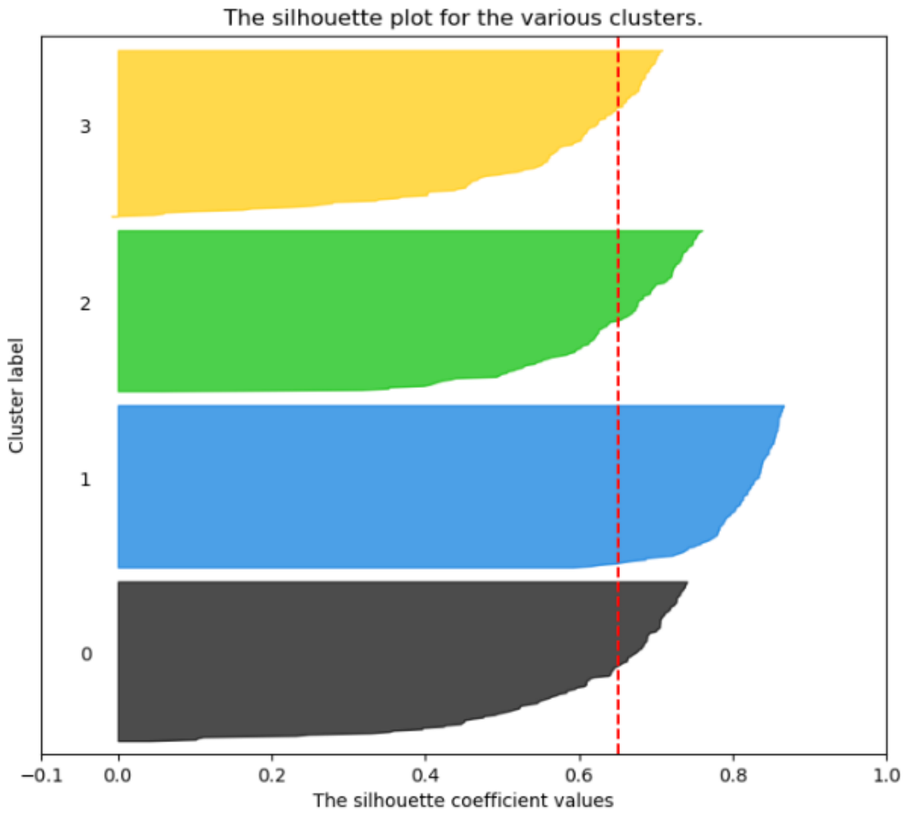
Silhouette Score

$$s(i) = \frac{(b(i)-a(i))}{(max(a(i),b(i)))}$$

a(i): 데이터 포인트 i가 속한 클러스터 내
데이터 포인트들과의 거리 평균

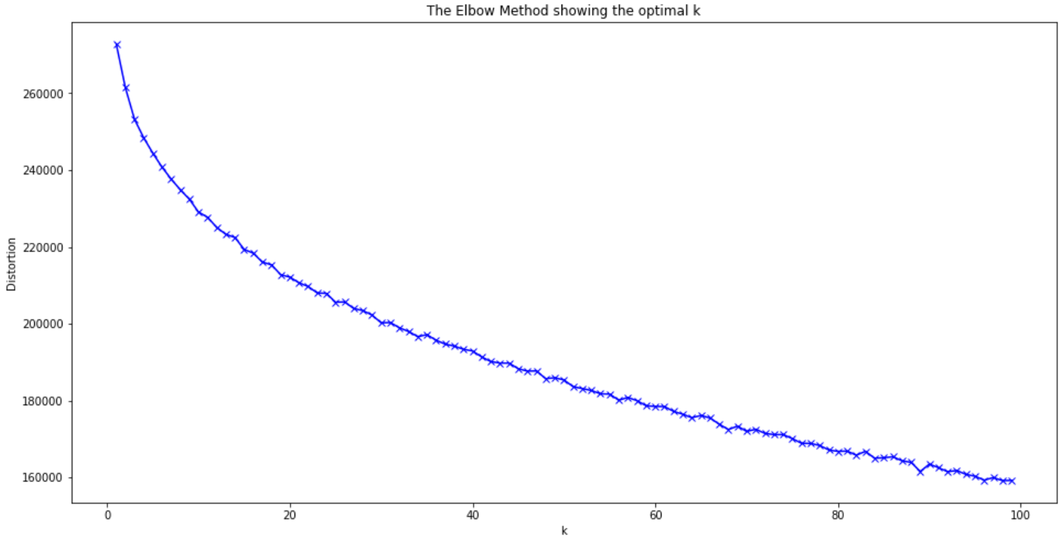
b(i): 데이터포인트 i가 속하지 않은 클러스터의
데이터 포인트들과 거리평균의 최솟값

silhouette score 가 1에 가까이 커질 수록 좋다고 판단

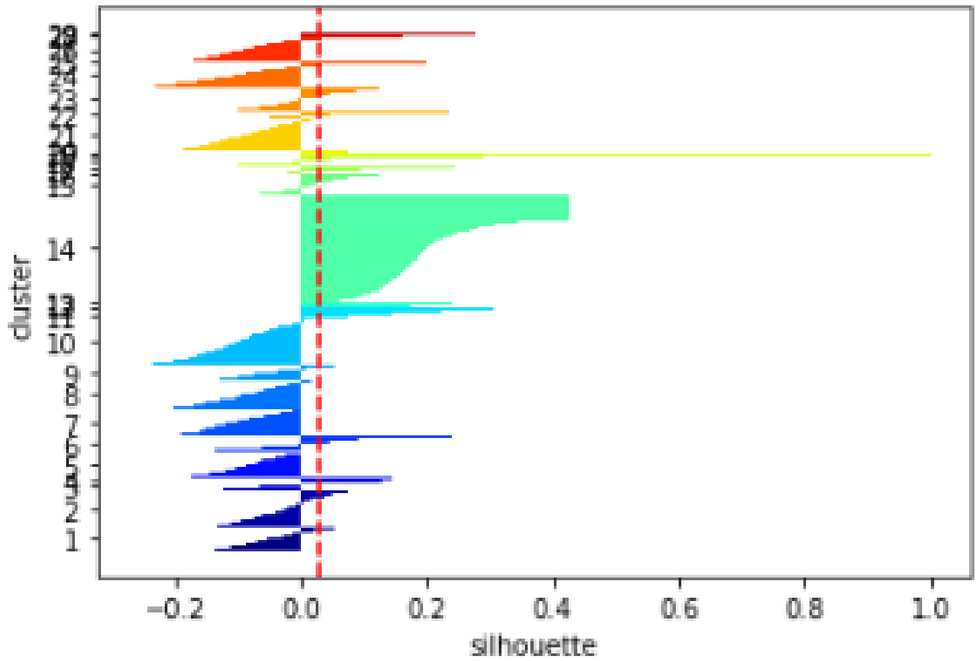


04-4. 군집화 : DBSCAN & K-means를 이용한 클러스터링

3. Kmeans를 이용한 클러스터링



Elbow Method의 경우 1~100개의 클러스터의 값을
그렸는데 급감하는 k 값을 찾기 애매

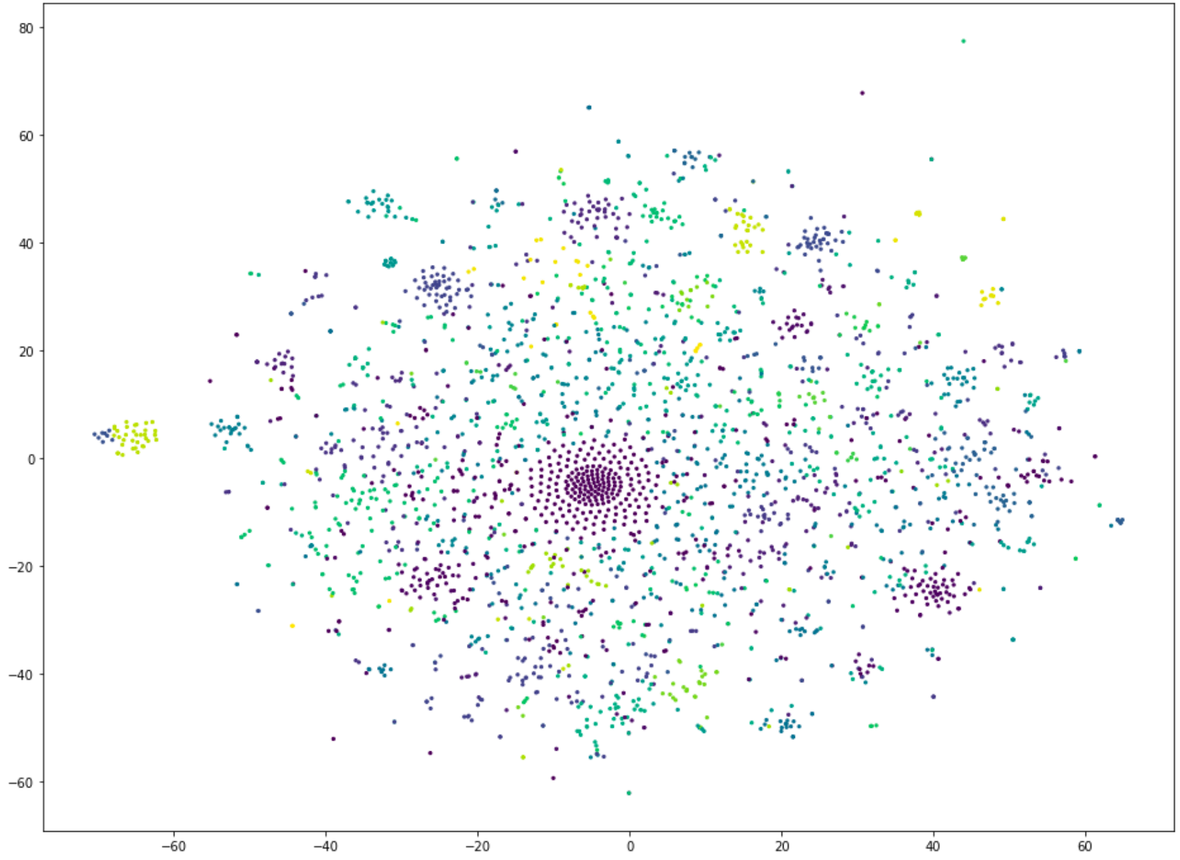


3개~ 30개의 클러스터를 비교했는데
결과가 좋지못함.

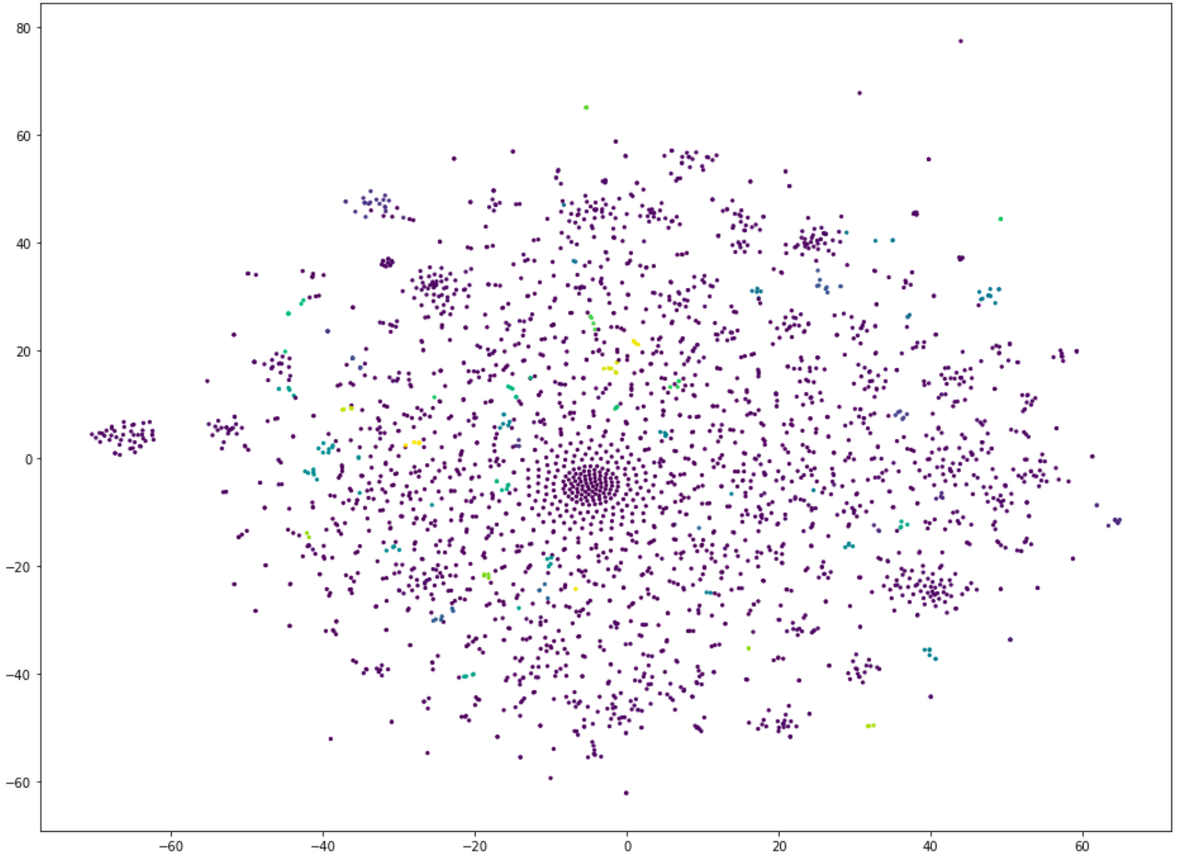
04-4. 군집화 : DBSCAN & K-means를 이용한 클러스터링

4. 클러스터링 시각화

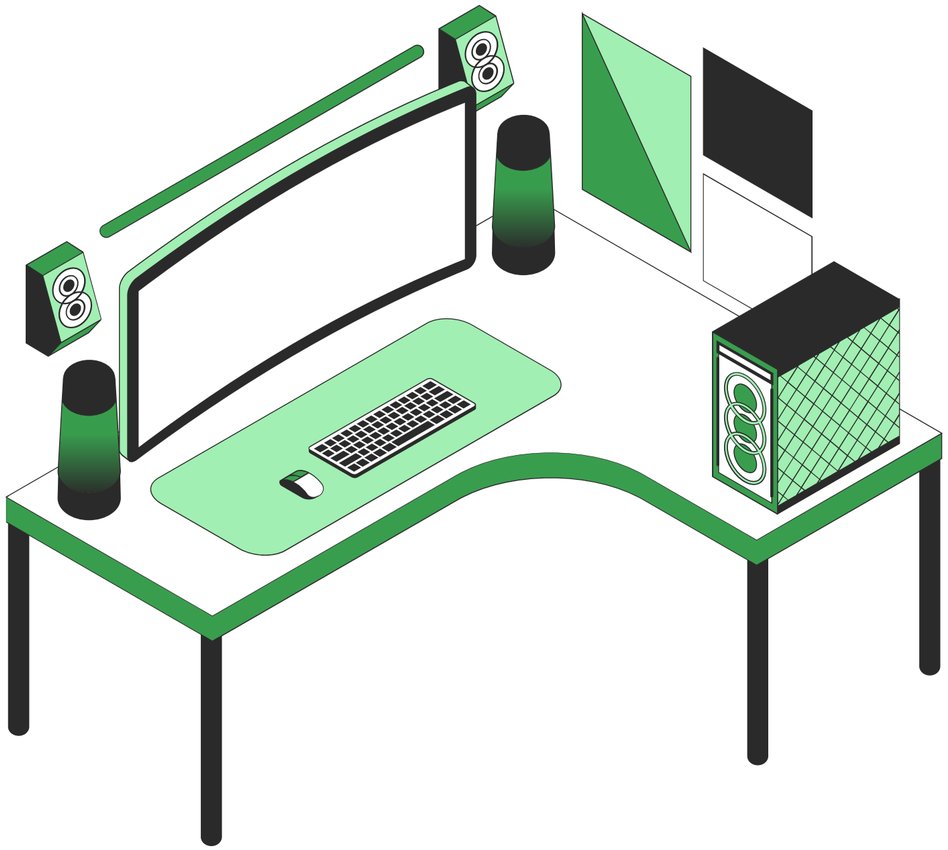
300 차원의 벡터를 눈으로 보기 쉽게 2차원으로 축소



k-means: k=100



DBSCAN: eps= 3.8



End

15기 이병주 16기 이은찬 최규빈 이영노