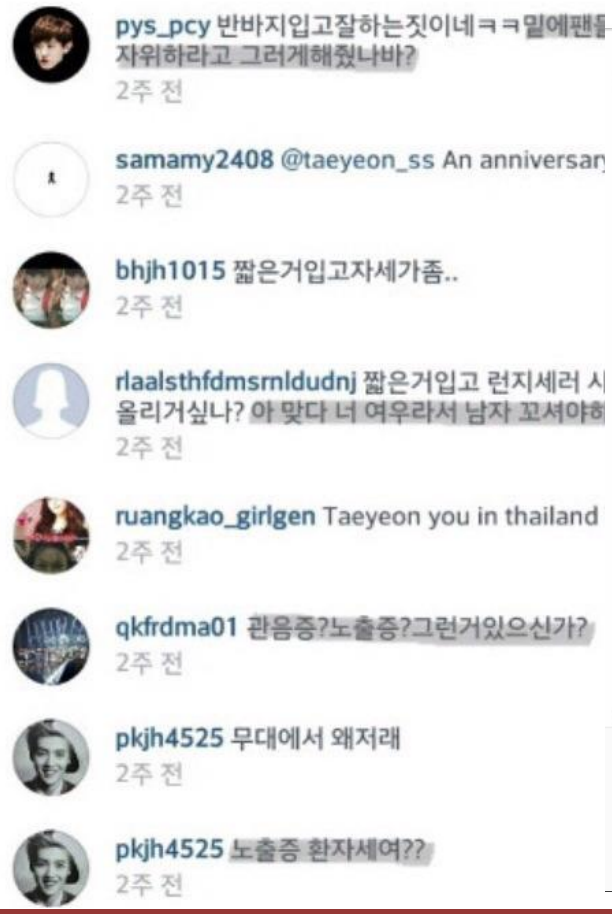
DSC DATA SCIENCE COMPETITION

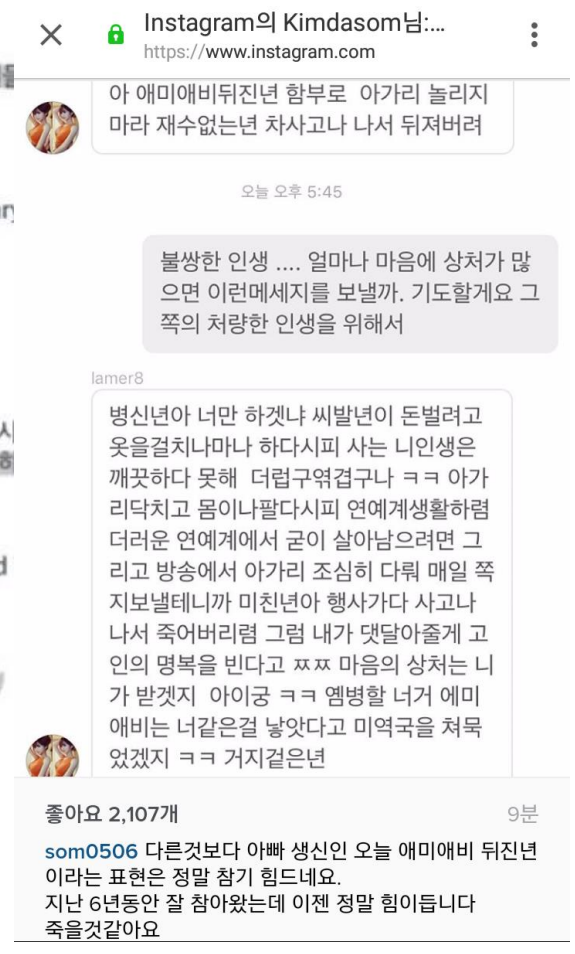
악플 타파

김서현 김지현 나혜지 박유진

“ 악플 타파를 하게 된 이유 ”


da****
장재인이 모두 저지른일임. 자기
고 직접 씀
전 | 신고
카톡내용만 정확히 따지면 양다리는 아니지
정확히 선 그었으니까 그리고 양다리 루머에
하고싶어도 폭탄같은 폭로녀분이 실시간으로
루머캡쳐 인스타스토리에 올리면서 협박했기
라고 사실이 아닌 루머로 상대방 명예훼손하
하 자격정지던데 남태현이 맘먹고 고소했으
인 자격정지도 가능했음
2일 전 | 신고
da****
개인 작업실 방송에 증거 다 있어요. 얼마나 들이대
했는지 그 방송 보면 다있는데 사람들이 속을거라
하지마세요. 그리고 카톡 내용에도 양다리가 아니
다 있고. 게다가 자신이 오해라고 직접 인스타에글
지 보고 기억하는 남태현 팬들 더 늘어날거고.. ㅈ
h****
스타 문자만으로 연애당사자
노출한 장재인을 극히 싫어
전 | 신고
니친구
성고다
성괴ㅋㅋ성형괴물녀
필러 코 눈 보조개

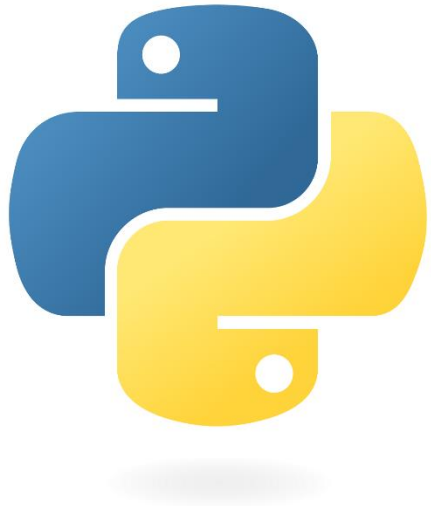

pys_pcy 반바지입고잘하는짓이네ㅋㅋ밑에팬들
자위하라고 그러게해줬나봐?
2주 전
samamy2408 @taeyeon_ss An anniversary
2주 전
bhjh1015 짧은거입고자세가좀..
2주 전
rlaalsthfdmsrnldudnj 짧은거입고 런지세러 시
올리거싶나? 아 맞다 너 여우라서 남자 꼬셔야하
2주 전
ruangkao_girlgen Taeyeon you in thailand
2주 전
qkfrdma01 관음증?노출증?그런거있으신가?
2주 전
pkjh4525 무대에서 왜저래
2주 전
pkjh4525 노출증 환자세여??
2주 전

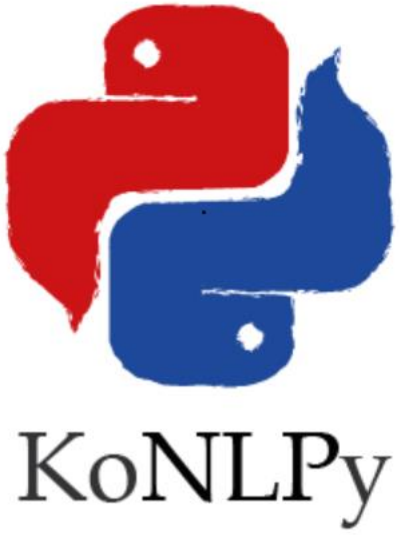

Instagram의 Kimdasom님:...
https://www.instagram.com
아 애미애비뒤진년 함부로 아가리 놀리지
마라 재수없는년 차사고나 나서 뒤져버려
오늘 오후 5:45
불쌍한 인생 .... 얼마나 마음에 상처가 많
으면 이런메세지를 보낼까. 기도할게요 그
쪽의 처량한 인생을 위해서
lamer8
병신년아 너만 하겟냐 씨발년이 돈벌려고
옷을걸치나마나 하다시피 사는 니인생은
깨끗하다 못해 더럽구엮겹구나 ㅋㅋ 아가
리닥치고 몸이나팔다시피 연예계생활하렴
더러운 연예계에서 굳이 살아남으려면 그
리고 방송에서 아가리 조심히 다뤄 매일 쪽
지보낼테니까 미친년아 행사가다 사고나
나서 죽어버리렴 그럼 내가 댓달아줄게 고
인의 명복을 빈다고 ㅉㅉ 마음의 상처는 니
가 받겟지 아이궁 ㅋㅋ 옘병할 너거 에미
애비는 너같은걸 낳앗다고 미역국을 쳐묵
었겠지 ㅋㅋ 거지같은년
좋아요 2,107개
9분
som0506 다른것보다 아빠 생신인 오늘 애미애비 뒤진년
이라는 표현은 정말 참기 힘드네요.
지난 6년동안 잘 참아왔는데 이젠 정말 힘이듭니다
죽을것같아요

“ 악플 타파를 하게 된 이유 ”


늘어난 사이버 명예훼손
1만 5,504건
1만 3,502
1만 5,840
일반 명예훼손
사이버 명예훼손
7,447
1만1,988
1만 4,661건
2014
2016
2018
자료 : 대검찰청

Data set





Data set

20대 미혼모 '3살 딸 학대치사' 119 신고자도 공범(종합)	강다니엘
철도노조 20일부터 무기한 총파업..4조 2교대제·SR 통합 요구	방탄소년단
수능 국어 25번, "1타 강사 믿고 풀었는데"	지효
박근혜 석달째 장기입원..법무부 "재수감 계획 전혀 검토 안해"	백현
"먹던 우물에 침뱉어"..한국당 영남다선들, 김세연에 반발 기류	태연
[단독] "포스터 곳곳 반칙 행적"..나경원 아들 '4저자'의 비밀	
제 발등 찍은 日.. 대일 무역적자 16년 만에 최저	
與86그룹, '기득권' 비판에 "모욕적..이해할 수 없어"	
"결혼 반대해서.." 남자친구와 아버지 살해한 딸, 징역 15년	
유니클로 '공짜 히트텍' 1시간만에 '순삭'	
고유정 "검사님 무서워 답변 못 하겠다"..재판 휴정	
폐암 여성의 90%, 담배 피운 적 없다는데.. 왜?	
고유정 "경찰이 시신 찾을 수 있을 줄 알았다"..검찰구형은 2주 연기	
"마킹 실수로 수정" 말했다가..수능 전 과목 '0점 처리'	
"이젠 유튜버 안받습니다"..'노튜버존' 선언하는 식당들	

Data set

In [2]:
```
def get_daum_news_comments(news_id):
    headers = {
    'Authorization' : 'Bearer eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJmb3J1bV9rZXkiOiJuZXdzIiwiZ3JhbnRfdHlwZSI6ImFsZXhfY3JIZGVudGlhbHMiLCJzY2
    'Origin' : 'https://news.v.daum.net',
    'Referer' : 'https://news.v.daum.net/v/20191118175606836',
    'User-Agent' : 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/78.0.3904.97 Safari/537.36'
    }
    url_template = 'https://comment.daum.net/apis/v1/posts/@{}/comments?parentId=0&offset={}&limit=10&sort=RECOMMEND&isInitial=false'
    offset = 0
    comments = []
    while True:
        url = url_template.format(news_id, offset)
        resp = requests.get(url, headers=headers)
        data = resp.json()
        if not data:
            break
        comments.extend(data) # data 역시 타입이 list 이기 때문에 extend 로
        offset += 10
    return comments
```

In [3]:
```
data= pd.DataFrame(get_daum_news_comments('20191118175606836'))
df=data[['id','content']]
data = (df['content']).map(lambda x: x.replace('₩u11a2','').replace('₩u11a2',''))
df.columns = ['id','content']
df.to_csv('크롤링결과1.csv',encoding='UTF-8')
```

Data set

In [2]:
```
def get_daum_news_comments(news_id):
    headers = {
    'Authorization' : 'Bearer eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJmb3J1bV9rZXkiOiJuZXdzIiwiZ3JhbnRfdHlwZSI6ImFsZXhfY3JIZGVudGlhbHMiLCJzY2
    'Origin' : 'https://news.v.daum.net',
    'Referer' : 'https://news.v.daum.net/v/20191118175606836',
    'User-Agent' : 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/78.0.3904.97 Safari/537.36'
    }
    url_template = 'https://comment.daum.net/apis/v1/posts/@{}/comments?parentId=0&offset={}&limit=10&sort=RECOMMEND&isInitial=false'
    offset = 0
    comments = []
    while True:
        url = url_template.format(news_id, offset)
        resp = requests.get(url, headers=headers)
        data = resp.json()
        if not data:
            break
        comments.extend(data) # data 역시 타입이 list 이기 때문에 extend 로
        offset += 10
    return comments
```

In [3]:
```
data= pd.DataFrame(get_daum_news_comments('20191118175606836'))
df=data[['id','content']]
data = (df['content']).map(lambda x: x.replace('₩u11a2','').replace('₩u11a2',''))
df.columns = ['id','content']
df.to_csv('크롤링결과1.csv',encoding='UTF-8')
```

Data set

In [1]:
```
import GetOldTweets3 as got
from bs4 import BeautifulSoup
```

In [2]:
```
import datetime

days_range = []

start = datetime.datetime.strptime("2014-06-19", "%Y-%m-%d")
end = datetime.datetime.strptime("2014-08-16", "%Y-%m-%d")
date_generated = [start + datetime.timedelta(days=x) for x in range(0, (end-start).days)]

for date in date_generated:
    days_range.append(date.strftime("%Y-%m-%d"))

print("=== 설정된 트윗 수집 기간은 {} 에서 {} 까지 입니다 ===".format(days_range[0], days_range[-1]))
print("=== 총 {}일 간의 데이터 수집 ===".format(len(days_range)))
```

```
=== 설정된 트윗 수집 기간은 2014-06-19 에서 2014-08-15 까지 입니다 ===
=== 총 58일 간의 데이터 수집 ===
```

Data set

```
# 특정 검색어가 포함된 트윗 검색하기 (quary search)
# 검색어 : 어벤져스, 스포

import time

# 수집 기간 맞추기
start_date = days_range[0]
end_date = (datetime.datetime.strptime(days_range[-1], "%Y-%m-%d")
            + datetime.timedelta(days=1)).strftime("%Y-%m-%d") # setUntil이 끝을 포함하지 않으므로, day + 1

# 트윗 수집 기준 정의
tweetCriteria = got.manager.TweetCriteria().setQuerySearch('태연')\
                                           .setSince(start_date)\
                                           .setUntil(end_date)\
                                           .setMaxTweets(-1)

# 수집 with GetOldTweet3
print("Collecting data start.. from {} to {}".format(days_range[0], days_range[-1]))
start_time = time.time()

tweet = got.manager.TweetManager.getTweets(tweetCriteria)

print("Collecting data end.. {0:0.2f} Minutes".format((time.time() - start_time)/60))
print("=== Total num of tweets is {} ===".format(len(tweet)))
```

```
Collecting data start.. from 2014-06-19 to 2014-08-15
```

Data set

```
# 원하는 변수 골라서 저장하기

from random import uniform
from tqdm import tqdm_notebook

# initialize
tweet_list = []

for index in tqdm_notebook(tweet):

    # 메타데이터 뽑기
    username = index.username
    link = index.permalink
    content = index.text
    tweet_date = index.date.strftime("%Y-%m-%d")
    tweet_time = index.date.strftime("%H:%M:%S")
    retweets = index.retweets
    favorites = index.favorites

    # 결과 합치기
    info_list = [tweet_date, tweet_time, username, content, link, retweets, favorites]
    tweet_list.append(info_list)

    # 휴식
    time.sleep(uniform(1,2))
```

```
# 파일 저장하기

import pandas as pd

twitter_df = pd.DataFrame(tweet_list,
                          columns = ["date", "time", "user_name", "text", "link", "retweet_counts", "favorite_counts"])

# csv 파일 만들기
twitter_df.to_csv("sample_twitter_data_{}_to_{}_태연.csv".format(days_range[0], days_range[-1]), index=False ,encoding='UTF-8')
print("=== {} tweets are successfully saved ===".format(len(tweet_list)))
```

Data set

```
import re
INPUT_FILE_NAME = '악플타파_합친거 최종본.txt'
OUTPUT_FILE_NAME = '악플타파_합친거 최종본_clean.txt'
def clean_text(text):
    cleaned_text = re.sub('[a-zA-Z]' , '', text)
    cleaned_text = re.sub('[ㅱ{ㅱ}ㅱ[ㅱ]ㅱ/?.,;:|ㅱ)*~`!^ㅱ-_+<>♡@ㅱ#$%&ㅱㅱ=ㅱ(ㅱ'ㅱ" · ♪ ♬ ♩ ★● ]',
                          '', cleaned_text)
    return cleaned_text
read_file = open(INPUT_FILE_NAME, 'r')
write_file = open(OUTPUT_FILE_NAME, 'w')
text = read_file.read()
text = clean_text(text)
write_file.write(text)
read_file.close()
write_file.close()
```

단어 빈도수 조사

```
f = open("BBD_.txt", "r")

lines = f.read()
```

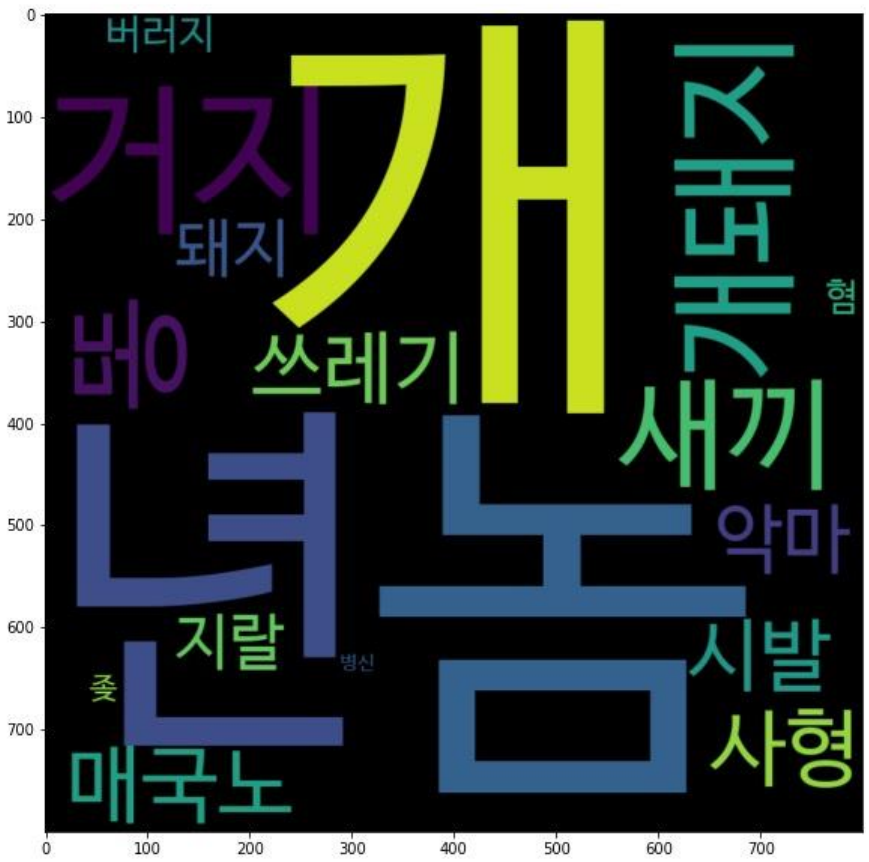
```
from konlpy.tag import Twitter

nlpy = Twitter()

nouns = nlpy.nouns(lines)

print(nouns)
```

```
['내', '짤', '백혐', '평화', '상징', '존', '한텐', '마음', '존', '조', '말', '그냥', '백혐', '뿐', '사랑', '인걸', '혼자', '대애',
'박', '그', '오크', '케이', '줬음', '백혐', '욕', '뭐', '난', '존나', '백혐', '말', '애', '서치', '나리', '존낰', '그게', '뭐', '또',
'같알', '존나', '정상', '멘탈', '이변', '백혐', '직접', '검색', '생각', '멋쟁이', '백혐', '백혐', '백혐', '백혐', '캡쳐', '엑퀴', '여
자', '사람', '여자', '사람', '관심', '인티', '트윗', '백혐', '덕', '구들', '캡쳐', '그', '중', '플텍', '스파이', '왜', '거짓말', '해',
'네', '먼저', '백혐', '시작', '자', '본인', '서치', '이상', '면전', '백혐', '멘탈', '덱', '심지어', '트위터', '지네', '구', '배',
'덱', '백혐', '것', '서치', '뭐', '니', '꼬우', '텅', '구라', '거', '서치', '왜', '니', '데', '현이', '보고', '백혐', '지랄', '태연',
```

단어 빈도수 조사

당가	1013
니엘	1002
백혐	623
강	413
빨탄	387
글자	204
존나	184
실트	183
왜	143
진짜	134
팬	131
방	131
뭐	107
거	107
시발	104
개	101
내	100
세례명	100
다니엘	91
나	82
우리	81
사람	71
니	68
애	63
생각	62
것	59
때	59
말	58
혐	55
더	54
웃기	54
지금	53
너	52
그	51
또	50
엑소	45
이	44
씨발	44
좀	43
소년단	43
이제	42
보고	41
지랄	41
새끼	41
생	41
상	41
안	40
임	40
베이컨	38
좇	37
친구	37
그냥	36
하나	35
팬덤	35
욕	33
서치	33
네	33
데	33
글	33
저	33
단어	33
효	33
줄	32
연애	31
이름	31
오늘	31
알	31
서방	30
난	29
오늘	31
알	31
서방	30
난	29
분리수거	29
얼굴	28
눈	28
공개	28
집	27
누가	26
가수	25
년	25
때문	25
수	25
대상	25
공카	25
백현	24
제발	24
누구	24
사진	24
표절	24
당나귀	24
의미	24
해	23
구	23
오빠	23
남	23
그게	22
태연	22
정말	22
함	22
잼	22
돈	22
도	21
건	21

시각화: WordCloud

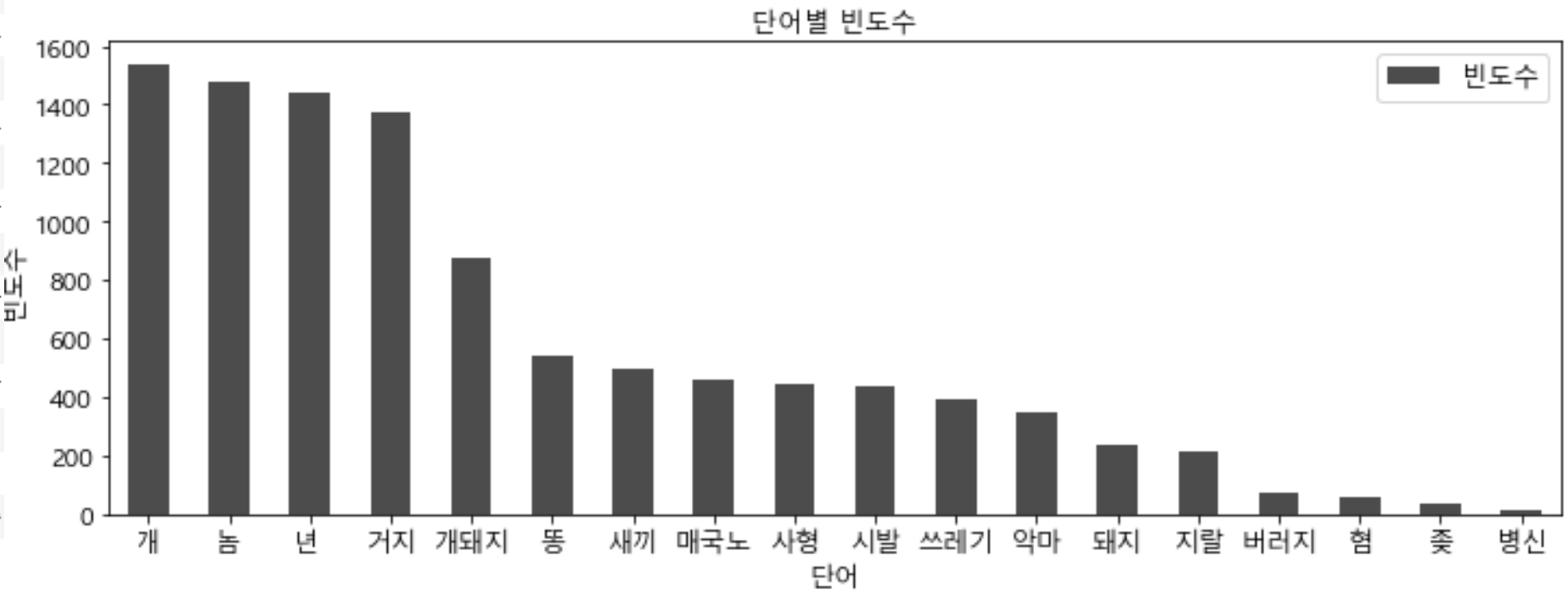


시각화: 바 그래프

	단어	빈도수
0	개	1539
1	놈	1480
2	년	1441
3	거지	1377
4	개돼지	876
5	똥	539
6	새끼	497
7	매국노	456
8	사형	447
9	시발	434
10	쓰레기	390
11	악마	345
12	돼지	239
13	지랄	213
14	버러지	70
15	혐	55
16	좆	37
17	병신	11

시각화: 바 그래프

	단어	빈도수
0	개	1539
1	놈	1480
2	년	1441
3	거지	1377
4	개돼지	876
5	똥	
6	새끼	
7	매국노	
8	사형	
9	시발	
10	쓰레기	
11	악마	
12	돼지	
13	지랄	
14	버러지	
15	혐	
16	좆	
17	병신	



해결책 제안 1

1.네이버, 다음, 유튜브, 트위터 등 정제된 악플들은 악플이라고 판단이 되면 바로 pdf로 id명으로 저장이 된다.

2.저장을 하는 공간은 각 포털에서 관리를 해주며 이는 2년 정도에 한 번씩 저장고를 청소하기 위해 초기화한다

해결책 제안 2

1. 인스타라이브, 브이앱, 유튜브 라이브 등 실시간으로 내용이 올라가는 악플의 경우 동영상을 저장하지 않는 경우에는 악플의 증거가 남지 않고 저장한 경우에도 일일이 시간 별로 찾아 캡쳐하여 증거를 남기고 있다.

2. 이를 방지하기 위해 실시간 라이브가 돌아갈 때 타이핑을 하여 글이 올라오면 악플이라 판단될 경우 바로 pdf파일로 넘기는 프로그램을 패치한다.

THANK YOU

발 표 자 박 유 진