2019년 하반기 인턴 개인과제

# 통신사별 고객 평판 분석

Web Crawling 활용 자연어 처리 및 Insight 도출

리텐션혁신팀 인턴 정원희



# 목차

01. 주제 선정 배경

## 02. 분석내용

- 방법론
- 자연어 데이터 수집 & 전처리
- 감성 분류
- 결과분석

03. Insight 도출

04. Appendix



## 인터넷가입 현명하게 하는법 - 자유게시판 - 체리피커

cpicker.com/136112 ▼

IT쪽 생활을 하다보면, 안타까운 풍경들을 자주 봅니다. 가장 대표적인 것이 **"인터넷가입"** 인데요.. 인터넷의 경우에는, 핸드폰에 비해서 현명하게 갈아타지 못하는 ...

## SK LG KT 인터넷가입 비교사이트, 요금절약과 사은품을 동시에

www.econovill.com/news/articleView.html?idxno=344205 ▼

2018. 8. 27. - [이코노믹리뷰=전진혁 기자] 초고속기가 **인터넷가입** 비교사이트를 통해 KT LG SK **인터넷가입**을 신청할 경우 사은품을 지원 받을 수 있으며 최저 ...





인터넷의 경우에는, 핸드폰에 비해서 현명하게 갈아타지 못하는 경우들을 자주 접합니다.

IT쪽 생활을 하다보면, 안타까운 풍경들을 자주 봅니다.

(아마 핸드폰에 비해 갈아타는 주기가 다소 길어 그런거 같습니다)

가장 대표적인 것이 "인터넷가입" 인데요..



- ① 나는 언제 약정이 끝나는가?
- ② 지금 해지하면 위약금이 얼마인가?

만약 약정기간(대부분 3년)이 만료된 상태라면,

이 질문 자체가 의미가 없겠죠?

만약 약정 만료가 아닌 경우에는 꼭! 확인해보셔야 합니다.



(어떤 경우의 수도 필요없이, 그냥 인터넷 통신사를 갈아타는게 100% 이득임)

#### 01. 주제 선정 배경

## 리텐션혁신팀

만약, 약정이 아직 끝나지 않은 상태인 경우에는

인터넷가입을 하는 게 아니라 **"해지신공"**을 펼치는 것이 훨씬 낫습니다.

여기서 해지신공이란,

통신사에 전화해서 "해지하려고 고민중인데... 내가 해지안하면, 나에게 어떤 혜택을 줄 수 있냐?" 라고

간접적으로 물어보는 행위를 의미합니다.

이게 왜 의미가 있냐하면, 통신사 고객센터는 어떻게 해서든 해지를 막아야 하는 매뉴얼이 있기 때문에,

해지방어를 위해 ① 요금 **할인, ② 상품권 지급** 등의 행동을 하기 때문입니다.

즉, 내가 가입한 통신사에 전화 한통만 해도..

요금이 할인되고 상품권이 생기는 마법을 경험하게 되는 것이지요.



만약, 약정이 아직 끝나지 않은 상태인 경우에는

인터넷가입을 하는 게 아니라 "해지신공"을 펼치는 것이 훨씬 낫습니다

여기서 해지신공이란

통신사에 전화해서 "해지하려고 고민중인데... 내가 해지안하면, 나에게 어떤 혜택을 줄 수 있냐?" 라고

## 정보에 민감한 체리피커인 커뮤니티 이용자들은

ᅃᄳ<mark>통신3사를 어떻게 인식</mark>하고 있을까? 때문에

해지방어를 위해 ① 요금 할인. ② 상품권 지급 등의 행동을 하기 때문입니다.

즉. 내가 가입한 통신사에 전화 한통만 해도.

요금이 할인되고 상품권이 생기는 마법을 경험하게 되는 것이지요

## Web Crawling을 통해 자연어 데이터를 수집 후 전처리 및 결과 분석함

#### 자연어 데이터 수집

Web Crawling을 활용해 통신사 관련 게시글 수집



#### 자연어 전처리 작업

자연어 처리 패키지에 적합한 데이터로 전처리



#### ▍ 감성 분류 및 결과분석 실시

긍정 / 부정 / 중립을 판별하고 도출된 결과 분석





## 타겟 커뮤니티 사이트를 선정하고 매크로 방식으로 Crawling 진행함

#### 타겟 커뮤니티 정보

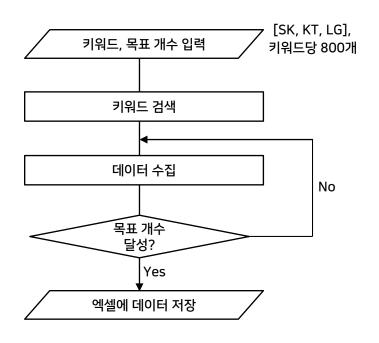
국내 최대 규모 커뮤니티 '디시인사이드'로 선정



고른 연령대 분포와 높은 이용량으로 분석에 용이함

#### 자연어 수집 순서도

키워드 검색 후 HTML 태그를 통해 데이터 수집하는 매크로 방식



## 통신사별 자연어 데이터(제목, 본문, 댓글, 날짜, 조회수)를 분류함

## 항목별 자연어 데이터 분류

수집한 데이터를 아래 양식에 맞춰서 전처리하기 위한 선행작업 실시

고객 자연어 데이터 <sup>1)</sup>				Index용 데이터	
제목	본문	댓글	날짜	조회수	
유선인터넷은 무조건 KT써야하는 이유.txt	SK브로드밴드랑 유플러스는 KT해외망 빌려써서 느릴수밖에 없음	1. 인정	2019.07.31	167	
SK 브로드밴드 어떠함?	요즘 쓸만함???	<ol> <li>그냥 뭐 좋지도 않고 그렇다고 막 엄청 나쁘지 않고 가족묶음 할인 아니면 kt 가야지</li> <li>옛날 두루넷선 쓰는데 좋겠냐?</li> <li>결합묶여있는 거 아니면 kt가라</li> </ol>	2019.07.26	102	
티브로드랑 sk브로드밴드랑 합병했는데	이렇게 합병되면 뭐 투자포인트 같은거 있냐? ● ●	1. 시너지 명목으로 요금올림 ㅋㅋㅋㅋㅋ 2. 다른 회사였음?	2019.04.27	128	

(+)

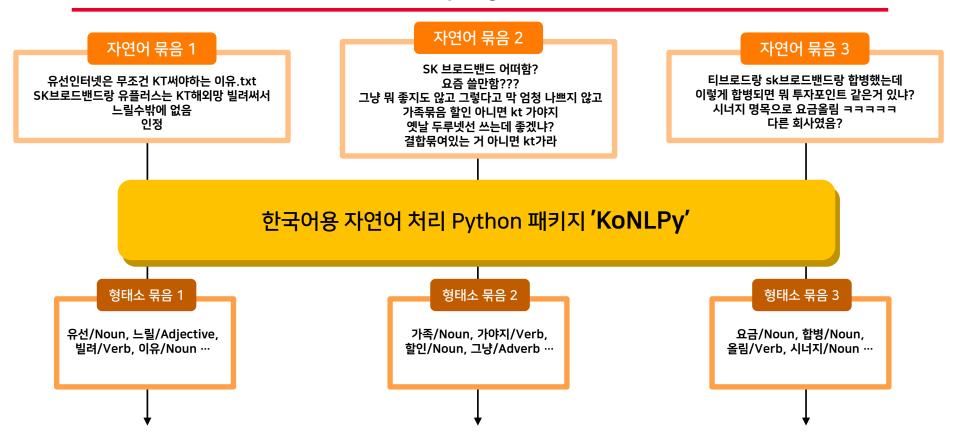
SK

KT

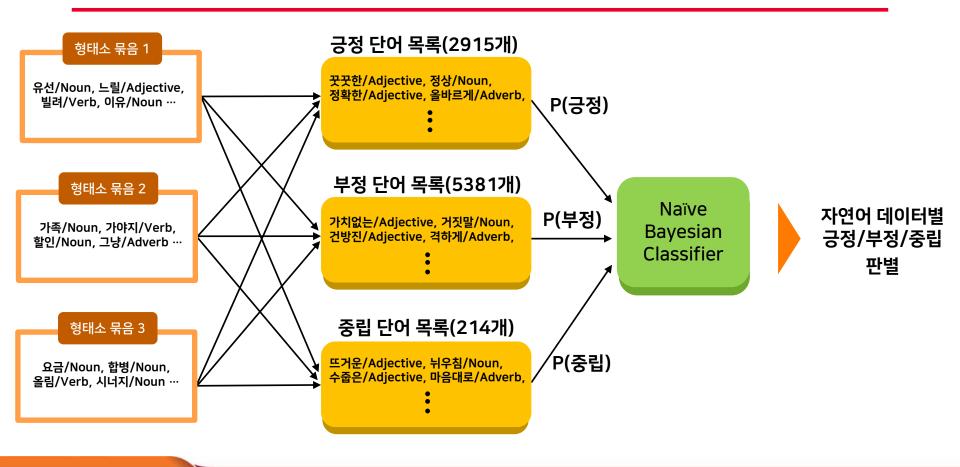
LG

통합

## 자연어 데이터들을 각각 Grouping하고 형태소 단위로 세부분류함

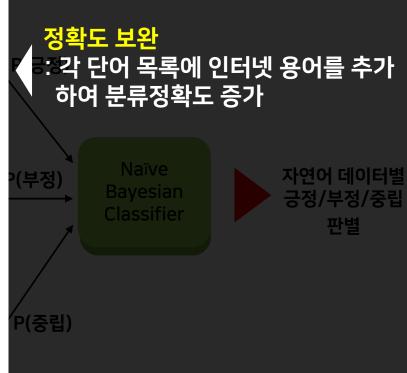


## 형태소 묶음을 감성 사전1)을 활용해 긍정/부정/중립으로 구분함



## 형태소 묶음을 감성 사전<sup>1)</sup>을 활용해 긍정/부정/중립으로 구분함





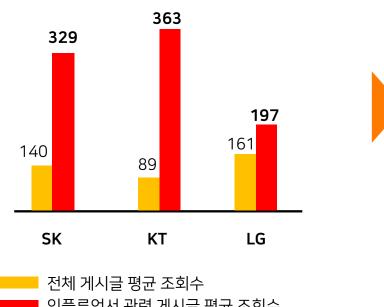
## 실제 자연어 데이터를 확인해보면 KT에 대해 긍정적임

통신사	제목	본문	댓글	분류결과
SK	SK브로드밴드는 왜 회선품질 개선이 안되냐	빠가야로	태생이 하나로텔레콤이여 전국에 전화국 가지고 있는 KT를 어케 이기냐 광케이블도 그렇고	긍정
SK	SK브로드밴드 디자인 이상해짐	개별로	- 오분류·	긍정
SK	SK브로드밴드 공유기 진짜 개꾸짐	20	-	긍정
KT	유선인터넷은 KT써라	제발	1. Sk브로드밴드가 제일 폐급이야 2. 헐 어제 sk브로드밴드에서 인터넷 요금 kt보다 싸게해준다고 넘어오라고 하더니만… 고민하다 안갔는데…	부정
SK	인터넷은 kt가 답이다	Sk랑 sk브로드밴드 인터넷도 역겨움 kt로 바꾸고 ㄹㅇ 신세계임	<del>-</del>	부정
LG	SK로 인터넷이랑 TV 옮겼다	원래 쓰던건 LG U+ 100메가 광랜 + IPTV. SK 100메가 광랜 + IPTV 현 금 65만 + 상품권 7만원 받고 옮김	1. Sk인터넷은 쓰면 안됨 2. 개마니주네 ㄷㄷ 쩐다	부정

## 그러나, 실제 자연어 데이터를 확인해보면 오히려 KT에 대해 긍정적임

SK	SK브로드밴드는 왜 회선품질 개선이 안되냐	빠가야로 실제 고객 평판	태생이 하나로텔레콤이어 KT를 어케 이기냐	i 전국에 전화국 가지고 있는 광케이블도 그렇고	긍정
SK	SK브로드밴드 디자인 이상해짐			- 오분류· <sup>·.·</sup>	긍정
SK	SK브로드밴드 공유기 진짜 개꾸짐	긍정 : KT > SK > L0	i	-	긍정
KT	유선인터넷은 KT써라	부정: KT < SK < L0 '유선은 KT, 무선은 SK가 좋다 반응이 지배적	- k브i H주디	ㅏ제일 폐급이야 로드밴드에서 인터넷 요금 ·고 넘어오라고 하더니만··· ···	부정
SK	인터넷은 kt가 답이다	다 이 기 기 기 기 기 기 기 기 기 기 기 기 기 기 기 기 기 기		-	부정
LG	SK로 인터넷이랑 TV 옮겼다	원래 쓰던건 LG U+ 100메가 광랜 + IPTV. SK 100메가 광랜 + IPTV 현 금 65만 + 상품권 7만원 받고 옮김	1. Sk인터넷은 쓰던 2. 개마니주네 ㄷㄷ		부정

▋통신사별 게시글 평균 조회수¹)



인플루언서 관련 내용이 포함된 게시글의 조회수와 고객평판이 매우 긍정적

인플루언서 관련 게시글 평균 조회수

SK

#### SK브로드밴드 광고 웃음지뢰

수퍼비랑 면도ㅋㅋㅋㅋㅋ





래퍼 <수퍼비>

래퍼 <myunDo>

수퍼비 X 면도 (feat.이날) 'Life Is Premium' 뮤직비디오 출현

SK브로드밴드 신입사원 이서진, 헨리.gif





배우 <이서진>

가수 <헨리>

- 둘다 좋아ㅋ 광고 찍고싶으면 예능 고고
- 헨리 꽃할배 나와도 되겠다 둘 케미 좋다

SK브로드밴드 광고모델 이서진, 헨리

SK는 연예인을 활용하여 광고



## **KT**

#### 릴카 슈기 kt이벤트에 참여





게임 유튜버 <릴카>

먹방 유튜버 <슈기>

- 대기업 이벤트에 참여하다니 대단하구만

KT 3D 영상통화앱 '나를' 통화이벤트 참여





뜨뜨뜨뜨 X KT 10GiGA 인터넷 광고(유튜브 영상 캡처)





## LG

#### U+ 5G 청하 AR 진짜 신기하다 ㅋㅋ



가수 <청하>

- 춤추는거 찍어서 영상 올리는 이벤트인듯 신기하당

U+ X 청하 AR 활용 '나만의 입체스타' 광고

#### 대도서관 이미지 좋긴 좋네



- 대외활동을 활발하게 함. 기자들도 팬이라더만

U+ X 대도서관 방구석 아티스트(오디션) 이벤트





## 인플루언서의 파급효과를 사용해 리텐션을 높일 방법을 찾아볼 수 있음

44 리텐션에 <mark>인플루언서</mark>를 활용하면 효과적일 것이다

예시 : 해지방어시 상품권뿐만 아니라 연예인 공연 티켓 제공 해지취약Seg들을 추첨하여 인플루언서 팬미팅 개최



# 감사합니다



#### Web Crawling 소스코드 (1/5)

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from openpyxl import load workbook #엑셀 읽기용
from openpyxl import Workbook #엑실 쓰기용
import time
load excel = load workbook("output.xlsx", data only=True)
load sheet = load excel['Sheet2']
path = "C:/chromedriver.exe"
driver = webdriver.Chrome(path)
driver.get(load sheet.cell(row=1, column=3).value)
1 1 1
query = driver.find element by name ("search")
query.send keys("sk므로드밴드")
query.send keys(Keys.RETURN)
morePosts = driver.find elements by tag name('a')
for i in morePosts:
   if i.text == '게시물 더보기':
        i.click()
       break
1 1 1
```



## Web Crawling 소스코드 (2/5)

```
column=2).value
isLastPage = False
db = [List('','','','','',0)]
blacklist = ['아우슈리네', '행복전도사', '초등교육학', '핑핑이', '보트피플', '게임방송', '잡담']
checkBlacklist = False
driver.refresh()
while curPageNum <= 33:
   driver.refresh()
    postLists =
driver.find_elements_by_xpath("//ul[@class='sch_result list']/li/a[@class='tit txt']")
    for j in postLists:
       driver.implicitly wait(1)
        for item in blacklist:
           if (j.text).find(item) >= 0:
               checkBlacklist = True
               continue
        if checkBlacklist:
           checkBlacklist = False
           continue
        j.click()
        driver.switch to.window(driver.window handles[-1])
```

idx, curPageNum = load sheet.cell(row=1, column=1).value, load sheet.cell(row=1,

#### Web Crawling 소스코드 (3/5)

if ( (driver.current url).find('no=') == -1):

```
driver.close()
            driver.switch to.window(driver.window handles[0])
            driver.implicitly wait(3)
            continue
        idx = idx + 1
        print(idx,'번째 글입니다')
        driver.refresh()
        title = driver.find element by xpath("//h3[@class='title ub-
word']/span[@class='title subject']")
        print(title.text)
        content = driver.find elements by xpath("//div[@style='overflow:hidden;']/p |
//div[@style='overflow:hidden;']/div")
        content tmp = ''
        for k in content:
            if(k.text != ''):
                #print(k.text)
                content tmp = content tmp + k.text + ' '
 comment = driver.find elements by xpath("//p[@class='usertxt ub-word']")
        comment tmp = ''
        for k in comment:
            if(k.text != ''):
                #print(k.text)
                comment tmp = comment tmp + k.text + ' '
```

#### Web Crawling 소스코드 (4/5)

```
date = driver.find element by xpath("//div[@class='fl']/span[@class='gall date']")
        view = driver.find element by xpath("//div[@class='fr']/span[@class='gall count']")
        load sheet.cell(row=idx + 2, column=2, value=idx)
        load sheet.cell(row=idx + 2, column=3, value=title.text)
        load sheet.cell(row=idx + 2, column=4, value=content tmp)
        load sheet.cell(row=idx + 2, column=7, value=comment tmp)
        load sheet.cell(row=idx + 2, column=5, value=date.text[:10])
        load sheet.cell(row=idx + 2, column=6, value=view.text[3:])
        load sheet.cell(row=1, column=1, value=idx)
        load excel.save(filename='output.xlsx')
        #db.append(List(driver.current url, title.text, content tmp, comment tmp,
date.text[:10], int(view.text[3:])))
        #print(idx, db[idx].url, db[idx].title, db[idx].date, db[idx].viewCount)
        #print(db[idx].contents)
        #print(db[idx].comments)
        #driver.implicitly wait(1)
        driver.close()
        driver.switch to.window(driver.window handles[0])
        #driver.implicitly wait(1)
```



#### Web Crawling 소스코드 (5/5)

```
curPageNum = curPageNum + 1
    load sheet.cell(row=1, column=2, value=curPageNum)
    if curPageNum % 10 == 1:
        isLastPage = True
    if isLastPage:
        driver.refresh()
       next =
driver.find elements by xpath("//div[@class='bottom paging box']/a[@class='b next']")
        for i in next:
            if i.text == '다음':
                i.click()
                isLastPage = False
                driver.implicitly wait(1)
                break
 else:
        pageLists = driver.find elements by xpath("//div[@class='bottom paging box']/a")
        for i in pageLists:
            if i.text == str(curPageNum):
                i.click()
                load sheet.cell(row=1, column=3, value=driver.current url)
                break
    #driver.implicitly wait(3)
    driver.refresh()
```

## 형태소 분석 소스코드(1/2)

```
from konlpy.tag import Twitter
from openpyxl import load workbook
List = ['SK', 'KT', 'LG', '통합']
load excel = load workbook(filename='output.xlsx', data only=True)
twit = Twitter()
for idx in range (3,4):
    nounList, nounCountList = set(), {}
    load sheet = load excel[List[idx]]
    if idx == 0:
        k = 744
    elif idx == 1:
        k = 769
    elif idx == 2:
        k = 764
    else:
        k = 2273
```



#### 형태소 분석 소스코드(2/2)

```
for i in range (3, k):
       print(i-2,'번째 문장입니다.')
        sentence = load sheet.cell(row=i, column=3).value + ' ' + str(load_sheet.cell(row=i,
column=4).value) + ' ' + str(load sheet.cell(row=i, column=7).value)
        tagging = twit.nouns(sentence)
       for j in tagging:
            if j not in nounList:
                #print(j, '는 없는 명사입니다.')
                nounList.add(j)
nounCountList[j] = 0
            nounCountList[j] = nounCountList[j] + 1
   result = sorted(nounCountList.items(), key=lambda i : i[1], reverse=True)
    read excel = load workbook(filename='result(Twitter).xlsx', data only=True)
   read sheet = read excel[List[idx]]
    for t, i in enumerate (result):
        read sheet.cell(row=t+1, column=1, value=i[0])
        read sheet.cell(row=t+1, column=2, value=i[1])
    read excel.save(filename='result(Twitter).xlsx')
```



## Naïve Bayesian Classifier 소스코드(1/4)

from konlpy.tag import Twitter

```
from openpyxl import load workbook
from openpyxl import Workbook
from konlpy.utils import pprint
from math import log
twit = Twitter()
def readDic(textfile, wordSet):
    f dic = open(textfile, 'r', encoding='UTF8')
    lines = f dic.readlines()
    for line in lines:
        pos items = twit.pos(line)
        for pos item in pos items:
            tmp = '/'.join(pos item)
            if tmp not in wordSet:
                wordSet.add(tmp)
    return wordSet
def makePos(sentence):
    wordSet = set()
    pos items = twit.pos(sentence)
    for pos item in pos items:
        tmp = '/'.join(pos item)
        if tmp not in wordSet:
            wordSet.add(tmp)
    return wordSet
```

#### Naïve Bayesian Classifier 소스코드(2/4)

```
def getProb(sentence,posList):
   prob, listCnt = 0, len(posList)
   items = makePos(sentence)
    for item in items:
        cnt = 1/countAll
        if item in posList:
           cnt += listCnt/countAll
        prob += log(cnt/(listCnt+countAll))
   return prob
def findMaxProb(pos, neg, mid):
   list = [pos, neg, mid]
   if max(list) == pos:
       print('긍정입니다')
       return '긍정'
   elif max(list) == neg:
       print('부정입니다')
        return '부정'
   else:
        print('중립입니다')
        return '중립'
```



#### Naïve Bayesian Classifier 소스코드(3/4)

```
def bayesian(sheetName, cnt):
    load sheet = load excel[sheetName]
    for i in range(3,cnt):
         target = load sheet.cell(row=i, column=3).value + ' ' \
                  + str(load sheet.cell(row=i, column=4).value) + ' ' \
                  + str(load sheet.cell(row=i, column=7).value)
    load excel.save()
emptySet = set()
f pos = readDic('pos pol word.txt',emptySet)
countPos = len(f pos)
emptySet = set()
f neg = readDic('neg pol word.txt',emptySet)
countNeg = len(f neg)
emptySet = set()
f mid = readDic('obj unknown pol word.txt',emptySet)
countMid = len(f mid)
countAll = 7649
load excel = load workbook(filename='output.xlsx', data only=True)
```



#### Naïve Bayesian Classifier 소스코드(4/4)

```
List = ['SK', 'KT', 'LG', '통합']
for idx in range (4):
    load sheet = load excel[List[idx]]
    if idx == 0:
        k = 745
    elif idx == 1:
        k = 770
    elif idx == 2:
        k = 765
    else:
        k = 2274
    for i in range (3, k):
        print(i-2,'번째 글입니다')
        sentence = load sheet.cell(row=i, column=3).value + ' ' + str(load sheet.cell(row=i,
column=4).value) + ' ' + str(load sheet.cell(row=i, column=7).value)
        posProb = getProb(sentence, f pos)
        negProb = getProb(sentence, f neg)
        midProb = getProb(sentence, f mid)
        load sheet.cell(row=i, column=8, value=findMaxProb(posProb,negProb,midProb))
        load excel.save(filename='output.xlsx')
```



## 엑셀에 전처리한 자연어 데이터 일부(욕설 필터링 미적용)

	Index	제목	본문	날짜	조회수	댓글	분류
A	В	C	D	E	F	G	Н
63	561	혹시 인터넷tv sk브로드밴드 쓰는 사람 있음??	캐치온 vod나 프리미어 월정액 7	2018.08.25	222	기본으로 쓰긴한디——먄 나 잘몰겠다	긍정
64	562	이번 에스케이 브로드밴드 씨엡 이상	뭘말하고 싶은거냐	2018.08.24	54		긍정
65	563	동남아 패킷뜨는사람	저만 패킷터짐??sk브로드밴드쓰	2018.08.24	201	나도 요즘 그렇다 kt쓰는데 - dc App	부정
66	564	진짜 개쓰레기 같은 sk 브로드밴드 시발련들	묘량 달빛 공명때 자꾸 트레이서	2018.08.24	90	역시인터넷은케이티지 ㅋㅋ 인터넷은 .	부정
67	565	175.113 sk브로드밴드 아이핀데 뭔 저격을 함	같은 회사면 대역도 비슷하겠지	2018.08.23	43		긍정
68	566	아이툴즈 해킹범 ip밴했다. 환불해달라하지마라	라이센스 1개 사갔으면 pc 1개어	2018.08.23	269	양심 ㅇㄷㅋㅋㅋㅋ 진짜 개념없네 ㅋㅋ	부정
69	567	헨리도 SK브로드밴드 모델임	헨리 완전 잘나가	2018.08.23	491	잘생이네 나혼산 다 슼 찍엇지않냐 이시	부정
70	568	이서진 광고 킹이네 sk브로드밴드 모델임	ССС	2018.08.23	608	사실상 현재 젤 이미지좋은 연옌 그 광.	부정
71	569	이서진 sk 광고찍엇네	sk브로드밴드	2018.08.23	491	이서진 광고 진짜 엄청나네ㄷㄷ 긷스타	부정
72	570	SK 브로드밴드 인터넷 좆구린데 어케 못하냐	해지하면 위약금 오지게 나오는	2018.08.22	100	나도 시발 원룸전체가 sk라 답도없다 =	부정
73	571	SK 브로드밴드 인터넷 좆구린데 해지하면 위약금 존나	어케 못하나3년 더 써야함	2018.08.22	148	ㅇㅇ 써야지 뭐 그래도 해외인터넷 존	부정
74	572	하우스 전편 구매 가능한 사이트	글들 읽어보니까 sk브로드밴드어	2018.08.20	380	토렌트를 이용하면 1080p가 공짜!	부정
75	573	7급과 초등교사 비교해준다	대기업의 인수합병을 막아버린 등	2018.08.18	2420	특히 최근 들어서는 국장급 이하 직원들	부정
76	574	폰사는거랑 통신사에서 고민중임	KT는 장기고객이 영 아닌데 홈파	2018.08.17	84	Sk가조은거같다 sk가즈아 근디묶음상품	부정
77	575	'몸값 비싼 팀장들'잘나가는 테크기업 연봉킹은?	인터넷과 게임 등 주요 테크 기입	2018.08.16	185		부정
78	576	SK 브로드밴드 절대로 쓰지 마라. AS가 개판이구나.	기가 프리미엄이라고 SK최초로	2018.08.15	153	ㄷㄷ - dc App 으휴 병신	부정
79	577	인랑 볼만하냐?	sk 브로드밴드 포인트 남아서 영	2018.08.15	29	클레멘타인 급 띵작임 보셈	부정
80	578	SK브로드밴드 때문에 자주 핑튀는 애들 참고해라	참고 하라고	2018.08.14	1817	쫌웃겼다 나쁜새끼네 이거 ㅆ1발	부정
81	579	SKT 자주 끊어지는분들 참조 하세요.	SK브로드밴드 기가인터넷 사용하	2018.08.14	307	skt알바임 내가봄 sk 전용공유기라하면	부정
82	580	SK 브로드밴드로 바꿨더니 핑그랩소리듣는다	갑자기 핑튀어서 내가 쫓던 생존	2018.08.12	203	병신 ㅋㅋ 자체 핑그랩 무엇 - dc App	부정
83	581	그립읍니다	SK 브로드밴드 기사 : 최정 - dc	2018.08.12	36		중립
84	582	sk 브로드밴드 공유기 맥어케 바꿈 ?_?	안바껴 :(	2018.08.12	107	http://ddf2net.tistory.com/24 맞나 모르	부정
→ →			,			: (	

