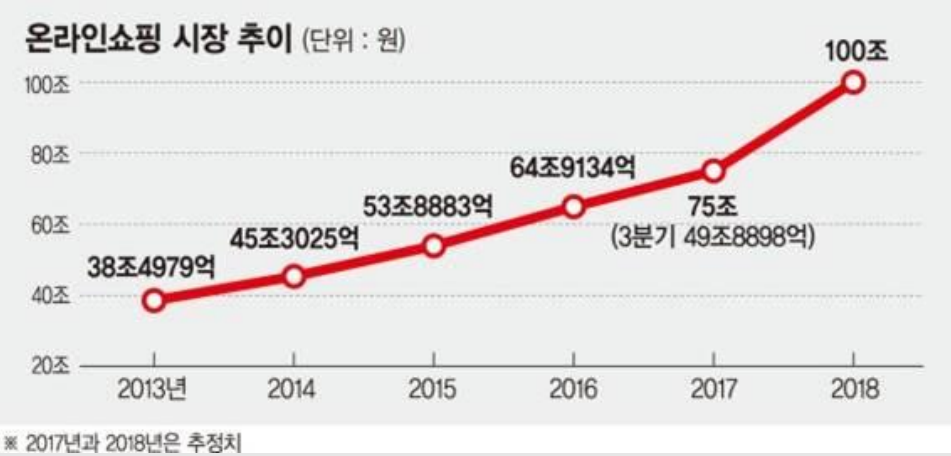
쿠팡 리뷰

키워드 · 긍부정 분석

캡스톤디자인 I
201804236
이 혜 규

목차

목차



- 온라인 쇼핑 시장이 점점 커지고 업체 등이 다양해짐

- 후기를 보고 구매 여부 결정.
- 하지만, 사용자와 리뷰가 늘어나면 하나 하나 읽기 힘들고 시간이 많이 소요.

후기를 한번에 보기 쉽게 할 방법 필요

목차

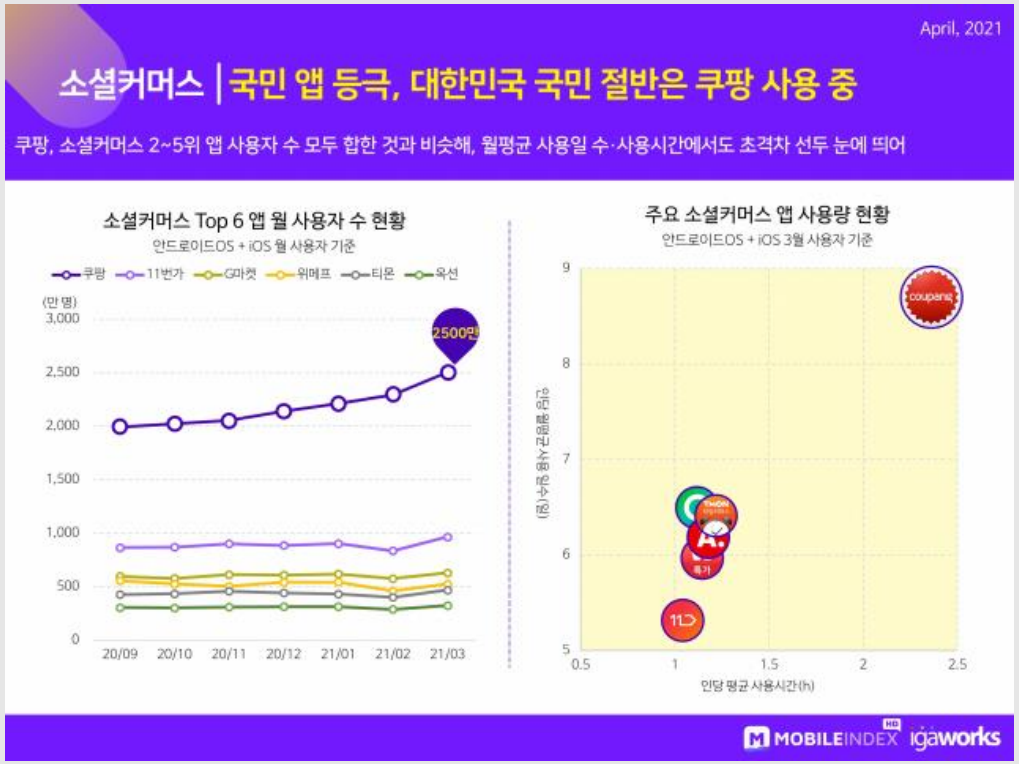
coupang

한국 상위 10 쇼핑 앱

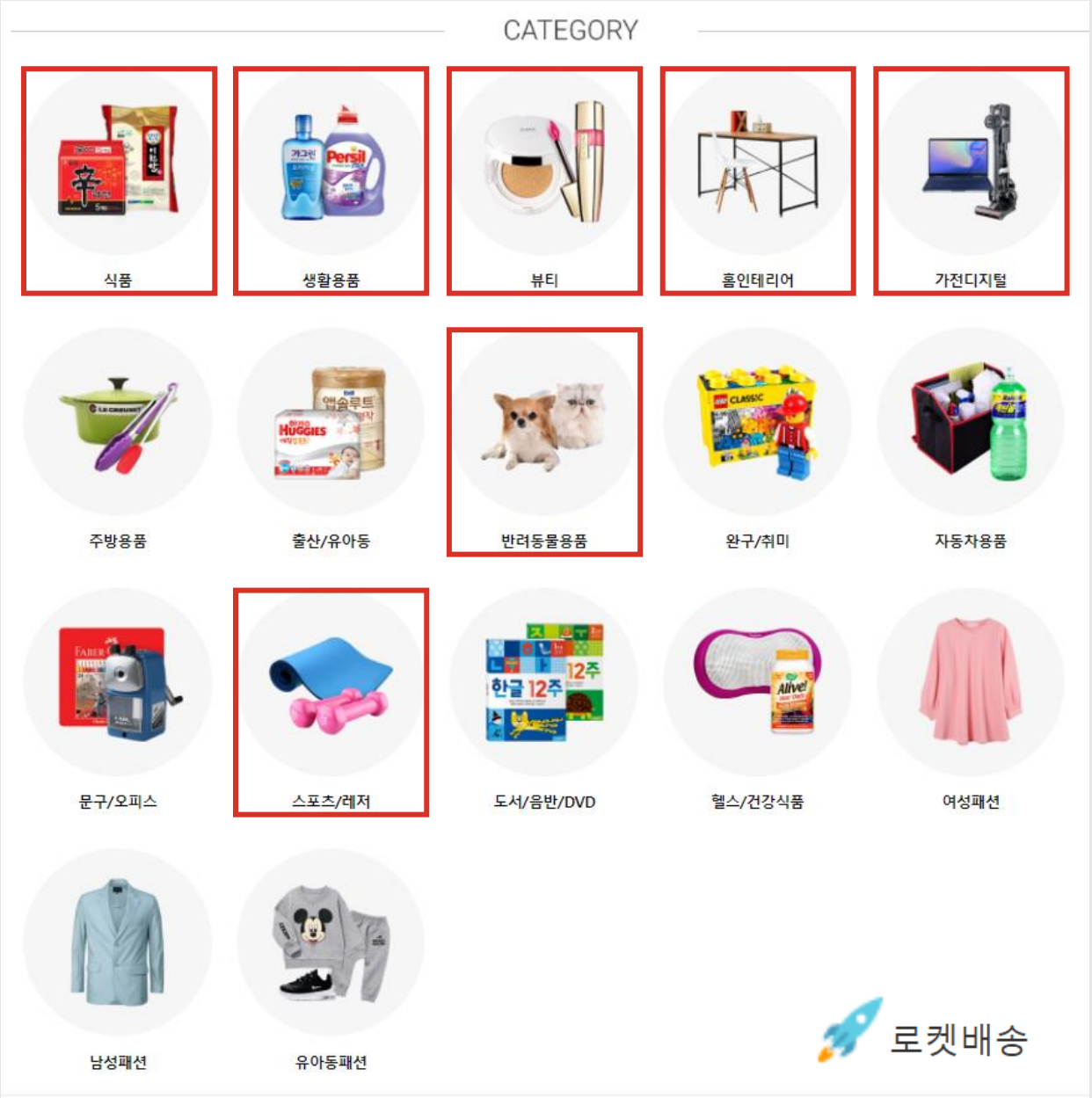
MAU | 2020월 11월 첫째 주

Rank	Name	Parent Company
1	쿠팡	Coupang
2	11번가	SK Group
3	지마켓	eBay
4	티몬	KKR
5	위메이크프라이스	WEMAKEPRICE
6	옥션	eBay
7	무신사	Musinsa
8	올리브영	CJ Group
9	SSG.COM	SHINSEGAE
10	인터파크 쇼핑	Interpark INT

Note: Combined iPhone and Android Phone



목차



목차

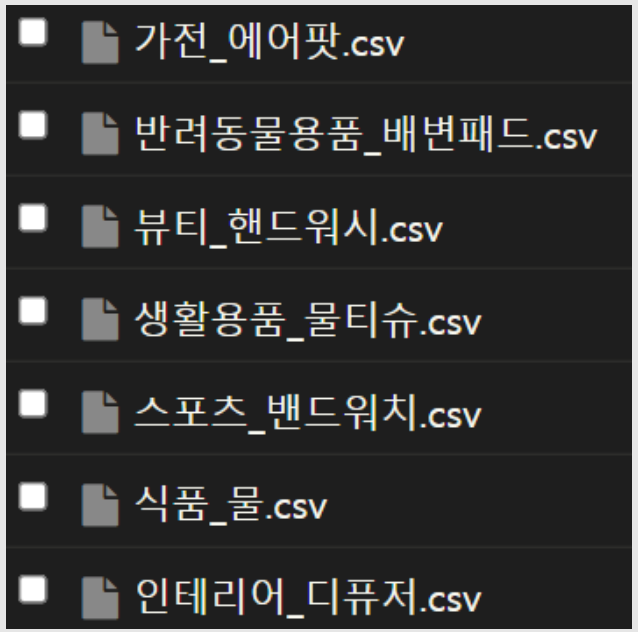

```
import pandas as pd
df = pd.DataFrame(columns=['name', 'date', 'review', 'score'])

url1='https://www.coupang.com/vp/product/reviews?productId=206038267&page='
url2='&size=100&sortBy=ORDER_SCORE_ASC&ratings='
url3='&q=&viRoleCode=3&ratingSummary=true'

nob="명에게 도움 됨"

for i in range(5, 0,-1):
    for k in range(1, 11):
        driver = webdriver.Chrome('/Users/user/Downloads/chromedriver.exe')
        url = url1 + str(k) + url2 + str(i) + url3
        driver.get(url)
        time.sleep(5)

        for j in range(1, 101):
            try:
                user=driver.find_element_by_xpath('''/html/body/article[%s]/div[1]/div[2]/span'''%j).text.strip()
                date=driver.find_element_by_xpath('''/html/body/article[%s]/div[1]/div[3]/div[2]'''%j).text.strip()
                rev=driver.find_element_by_xpath('''/html/body/article[%s]/div[4]/div'''%j).text.strip()
                #영상 올리거나 사진 많이 올리면 xpath가 다름
                if nob in rev:
                    rev=driver.find_element_by_xpath('''/html/body/article[%s]/div[3]/div'''%j).text.strip()
                    if nob in rev:
                        rev=driver.find_element_by_xpath('''/html/body/article[%s]/div[3]'''%j).text.strip()

                df=df.append({'name' : user, 'date' : date, 'review' : rev, 'score' : i} , ignore_index=True)
            except:
                pass
        driver.close()
```

7개의 productId만 변경하며 진행

- 가전_에어팟.csv
- 반려동물용품_배변패드.csv
- 뷰티_핸드워시.csv
- 생활용품_물티슈.csv
- 스포츠_밴드워치.csv
- 식품_물.csv
- 인테리어_디퓨저.csv

목차

- 1점 ~ 5점 각 1000개의 리뷰, 최대 4995개 수집하여 파일 저장
- 중복된 데이터는 제거하여 저장



카테고리	수집한 데이터	전체 리뷰
가전_에어팟	3942	53,803
반려동물용품_배변패드	4343	45,111
뷰티_핸드워시	2556	27,042
생활용품_물티슈	4844	81,505
스포츠_밴드워치	2199	12,145
식품_물	4895	518,921
인테리어_디퓨저	4998	54,185

목차

- 데이터 불러오기

```
import pandas as pd
df=pd.read_csv('인테리어_디퓨저.csv', header=0, encoding='utf-8', index_col=0)
df
```

	name	date	review	score
0	이은혜	2021.03.24	솔직 구매 후기예요 !!₩n₩n구매시기는 뉴코코도르 되기전 좀 오래전 구매 , 사용...	5
1	심콩수리	2021.02.07	좋은 평판 믿고 화장실에 두려고 코코도르 디퓨저 500ml 대용량을 구매해 사용하는...	5
2	허*미	2021.04.06	봄바람 휘날리며~~~₩n그대여그대여 오늘은 우리 같이 걸어요~~ㅋㅋ₩n₩n어느새 집...	5
3	KMJ71	2020.05.20	국민 디퓨저라고 하더라고요 코코도르^^₩n디퓨저가 갑자기 붐을 일으켰을때 가격대가 ...	5
4	최지우	2017.08.22	코코도르 디퓨저는 디퓨저용기도 깔끔하고 심플해서₩n여러번구매를했었었습니다 :)₩n이...	5
5	이니짱	2017.06.04	● 상품명 : 코코도르 디퓨저 200ml 2개₩n₩n₩n● 향 : 프렌치라벤더₩n₩...	5

```
df.info()
```

4998

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4998 entries, 0 to 4999
Data columns (total 4 columns):
name      4998 non-null object
date      4998 non-null object
review    4862 non-null object
score     4998 non-null int8
dtypes: int8(1), object(3)
memory usage: 161.1+ KB
```

NA값 제거

```
df.info()
```

4862

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4862 entries, 0 to 4861
Data columns (total 4 columns):
name      4862 non-null object
date      4862 non-null object
review    4862 non-null object
score     4862 non-null int8
dtypes: int8(1), object(3)
memory usage: 118.8+ KB
```

목차

- 특수기호 삭제 한 후 `doc' 열로 저장

```
df['doc'] = df['review'].str.replace("[^ㄱ-하-ㅣ가-힣]", " ")
df
```

	name	date	review	score	doc
0	이은혜	2021.03.24	솔직 구매 후기예요 !!₩n₩n구매시기는 뉴코코도르 되기 전 좀 오래전 구매 , 사용...	5	솔직 구매 후기예요 구매시기는 뉴코코도르 되기전 좀 오래전 구매 사용했던...
1	심종수리	2021.02.07	좋은 평판 믿고 화장실에 두려고 코코도르 디퓨저 500ml 대용량을 구매해 사용하는...	5	좋은 평판 믿고 화장실에 두려고 코코도르 디퓨저 대용량을 구매해 사용하는...
2	허*미	2021.04.06	봄바람 휘날리며~~~₩n그대여그대여 오늘은 우리 같이 걸어요~~ㅋㅋ₩n₩n어느새 집...	5	봄바람 휘날리며 그대여그대여 오늘은 우리 같이 걸어요 ㅋㅋ 어느새 집앞 공...
3	KMJ71	2020.05.20	국민 디퓨저라고 하더라고요 코코도르^^₩n디퓨저가 갑자기 붐을 일으켰을때 가격대가 ...	5	국민 디퓨저라고 하더라고요 코코도르 디퓨저가 갑자기 붐을 일으켰을때 가격대가 사...
4	최지우	2017.08.22	코코도르 디퓨저는 디퓨저용기도 깔끔하고 심플해서₩n여러번구매를했었었습니다 :)₩n이...	5	코코도르 디퓨저는 디퓨저용기도 깔끔하고 심플해서 여러번구매를했었었습니다 이번에...

- 불용어 설정

```
from konlpy.tag import Okt

f = open("korean_stopwords.txt", 'r', encoding='UTF8')
read = f.read()
stop_words = read.split()

name ="뉴 코코도르 디퓨저 200ml 2개"
okt = Okt()
st= okt.morphs(name)

for i in range(len(stop_words)):
    st.append(stop_words[i])
st
```

"아", "예를 들면", "은/는/이/가", "을/를", "이/그/저" …

목차


```
def clean(dfdf):
    okt = Okt()
    len_df = len(dfdf)
    dfdf=reduce_mem_usage(dfdf)
    dfdf['word']=0
    for i in range(len_df):
        lists=[]
        t_words = okt.nouns(dfdf['doc'].iloc[i])
        for w in t_words:
            if w not in st:
                lists.append(w)
        dfdf['word'].loc[i]=lists
    return dfdf
```

- 명사만 리스트로 구분하여 `word'열로 저장

	name	date	review	score	doc	word	rate
0	이은혜	2021.03.24	솔직 구매 후기예요 !!₩n₩n구매시기는 뉴코코도르 되기전 좀 오래전 구매 , 사용...	5	솔직 구매 후기예요 구매시기는 뉴코코도르 되기전 좀 오래전 구매 사용했던...	[시기, 기전, 사용, 터, 사진, 리뉴, 상품, 사진, 참고, 집안, 향, 바, ...	1
1	심콩수리	2021.02.07	좋은 평판 믿고 화장실에 두려고 코코도르 디퓨저 500ml 대용량을 구매해 사용하는...	5	좋은 평판 믿고 화장실에 두려고 코코도르 디퓨저 대용량을 구매해 사용하는...	[평판, 화장실, 려고, 용량, 사용, 중이, 정말, 집, 화장, 실용, 커서, 용...	1
2	허*미	2021.04.06	봄바람 휘날리며~~~₩n그대여그대여 오늘은 우리 같이 걸어요~~ㅋㅋ₩n₩n어느새 집...	5	봄바람 휘날리며 그대여그대여 오늘은 우리 같이 걸어요 ㅋㅋ 어느새 집 앞 공...	[봄바람, 대여, 대여, 오늘, 집 앞, 공원, 벚꽃, 활, 짝, 계절, 집안, 분위...	1
3	KMJ71	2020.05.20	국민 디퓨저라고 하더라고요 코코도르 ^^₩n디퓨저가 갑자기 붐을 일으켰을때 가격대가 ...	5	국민 디퓨저라고 하더라고요 코코도르 디퓨저가 갑자기 붐을 일으켰을 때 가격대가 사...	[국민, 갑자기, 붐, 가격, 대가, 방향, 제나, 사후, 딸, 얘길, 사춘기, 여...	1
4	최지우	2017.08.22	코코도르 디퓨저는 디퓨저용기도 깔끔하고 심플해서₩n여러번구매를했었었습니다 :)₩n이...	5	코코도르 디퓨저는 디퓨저용기도 깔끔하고 심플해서 여러번구매를했었었습니다 이번에...	[용기, 심플, 번, 향, 중, 제일, 향, 에어, 프릴, 구입, 도움, 평, 배송...	1

목차

- 단어의 빈도가 많은 순서대로 정렬

```
from collections import Counter
counter = Counter(word)
most_common = counter.most_common(15)
most_common
```

```
[('향', 8078),
 ('냄새', 2259),
 ('스틱', 1291),
 ('화장실', 789),
 ('향기', 772),
 ('사용', 738),
 ('집', 716),
 ('안나', 702),
 ('방', 700),
 ('생각', 661),
 ('정도', 560),
 ('발향', 559),
 ('가격', 555),
 ('처음', 527),
 ('체리', 524)]
```

향
냄새
스틱
화장실
향기

목차

- TF-IDF(Term Frequency + Inverse Document Frequency)
  - ✓ 검색과 텍스트 마이닝에서 주로 이용하는 가중치 방법
  - ✓ 여러 문서로 이루어진 문서에 있을 때 어떤 단어가 특정 문서내에서의 중요도 계산
  - ✓ 주로 문서의 핵심어 추출, 검색엔진에서 검색 순위 결정 등에 사용

목차

- 원래 리뷰의 점수에서 정확도 측정을 위해 임시로 구분

```
import numpy as np
conditionlist = [(df_clean['score'] >= 4), (df_clean['score'] == 3), (df_clean['score'] <= 2)]
choicelist = [1, 0, -1]
df_clean['rate'] = np.select(conditionlist, choicelist, default='Not Specified')
df_clean
```

	name	date	review	score	doc	word	rate
0	이은혜	2021.03.24	솔직 구매 후기예요 !!₩n₩n구매시기는 뉴코코도르 되기전 좀 오래전 구매 , 사용...	5	솔직 구매 후기예요 구매시기는 뉴코코도르 되기전 좀 오래전 구매 사용했던...	[시기, 기전, 사용, 터, 사진, 리뉴, 상품, 사진, 참고, 집안, 향, 바, ...	1
1	심콩수리	2021.02.07	좋은 평판 믿고 화장실에 두려고 코코도르 디퓨저 500ml 대용량을 구매해 사용하는...	5	좋은 평판 믿고 화장실에 두려고 코코도르 디퓨저 대용량을 구매해 사용하는...	[평판, 화장실, 려고, 용량, 사용, 중이, 정말, 집, 화장, 실용, 커서, 용...	1
2	허*미	2021.04.06	봄바람 휘날리며~~~₩n그대여그대여 오늘은 우리 같이 걸어요~~ㅋㅋ₩n₩n어느새 집...	5	봄바람 휘날리며 그대여그대여 오늘은 우리 같이 걸어요 ㅋㅋ 어느새 집앞 공...	[봄바람, 대여, 대여, 오늘, 집앞, 공원, 벚꽃, 활, 짝, 계절, 집안, 분위...	1
3	KMJ71	2020.05.20	국민 디퓨저라고 하더라고요 코코도르 ^^₩n디퓨저가 갑자기 붐을 일으켰을 때 가격대가 ...	5	국민 디퓨저라고 하더라고요 코코도르 디퓨저가 갑자기 붐을 일으켰을때 가격대가 사...	[국민, 갑자기, 붐, 가격, 대가, 방향, 제나, 사후, 딸, 얘길, 사춘기, 여...	1

긍정 4, 5 -> 1 -> train

모호 3 -> 0 -> test

부정 1, 2 -> -1 -> train

```
df["rate"].value_counts()
```

```
1     2000
-1    1862
0     1000
Name: rate, dtype: int64
```

목차

- Train / test 구분

```
from sklearn.feature_extraction.text import TfidfTransformer

bow_vect = vect.fit_transform(df_train['doc'].tolist())
tfidf_vectorizer = TfidfTransformer()
tf_idf_vect = tfidf_vectorizer.fit_transform(bow_vect)
```

```
from sklearn.model_selection import train_test_split

x = tf_idf_vect
y = df_train['rate']
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state=42)
```

```
print("x_train", x_train.shape)
print("x_test", x_test.shape)
print("y_train", y_train.shape)
print("y_test", y_test.shape)
```

```
x_train (3089, 4022)
x_test (773, 4022)
y_train (3089,)
y_test (773,)
```

목차

- 모델링 및 모델 평가

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.pipeline import Pipeline

lr = LogisticRegression()
lr.fit(x_train, y_train)

y_pred = lr.predict(x_test)
```

```
print('정확도: %.2f' % accuracy_score(y_test, y_pred))
```

```
정확도: 0.84
```

목차

- 긍정에서 영향을 많이 준 단어 / 부정에서 영향을 많이 준 단어

```
pos=[]
for coef in coef_pos_index[:10]:
    pos.append(invert_index_vectorizer[coef[1]])

neg=[]
for coef in coef_neg_index[:10]:
    neg.append(invert_index_vectorizer[coef[1]])

final = pd.DataFrame({"긍정" : pos, "부정" : neg})
final
```

	긍정	부정
0	달달	안나
1	기분	별로
2	가성	안남
3	사용	전혀
4	배송	반품
5	용량	최악
6	걱정	싸구려
7	퓨저	알콜
8	집안	대야
9	친구	사지

목차

- Train / test 구분

```
X_train = []
for i in range(len(df_train)):
    X_train.append(df_train['word'].iloc[i])
```

```
X_test = []
for i in range(len(df_test)):
    X_test.append(df_test['word'].iloc[i])
```

- 인코딩

```
from tensorflow.keras.preprocessing.text import Tokenizer
tokenizer = Tokenizer()
tokenizer.fit_on_texts(X_train)
```

목차

- 샘플 길이 조절

```
def below_threshold_len(max_len, nested_list):
    cnt = 0
    for s in nested_list:
        if(len(s) <= max_len):
            cnt = cnt + 1
    print('전체 샘플 중 길이가 %s 이하인 샘플의 비율: %s'%(max_len, (cnt / len(nested_list))*100))
```

```
below_threshold_len(30, X_train)
```

```
전체 샘플 중 길이가 30 이하인 샘플의 비율: 93.26773692387364
```

- 모델링

```
model = Sequential()
model.add(Embedding(vocab_size, 100))
model.add(LSTM(10))
model.add(Dense(1, activation='relu'))
```

목차

- 긍정/부정 분류

```
for i in range(len(df_test)):
    score = float(model.predict(X_test))
    if(score >= 0.5):
        df_test.rate.iloc[i] = 1
        if(score < 0.5):
            df_test.rate.iloc[i] = -1
```

- 긍정/부정 리뷰 개수

```
df["rate"].value_counts()
```

```
 1    2000
-1    1862
-1    1000
Name: rate, dtype: int64
```

- 데이터 합치기

	name	date	review	score	rate
0	이은혜	2021.03.24	솔직 구매 후기예요 !!₩n₩n구매시기는 뉴코코도르 되기전 좀 오래전 구매 , 사용...	5	1
1	심콩수리	2021.02.07	좋은 평판 믿고 화장실에 두려고 코코도르 디퓨저 500ml 대용량을 구매해 사용하는...	5	1
2	허*미	2021.04.06	봄바람 휘날리며~~~₩n그대여그대여 오늘은 우리 같이 걸어요~~ㅋㅋ₩n₩n어느새 집...	5	1
3	KMJ71	2020.05.20	국민 디퓨저라고 하더라고요 코코도르^^₩n디퓨저가 갑자기 봄을 일으켰을때 가격대가 ...	5	1
4	최지우	2017.08.22	코코도르 디퓨저는 디퓨저용기도 깔끔하고 심플해서₩n여러번구매를했었었습니다 :)₩n이...	5	1

목차

빈도수 기준	긍정 / 부정 기준
향 냄새 스틱 화장실 향기	긍정 부정 0 달달 안나 1 기분 별로 2 가성 안남 3 사용 전혀 4 배송 반품
⇒ 스틱이 필요하고 화장실에서 향기가 난다.	– 긍정 : 달달한 향이고 사용할 때 기분에 영향을 준다. 가성비와 배송의 문제는 확인이 필요하다. – 부정 : 향이 전혀 안나기도 하고 별로 안난다. 반품의 의사가 있다.

목차

긍정 리뷰 수	2000
부정 리뷰 수	2862

부정적인 리뷰 수가 많기 때문에 추천하는 상품은 아니다.

목차

- 처음 해본 텍스트/감성 분류 분석이였다. 재밋고 신기했다.

- 리뷰가 전체적으로 부정적으로 되어있다. 모델링 과정에서의 문제 같아서 아쉽다.
- 페이크 리뷰에 대한 신뢰성 문제는 해결하지 못했다.

목차

✓ https://brunch.co.kr/@oms1225/46

✓ https://www.i-boss.co.kr/ab-6141-40891

✓ https://www.sedaily.com/NewsView/22L8G90XX8

✓ http://www.discoverynews.kr/sub_read.html?uid=331884