영화리뷰 데이터에서 긍정키워드, 부정키워드 추출

전처리 및 토큰화

[16] 34초
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

[17] 3초
```
import pandas as pd
df = pd.read_csv('/content/drive/MyDrive/강의자료/핀인사이트_강의자료 /2023 강의자료 업데이트/자연어처리/실습/data/review.csv',
                 sep = '\t')
# df = df[:500]
df = (df.dropna()).reset_index(drop = True)
df.head()
```

	id	document	label
0	9976970	아 더빙.. 진짜 짜증나네요 목소리	0
1	3819312	흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나	1
2	10265843	너무재밋었다그래서보는것을추천한다	0
3	9045019	교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정	0
4	6483659	사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 ...	1

[18] 0초
```
# !pip install konlpy
```

데이터 전처리 및 토큰화

0초
```
# 경고메세지 끄기
import warnings
warnings.filterwarnings(action='ignore')

import re
import numpy as np

from konlpy.tag import Okt
```

[20] 0초
```
# 전처리 전 데이터프레임
df.head(3)
```

[20]

	id	document	label
0	9976970	아 더빙.. 진짜 짜증나네요 목소리	0
1	3819312	흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나	1
2	10265843	너무재밋었다그래서보는것을추천한다	0

[21]

```
# '한글'을 제외한 다른 문자 모두 제거
remove_except_ko = re.compile(r"[^가-힣ㄱ-ㅎㅏ-ㅣ\\s]")
def preprocess_remove(text):
    text = re.sub(remove_except_ko,' ',text).strip()
    return text

df['document'] = df['document'].map(lambda x : preprocess_remove(x))
df.head(3)
```

	id	document	label
0	9976970	아 더빙 진짜 짜증나네요 목소리	0
1	3819312	흠 포스터보고 초딩영화줄 오버연기조차 가볍지 않구나	1
2	10265843	너무재밋었다그래서보는것을추천한다	0

```
# 토큰화 / 불용어처리
stop_pos = ['Josa','Eomi','Punctuation','Foreign','Number','Unknown','KoreanParticle']
stop_word = ['영화','정말','진짜']

tokenizer= Okt()
df['morphs'] = None
for i,row in df.iterrows():
  tokens = tokenizer.pos(row['document'])
  token_ls = []
  for token in tokens:
    if len(token[0]) > 1:
      if token[0] not in stop_word:
        if token[1] not in stop_pos:
          token_ls.append(token[0])
  # print(token_ls)
  df['morphs'][i] = ' '.join(token_ls)
df.head()
```

[22]

	id	document	label	morphs
0	9976970	아 더빙 진짜 짜증나네요 목소리	0	더빙 짜증나네요 목소리
1	3819312	흠 포스터보고 초딩영화줄 오버연기조차 가볍지 않구나	1	포스터 보고 초딩 오버 연기 가볍지 않구나
2	10265843	너무재밋었다그래서보는것을추천한다	0	무재 밋었 다그 래서 보는것을 추천
3	9045019	교도소 이야기구먼 솔직히 재미는 없다 평점 조정	0	교도소 이야기 구먼 솔직히 재미 없다 평점 조정
4	6483659	사이몬페그의 익살스런 연기가 돋보였던 영화 스파이더맨에서 늙어보이기만 했던 커스틴 ...	1	사이 몬페 익살스런 연기 돋보였던 스파이더맨 늙어 보이기만 했던 커스틴 던스트 너무...

키워드 추출

[23]

```python
from sklearn.feature_extraction.text import TfidfVectorizer

def get_keyword(text):
  tfidf_vect = TfidfVectorizer()
  tfidf_v = tfidf_vect.fit_transform([text])

  keyword = tfidf_v.tocoo()

  sorted_words = sorted(zip(keyword.col, keyword.data), key=lambda x:(x[1], x[0]), reverse=True)
  feature_name = tfidf_vect.get_feature_names_out()
  keywords = [(feature_name[i], score) for i, score in sorted_words[:5]]

  word_ls = []
  for word in keywords:
    word_ls.append(word[0])

  return word_ls
```

34초

```python
text = df['morphs'][59]
df['keywords'] = None
for i, row in df.iterrows():
  # row['morphs']
  if len(row['morphs'])>1:
    df['keywords'][i] = ' '.join(get_keyword(row['morphs']))
df.head(10)
```

	id	document	label	morphs	keywords
0	9976970	아 더빙 진짜 짜증나네요 목소리	0	더빙 짜증나네요 목소리	짜증나네요 목소리 더빙
1	3819312	흠 포스터보고 초딩영화줄 오버연기조차 가볍지 않구나	1	포스터 보고 초딩 오버 연기 가볍지 않구나	포스터 초딩 오버 연기 않구나
2	10265843	너무재밋었다그래서보는것을추천한다	0	무재 밋었 다그 래서 보는것을 추천	추천 보는것을 밋었 무재 래서
3	9045019	교도소 이야기구먼 솔직히 재미는 없다 평점 조정	0	교도소 이야기 구먼 솔직히 재미 없다 평점 조정	평점 조정 재미 이야기 없다
4	6483659	사이몬페그의 익살스런 연기가 돋보였던 영화 스파이더맨에서 늙어보이기만 했던 커스틴 ...	1	사이 몬페 익살스런 연기 돋보였던 스파이더맨 늙어 보이기만 했던 커스틴 던스트 너무...	했던 커스틴 익살스런 이뻐 연기
5	5403919	막 걸음마 뗀 세부터 초등학교 학년생인 살용영화 ㅋㅋㅋ 별반개도 아까움	0	걸음 초등학교 학년 생인 살용 반개 아까	학년 초등학교 아까 생인 살용

부정 키워드추출

[25]
```
df = df.dropna()
neg = df[df['label']==0]

total_neg=''
for i, row in neg.iterrows():
  total_neg += ' '.join(row['keywords'].split(' '))
  total_neg += ' '
```

[26]
```
total_neg
```

'짜증나네요 목소리 더빙 추천 보는것을 밎었 무재 래서 평점 조정 재미 이야기 없다 학년 초등학교 아까 생인 살용 했다 제대로 원작 살려내지못 긴장감 반복 해도 하는 이응경 연기 횡단보도 쳐나 이범수 울면 연기 어거지 감동 취향 존중 스토리 하는데 표절 재미 이해 없어지냐 음식 별로 바베트 만찬 할랫 지루하다 중반 주제 좋은데 짤랐을꺼 없었던거야 납득 꺼야 그럴꺼야 카밀라 발연기 진부하고말 아까워 쓰레기 시간 도안 죄인 입니다 기대했던 일까 했던건 하고자 틱장애 키이라 포스터 있어 보이는데 관객 완전 없고 하나 웃긴거도 없음 낭비 평점 시간 속지 마시길 이민기 공감 파손 툭하면 캐릭터 하더군 이런거 없냐 수준 북한 합니다 우리 사랑 포퐁 진호 저그 작은 세르게이 혼자 했더니 원한 어쩌라고 애가 최고 심심한 카리스마 예측 없는 악역 보는내내 중간 아무튼 불알 당황 느낌 평범함 평범한 조금 일상 밋밋한게 빨리 한두 짜증 주인공 전개도 캐릭터 중반 재탕 작품 이제 초반 짜리 지루하고 연기력 손예진 뽑는게 박시환 맞냐 망신 노래실력 일본 이런 유치하다 건가 졸작 전개 어이없는 어설픈 결말 했던 항거 폭도 왕조 온몸 실망 매우 갈등 흥행 화해 한국영 평점 시작 사진 불안하더니만 분만 보며 진창 하는지도 엉망 모르겠고 먹고 이건 우뢰매 입니다 쓰레기 였다 별루 최악 성룡 화신 클라라 아닌데 볼라 본거 캐스팅 진심 재미없음 이훨 쓰면 봅니다 고은님 감독 않았고 스토리 무섭지도 했으면 있게 은은하고 억지스럽고 어땠을까 킬링타임 태어나서 처음 중간 불륜 로맨스 했구나 하아 하네 짬뽕 짜증 시간 솔직히 별루더 느낌 낭비 이상해 내용 전개 느리다 너무나 내용 흥미 전개 이야기 안되는군 소재 절대 쓰레기 보지마라 허풍 포장 특유 중국인 있어 좋았을텐데 잡는 이건 우상화 설정 별로 일까 아니 연기자 문제 했음 진심 전혀 제대로 욕심 어느 보여줬다면 많았던 빵점 않다 아주 모자라진 하고 믿어지냐 뫼비우스 만들어졌다는게 도둑 케이블 줬으면 나와 그만 하나 차이밍량 짬뽕 이냐리 섞인 없음 ...'

```
from collections import Counter

result = Counter(total_neg.split())
result
```

```
 '억지': 12,
 '질질': 14,
 '끌어': 1,
 '터져': 4,
 '평정': 1,
 '졸면서': 4,
 '이리': 13,
 '아오': 15,
 '카톡': 1,
 '잇음': 3,
 '있지만': 6,
 '잃은': 2,
 '생동감': 1,
 '살아는': 1,
 '박물관': 1,
 '한심한': 8,
 '추억': 7,
 '미화': 10,
 '노땅': 1,
 '인종차별': 1,
 '유괴': 1,
 '불협화음': 1,
 '후자': 2,
 '없어': 19,
 '추가': 1,
 '좋았을것': 1,
 '전작': 17,
```

```
[28]  1 word = []
      2 count = []
      3 for x in result:
      4   # print(x,result[x])
      5   word.append(x)
      6   count.append(result[x])
      7
      8 set_neg = dict(zip(word,count))
      9 print(len(set_neg))
```

```
7693
```

```
 1 # !pip install wordcloud
 2 # 한글 폰트 설정
 3 # 설치하고 한글 적용이 안된다면 , 런타임 > 런타임 다시시작 ; 하고 설치코드 제외하고 나머지 코드 실행
 4 !sudo apt-get install -y fonts-nanum
 5 !sudo fc-cache -fv
 6 !rm ~/.cache/matplotlib -rf
```

```
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
  fonts-nanum
0 upgraded, 1 newly installed, 0 to remove and 18 not upgraded.
Need to get 10.3 MB of archives.
After this operation, 34.1 MB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/universe amd64 fonts-nanum all 20200506-1 [10.3 MB]
Fetched 10.3 MB in 0s (21.1 MB/s)
debconf: unable to initialize frontend: Dialog
debconf: (No usable dialog-like program is installed, so the dialog based frontend cannot be used. at /usr/share/perl5/Debconf/FrontEnd/Dialog.pm line 78, <> line 1.)
debconf: falling back to frontend: Readline
debconf: unable to initialize frontend: Readline
debconf: (This frontend requires a controlling tty.)
debconf: falling back to frontend: Teletype
dpkg-preconfigure: unable to re-open stdin:
Selecting previously unselected package fonts-nanum.
(Reading database ... 120876 files and directories currently installed.)
Preparing to unpack .../fonts-nanum_20200506-1_all.deb ...
Unpacking fonts-nanum (20200506-1) ...
Setting up fonts-nanum (20200506-1) ...
Processing triggers for fontconfig (2.13.1-4.2ubuntu5) ...
/usr/share/fonts: caching, new cache contents: 0 fonts, 1 dirs
/usr/share/fonts/truetype: caching, new cache contents: 0 fonts, 3 dirs
/usr/share/fonts/truetype/humor-sans: caching, new cache contents: 1 fonts, 0 dirs
/usr/share/fonts/truetype/liberation: caching, new cache contents: 16 fonts, 0 dirs
/usr/share/fonts/truetype/nanum: caching, new cache contents: 12 fonts, 0 dirs
/usr/local/share/fonts: caching, new cache contents: 0 fonts, 0 dirs
/root/.local/share/fonts: skipping, no such directory
/root/.fonts: skipping, no such directory
/usr/share/fonts/truetype: skipping, looped directory detected
/usr/share/fonts/truetype/humor-sans: skipping, looped directory detected
/usr/share/fonts/truetype/liberation: skipping, looped directory detected
/usr/share/fonts/truetype/nanum: skipping, looped directory detected
/var/cache/fontconfig: cleaning cache directory
/root/.cache/fontconfig: not cleaning non-existent cache directory
```

[30]
```python
# 폰트경로확인
import matplotlib.font_manager as fm
sys_font = fm.findSystemFonts()
[f for f in sys_font if 'Nanum' in f]
```

```
['/usr/share/fonts/truetype/nanum/NanumMyeongjoBold.ttf',
 '/usr/share/fonts/truetype/nanum/NanumBarunGothic.ttf',
 '/usr/share/fonts/truetype/nanum/NanumSquareR.ttf',
 '/usr/share/fonts/truetype/nanum/NanumGothicCodingBold.ttf',
 '/usr/share/fonts/truetype/nanum/NanumSquareRoundB.ttf',
 '/usr/share/fonts/truetype/nanum/NanumMyeongjo.ttf',
 '/usr/share/fonts/truetype/nanum/NanumSquareRoundR.ttf',
 '/usr/share/fonts/truetype/nanum/NanumGothicCoding.ttf',
 '/usr/share/fonts/truetype/nanum/NanumSquareB.ttf',
 '/usr/share/fonts/truetype/nanum/NanumBarunGothicBold.ttf',
 '/usr/share/fonts/truetype/nanum/NanumGothic.ttf',
 '/usr/share/fonts/truetype/nanum/NanumGothicBold.ttf']
```

4초
```python
from wordcloud import WordCloud
import matplotlib.pyplot as plt

font_path = '/usr/share/fonts/truetype/nanum/NanumGothic.ttf'

wc = WordCloud(font_path = font_path,
               width=1200, height=800,
               scale=2.0, max_font_size=250,
               background_color = 'white')
gen = wc.generate_from_frequencies(set_neg)
plt.figure()
plt.imshow(gen)
```

```
<matplotlib.figure.Figure at 0x79be5906c9d0>
```

긍정키워드 추출하기

[33]
```
pos = df[df['label']==1]

total_pos=''
for i, row in pos.iterrows():
  total_pos += ' '.join(row['keywords'].split(' '))
  total_pos += ' '
```

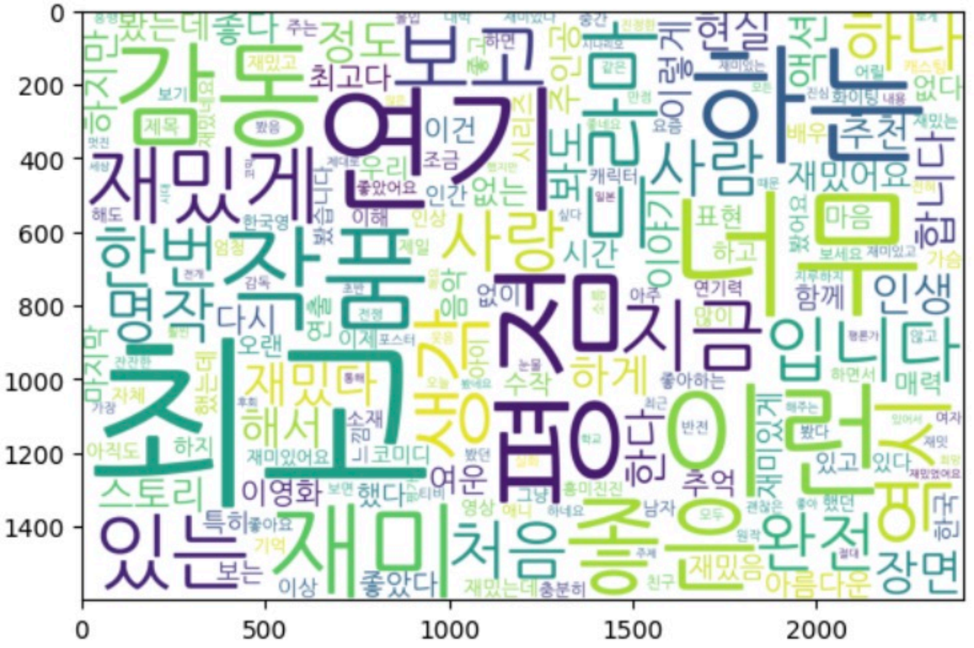
[34]
```
from collections import Counter

result = Counter(total_pos.split())
```

```
word = []
count = []
for x in result:
  # print(x,result[x])
  word.append(x)
  count.append(result[x])

set_pos = dict(zip(word,count))
print(len(set_pos))
```

```
7211
```

[36]
```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

wc = WordCloud(font_path = font_path,
               width=1200, height=800,
               scale=2.0, max_font_size=250,
               background_color = 'white')
gen = wc.generate_from_frequencies(set_pos)
plt.figure()
plt.imshow(gen)
```

<matplotlib.image.AxesImage at 0x79be549f99f0>



비교

```
1 # 부정
2 gen = wc.generate_from_frequencies(set_neg)
3 plt.figure()
4 plt.imshow(gen)
```

```
<matplotlib.image.AxesImage at 0x79be549f8e50>
```


```
[38] 1 # 긍정
2 gen = wc.generate_from_frequencies(set_pos)
3 plt.figure()
4 plt.imshow(gen)
```

```
<matplotlib.image.AxesImage at 0x79be53e31d50>
```