자연어 처리 실습 4

과목명 자연어 처리

담당교수 김낙현교수님

제출일 20211115

전공 컴퓨터전자시스템

학번 201904458

이름 이준용

한국외국어대학교

HANKUK UNIVERSITY OF FOREIGN STUDIES

```
pip install sentencepiece
```

```
Collecting sentencepiece
  Downloading sentencepiece-0.1.96-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.2 MB)
     |████████████████████████████████| 1.2 MB 5.0 MB/s
Installing collected packages: sentencepiece
Successfully installed sentencepiece-0.1.96
```

```
import pandas as pd
import sentencepiece as spm
import urllib.request
import csv
urllib.request.urlretrieve("https://raw.githubusercontent.com/e9t/nsmc/master/ratings.txt", filename="ratings.txt")
naver_df = pd.read_table('ratings.txt')
naver_df = naver_df.dropna(how = 'any') # Null 값이 존재하는 행 제거
# 결과를 naver_review.txt 파일에 저장
with open('naver_review.txt', 'w', encoding='utf8') as f: f.write('₩n'.join(naver_df['document']))
```

```
spm.SentencePieceTrainer.Train('--input=naver_review.txt --model_prefix=naver --vocab_size=10000 --model_type=bpe --m
```

```
naver_df['document'][range(80001,80011)]
```

```
80001                                이건 얄짤없이 10점이긴한디 재개봉인가요?
80002    너무 훌륭한 동물농장! 몇년째 보는 일요일 프로그램인데 오늘 정말 방치되어 굶어죽는...
80003                                        난 재밌던데?....
80004                                  귀엽고 멋지고 재미있는. 매력 덩어리
80005                          이방인과 현지인, 그들이 하나가 되는 순간의 코인로커
80006                                             ♥
80007      월드컵 기간에 보기엔 딱 좋은 영화! 이 영화만 100번 넘게 볼 정도로 강력 추천!
80008                                5점준놈 생각좀해 78년작인데 ㅡㅡ 멀더바래
80009                                  잘 찍었네요..하지만 흥행과는 무관할 듯
80010                                  이게 왜 8점대야 최소 9점대는 돼야지
Name: document, dtype: object
```

```
sp = spm.SentencePieceProcessor()
vocab_file = "naver.model"
sp.load(vocab_file)
lines = naver_df['document'][range(80001,80011)]

for line in lines:
  print(line)
  print(sp.encode_as_pieces(line))
  print(sp.encode_as_ids(line))
  print()
```

```
이건 얄짤없이 10점이긴한디 재개봉인가요?
['▁이건', '▁얄', '짤', '없이', '▁10', '점이', '긴', '한', '디', '▁재개봉', '인가요', '?']
[196, 7076, 9408, 370, 135, 498, 8505, 8291, 8431, 5607, 3025, 8329]

너무 훌륭한 동물농장! 몇년째 보는 일요일 프로그램인데 오늘 정말 방치되어 굶어죽는 강아지들 그리고 인간이 아닌것
['▁너무', '▁훌륭한', '▁동물', '농', '장', '!', '▁몇년', '째', '▁보는', '▁일', '요일', '▁프로그램', '인데', '▁
[23, 1617, 3037, 9129, 8344, 8303, 6196, 8872, 157, 104, 3723, 2401, 242, 952, 42, 443, 8400, 1755, 8275, 9842,

난 재밌던데?....
['▁난', '▁재밌던데', '?', '....']
[205, 4037, 8329, 47]

귀엽고 멋지고 재미있는. 매력 덩어리
['▁귀엽고', '▁멋지고', '▁재미있는', '.', '▁매력', '▁', '덩어리']
[2670, 5253, 1485, 8276, 396, 8275, 6715]

이방인과 현지인, 그들이 하나가 되는 순간의 코인로커
```

```
['▁이', '방', '인과', '▁현', '지', '인', ',', '▁그들이', '▁하나가', '▁되는', '▁순간', '의', '▁코', '인', '로',
[6, 8541, 3941, 240, 8281, 8308, 8315, 5884, 7408, 844, 1721, 8294, 215, 8308, 8299, 8767]

♥
['▁♥']
[6314]

월드컵 기간에 보기엔 딱 좋은 영화! 이 영화만 100번 넘게 볼 정도로 강력 추천!
['▁월드', '컵', '▁기', '간에', '▁보기엔', '▁딱', '▁좋은', '▁영화', '!', '▁이', '▁영화만', '▁100', '번', '▁넘기
[7116, 9619, 49, 5671, 2654, 547, 179, 5, 8303, 6, 4734, 1311, 8480, 4460, 97, 1450, 3969, 531, 8303]

5점준놈 생각좀해 78년작인데 ㅡㅡ 멀더바래
['▁5', '점준', '놈', '▁생각', '좀', '해', '▁7', '8', '년작', '인데', '▁ㅡㅡ', '▁멀', '더', '바', '래']
[543, 2110, 8765, 83, 8467, 8323, 536, 8619, 4451, 242, 488, 2076, 8366, 8448, 8412]

잘 찍었네요..하지만 흥행과는 무관할 듯
['▁잘', '▁찍', '었네요', '..', '하지만', '▁흥행', '과는', '▁무', '관', '할', '▁듯']
[63, 538, 2245, 3, 408, 1602, 2511, 58, 8486, 8391, 485]

이게 왜 8점대야 최소 9점대는 돼야지
['▁이게', '▁왜', '▁8', '점대', '야', '▁최소', '▁9', '점대는', '▁돼', '야지']
[244, 84, 497, 970, 8357, 5818, 486, 4341, 2616, 1155]
```

```
spm.SentencePieceTrainer.Train('--input=naver_review.txt --model_prefix=naver --vocab_size=20000 --model_type=bpe --m
```

```
sd = spm.SentencePieceProcessor()
vocab_file = "naver.model"
sd.load(vocab_file)
lines = naver_df['document'][range(80001,80011)]

for line in lines:
  print(line)
  print(sd.encode_as_pieces(line))
  print(sd.encode_as_ids(line))
  print()
```

```
이건 얄짤없이 10점이긴한디 재개봉인가요?
['▁이건', '▁얄', '짤', '없이', '▁10', '점이', '긴', '한', '디', '▁재개봉', '인가요', '?']
[196, 7076, 19408, 370, 135, 498, 18505, 18291, 18431, 5607, 3025, 18329]

너무 훌륭한 동물농장! 몇년째 보는 일요일 프로그램인데 오늘 정말 방치되어 굶어죽는 강아지들 그리고 인간이 아닌것
['▁너무', '▁훌륭한', '▁동물', '농', '장', '!', '▁몇년', '째', '▁보는', '▁일요일', '▁프로그램', '인데', '▁오늘
[23, 1617, 3037, 19129, 18344, 18303, 6196, 18872, 157, 9983, 2401, 242, 952, 42, 443, 18400, 1755, 18275, 1984

난 재밌던데?....
['▁난', '▁재밌던데', '?....']
[205, 4037, 16845]

귀엽고 멋지고 재미있는. 매력 덩어리
['▁귀엽고', '▁멋지고', '▁재미있는', '.', '▁매력', '▁', '덩어리']
[2670, 5253, 1485, 18276, 396, 18275, 6715]

이방인과 현지인, 그들이 하나가 되는 순간의 코인로커
['▁이', '방', '인과', '▁현', '지', '인', ',', '▁그들이', '▁하나가', '▁되는', '▁순간의', '▁코', '인', '로', '커
[6, 18541, 3941, 240, 18281, 18308, 18315, 5884, 7408, 844, 17975, 215, 18308, 18299, 18767]

♥
['▁♥']
[6314]

월드컵 기간에 보기엔 딱 좋은 영화! 이 영화만 100번 넘게 볼 정도로 강력 추천!
['▁월드컵', '▁기', '간에', '▁보기엔', '▁딱', '▁좋은', '▁영화', '!', '▁이', '▁영화만', '▁100', '번', '▁넘게',
[15237, 49, 5671, 2654, 547, 179, 5, 18303, 6, 4734, 1311, 18480, 4460, 97, 1450, 3969, 531, 18303]

5점준놈 생각좀해 78년작인데 ㅡㅡ 멀더바래
['▁5', '점준', '놈', '▁생각', '좀', '해', '▁7', '8', '년작', '인데', '▁ㅡㅡ', '▁멀', '더', '바', '래']
```

```
[543, 2110, 18765, 83, 18467, 18323, 536, 18619, 4451, 242, 488, 2076, 18366, 18448, 18412]

잘 찍었네요..하지만 흥행과는 무관할 듯
['_잘', '_찍', '었네요', '..', '하지만', '_흥행', '과는', '_무관', '할', '_듯']
[63, 538, 2245, 3, 408, 1602, 2511, 9718, 18391, 485]

이게 왜 8점대야 최소 9점대는 돼야지
['_이게', '_왜', '_8', '점대', '야', '_최소', '_9', '점대는', '_돼', '야지']
[244, 84, 497, 970, 18357, 5818, 486, 4341, 2616, 1155]
```

```
pip install konlpy
```

```
Collecting konlpy
  Downloading konlpy-0.5.2-py2.py3-none-any.whl (19.4 MB)
     |████████████████████████████████| 19.4 MB 1.2 MB/s
Requirement already satisfied: lxml>=4.1.0 in /usr/local/lib/python3.7/dist-packages (from konlpy) (4.2.6)
Collecting beautifulsoup4==4.6.0
  Downloading beautifulsoup4-4.6.0-py3-none-any.whl (86 kB)
     |████████████████████████████████| 86 kB 4.6 MB/s
Requirement already satisfied: numpy>=1.6 in /usr/local/lib/python3.7/dist-packages (from konlpy) (1.19.5)
Collecting colorama
  Downloading colorama-0.4.4-py2.py3-none-any.whl (16 kB)
Collecting JPype1>=0.7.0
  Downloading JPype1-1.3.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.whl (448 kB)
     |████████████████████████████████| 448 kB 58.8 MB/s
Requirement already satisfied: tweepy>=3.7.0 in /usr/local/lib/python3.7/dist-packages (from konlpy) (3.10.0)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.7/dist-packages (from JPype1>=0.7.0-
Requirement already satisfied: requests-oauthlib>=0.7.0 in /usr/local/lib/python3.7/dist-packages (from tweepy>
Requirement already satisfied: six>=1.10.0 in /usr/local/lib/python3.7/dist-packages (from tweepy>=3.7.0->konlp
Requirement already satisfied: requests[socks]>=2.11.1 in /usr/local/lib/python3.7/dist-packages (from tweepy>=
Requirement already satisfied: oauthlib>=3.0.0 in /usr/local/lib/python3.7/dist-packages (from requests-oauthli
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests[socks
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests[sock
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-package
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests[socks]>=2.
Requirement already satisfied: PySocks!=1.5.7,>=1.5.6 in /usr/local/lib/python3.7/dist-packages (from requests[
Installing collected packages: JPype1, colorama, beautifulsoup4, konlpy
  Attempting uninstall: beautifulsoup4
    Found existing installation: beautifulsoup4 4.6.3
    Uninstalling beautifulsoup4-4.6.3:
      Successfully uninstalled beautifulsoup4-4.6.3
Successfully installed JPype1-1.3.0 beautifulsoup4-4.6.0 colorama-0.4.4 konlpy-0.5.2
```

```
from collections import Counter
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import re
import urllib.request
from konlpy.tag import Okt
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
urllib.request.urlretrieve("https://raw.githubusercontent.com/e9t/nsmc/master/ratings_train.txt",
filename="ratings_train.txt")
f = open("ratings_train.txt",'r',encoding='utf-8')
result = f.read()

stopwords =['의','가','이','은','들','는','좀','잘','걍','과','도','를','으로','자','에','와','한','하다']
okt=Okt()
noun = okt.nouns(result)
noun = [word for word in noun if not word in stopwords] # 불용어 제거
count = Counter(noun)
noun_list = count.most_common(100)
```

```
for i in noun_list:
  print(i)
```

```
('장면', 2436)
('액션', 2397)
('주인공', 2382)
('걸', 2328)
('최악', 2268)
('지금', 2206)
('돈', 2205)
('이야기', 2174)
('별로', 2143)
('임', 2130)
('느낌', 2095)
('연출', 2082)
('개', 2064)
('끝', 2047)
('명작', 2041)
('듯', 2036)
('역시', 1994)
('이해', 1906)
('이영화', 1824)
('안', 1804)
('또', 1784)
('여자', 1742)
('때문', 1735)
('난', 1663)
('중', 1642)
('꼭', 1634)
('편', 1620)
('보기', 1611)
('기억', 1596)
('결말', 1579)
('마음', 1553)
('인생', 1541)
('소재', 1511)
('애', 1486)
('못', 1480)
('수준', 1448)
('현실', 1418)
('한번', 1402)
('가장', 1396)
('반전', 1382)
('매력', 1372)
('전개', 1368)
('남자', 1338)
('한국', 1337)
('가슴', 1331)
('저', 1300)
('음악', 1288)
('알', 1288)
('아이', 1279)
('원작', 1252)
('줄', 1240)
('인간', 1212)
('무슨', 1203)
('우리', 1199)
('추천', 1193)
('함', 1185)
('눈물', 1181)
('만', 1170)
('게', 1164)
```

```
stopwords =['의','가','이','은','들','는','좀','잘','걍','과','도','를','으로','자','에','와
okt = Okt() # 형태소 분석기
X_train = []
k = 0
for sentence in lines:
  k = k+1
  if k % 5000 == 0: print(k)
  temp_X = []
```

```
  temp_X = []
  temp_X = okt.morphs(sentence, stem=True) # 토큰화
  temp_X = [word for word in temp_X if not word in stopwords] # 불용어 제거
  X_train.append(temp_X)
```

```
for line in lines:
  print(line)
  noun = okt.nouns(line)
  count = Counter(noun)
  noun_list = count.most_common(100)
  for i in noun_list:
    print(i)
  print()
```

```
이건 얄짤없이 10점이긴한디 재개봉인가요?
('이건', 1)
('얄짤없', 1)
('점', 1)
('디', 1)
('재', 1)
('개봉', 1)

너무 훌륭한 동물농장! 몇년째 보는 일요일 프로그램인데 오늘 정말 방치되어 굶어죽는 강아지들 그리고 인간이 아
('동물농장', 2)
('년', 1)
('일요일', 1)
('프로그램', 1)
('오늘', 1)
('정말', 1)
('방치', 1)
('강아지', 1)
('인간', 1)
('견주', 1)
('다시', 1)
('한번', 1)
('감사', 1)
('일', 1)
('계세', 1)
('항상', 1)
('마음', 1)
('시청', 1)
('홧팅', 1)

난 재밌던데?....
('난', 1)

귀엽고 멋지고 재미있는. 매력 덩어리
('매력', 1)
('덩어리', 1)

이방인과 현지인, 그들이 하나가 되는 순간의 코인로커
('이방인', 1)
('현지', 1)
('그', 1)
('하나', 1)
('순간', 1)
('코인', 1)
('로커', 1)

♥

월드컵 기간에 보기엔 딱 좋은 영화! 이 영화만 100번 넘게 볼 정도로 강력 추천!
('영화', 2)
('월드컵', 1)
('기간', 1)
('보기', 1)
('이', 1)
('번', 1)
```

```
('볼', 1)
('정도', 1)
('강력', 1)
('추천', 1)
```

✓ 0초 오후 11:01에 완료됨