In [1]:
```
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']))
```

In [2]:
```
from collections import Counter

F = open('naver_review.txt', 'r', encoding='utf-8')

list_a = []
sen = F.readlines()
print("문장 수: ",len(sen))
sen = str(sen)
word = sen.split(' ');
print("단어의 수: ", len(word))

uw = set(word)
print("고유단어의 수: ", len(uw))
```

```
문장 수:  199992
단어의 수:  1518208
고유단어의 수:  500363
```

In [3]:
```
spm.SentencePieceTrainer.Train('--input=naver_review.txt --model_prefix=naver --vocab_size=10000 --model_type=bpe

sp = spm.SentencePieceProcessor()
vocab_file = "naver.model"
sp.load(vocab_file)

lines = [
    "난 재밌던데?....",
    "귀엽고 멋지고 재미있는. 매력 덩어리",
    "이방인과 현지인, 그들이 하나가 되는 순간의 코인로커",
    "♥",
    "월드컵 기간에 보기엔 딱 좋은 영화! 이 영화만 100번 넘게 볼 정도로 강력 추천!",
    "5점준놈 생각좀해 78년작인데 ㅡㅡ 멀더바래",
    "잘 찍었네요..하지만 흥행과는 무관할 듯",
    "이게 왜 8점대야 최소 9점대는 돼야지",
    "빌리 귀엽다잉 ㅋㅋㅋㅋㅋㅋㅋㅋㅋ 캐릭터 귀요미들",
    "허큘리스 개재밌음 난 학교에서 봤는데 왜 이걸 극장에서 못 봤는지 의문임 그리고 평점은 또 왜 이렇게 낮음ㅋㅋ",
]

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

```
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', '커']
[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', '▁넘게', '▁볼',
'▁정도로', '▁강력', '▁추천', '!']
[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', '▁□□', '▁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]

빌리 귀엽다잉 ㅋㅋㅋㅋㅋㅋㅋㅋㅋ 캐릭터 귀요미들
['▁빌', '리', '▁귀엽다', '잉', '▁ㅋㅋㅋㅋㅋㅋ', 'ㅋㅋㅋ', '▁캐릭터', '▁귀', '요', '미', '들']
[1636, 8298, 5299, 9098, 4258, 326, 542, 345, 8305, 8317, 8307]

허큘리스 개재밌음 난 학교에서 봤는데 왜 이걸 극장에서 못 봤는지 의문임 그리고 평점은 또 왜 이렇게 낮음ㅋㅋ
['▁허', '큘', '리스', '▁개', '재밌음', '▁난', '▁학교에서', '▁봤는데', '▁왜', '▁이걸', '▁극장에서', '▁못', '▁봤는지', '▁의문
', '임', '▁그리고', '▁평점은', '▁또', '▁왜', '▁이렇게', '▁낮', '음', 'ㅋㅋ']
[265, 9777, 693, 74, 3464, 205, 3659, 191, 84, 594, 1095, 89, 7912, 2971, 8440, 319, 1822, 257, 84, 246, 417, 833
0, 12]
```

In [4]:

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

sp = spm.SentencePieceProcessor()
vocab_file = "naver.model"
sp.load(vocab_file)

lines = [
    "난 재밌던데?....",
    "귀엽고 멋지고 재미있는. 매력 덩어리",
    "이방인과 현지인, 그들이 하나가 되는 순간의 코인로커",
    "♥",
    "월드컵 기간에 보기엔 딱 좋은 영화! 이 영화만 100번 넘게 볼 정도로 강력 추천!",
    "5점준놈 생각좀해 78년작인데 ㅡㅡ 멀더바래",
    "잘 찍었네요..하지만 흥행과는 무관할 듯",
    "이게 왜 8점대야 최소 9점대는 돼야지",
    "빌리 귀엽다잉 ㅋㅋㅋㅋㅋㅋㅋㅋㅋ 캐릭터 귀요미들",
    "허큘리스 개재밌음 난 학교에서 봤는데 왜 이걸 극장에서 못 봤는지 의문임 그리고 평점은 또 왜 이렇게 낮음ㅋㅋ",
]

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

```
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', '▁볼', '▁정
도로', '▁강력', '▁추천', '!']
[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', '▁□□', '▁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]

빌리 귀엽다잉 ㅋㅋㅋㅋㅋㅋㅋㅋㅋ 캐릭터 귀요미들
['▁빌리', '▁귀엽다', '잉', '▁ㅋㅋㅋㅋㅋㅋㅋㅋㅋ', '▁캐릭터', '▁귀요미', '들']
[11872, 5299, 19098, 8472, 542, 12261, 18307]

허큘리스 개재밌음 난 학교에서 봤는데 왜 이걸 극장에서 못 봤는지 의문임 그리고 평점은 또 왜 이렇게 낮음ㅋㅋ
['▁허', '큘', '리스', '▁개', '재밌음', '▁난', '▁학교에서', '▁봤는데', '▁왜', '▁이걸', '▁극장에서', '▁못', '▁봤는지', '▁의문
```

```
', '임', '_그리고', '_평점은', '_또', '_왜', '_이렇게', '_낮음', 'ㅋㅋ']
[265, 19777, 693, 74, 3464, 205, 3659, 191, 84, 594, 1095, 89, 7912, 2971, 18440, 319, 1822, 257, 84, 246, 11136,
12]
```

In [5]:

```python
from konlpy.tag import Okt
from collections import Counter
okt = Okt()

#한번 나온 단어가 고유단어라고 생각하여 풀었습니다.
#또한 morphs를 쓴 이유는 사전 의미상으로
#단어는 명사뿐만이 아닌 다른 것도 가능하다고 나와있어
#nouns를 쓰지 않았습니다.

F = open('naver_review.txt', 'r', encoding='utf-8')
data = F.read()
words = okt.morphs(data)

vocab = Counter(words)
final_result = []
tu_voca = list(vocab.items())

#print(len(vocab))
#print(len(tu_voca))

for i in tu_voca:
    if i[1] == 1:
        final_result.append(i)

final_result = sorted(sorted(final_result), key = lambda x: x[1], reverse = True)

print("고유단어수: ", len(final_result))
```

```
고유단어수:  66854
```

In [6]:

```python
lines = [
    "난 재밌던데?....",
    "귀엽고 멋지고 재미있는. 매력 덩어리",
    "이방인과 현지인, 그들이 하나가 되는 순간의 코인로커",
    "♥",
    "월드컵 기간에 보기엔 딱 좋은 영화! 이 영화만 100번 넘게 볼 정도로 강력 추천!",
    "5점준놈 생각좀해 78년작인데 ㅡㅡ 멀더바래",
    "잘 찍었네요..하지만 흥행과는 무관할 듯",
    "이게 왜 8점대야 최소 9점대는 돼야지",
    "빌리 귀엽다잉 ㅋㅋㅋㅋㅋㅋㅋㅋ 캐릭터 귀요미들",
    "허큘리스 개재밌음 난 학교에서 봤는데 왜 이걸 극장에서 못 봤는지 의문임 그리고 평점은 또 왜 이렇게 낮음ㅋㅋ",
]

str_lines = str(lines)
words2 = okt.morphs(str_lines)
vocab2 = Counter(words2)

print(vocab2)
```

```
Counter({"'": 9, "',": 7, '이': 4, '왜': 3, '난': 2, '들': 2, '영화': 2, '점': 2, '대': 2, '에서': 2, "['": 1, '재밌
던데': 1, "?....',": 1, '귀엽고': 1, '멋지고': 1, '재미있는': 1, '.': 1, '매력': 1, '덩어리': 1, '이방인': 1, '과': 1, '
현지': 1, '인': 1, ',': 1, '그': 1, '하나': 1, '가': 1, '되는': 1, '순간': 1, '의': 1, '코인': 1, '로커': 1, '♥': 1, '
월드컵': 1, '기간': 1, '에': 1, '보기': 1, '엔': 1, '딱': 1, '좋은': 1, '!': 1, '만': 1, '100': 1, '번': 1, '넘게': 1,
'볼': 1, '정도': 1, '로': 1, '강력': 1, '추천': 1, "!',": 1, '5': 1, '점준놈': 1, '생각': 1, '좀해': 1, '78년': 1, '작
인데': 1, 'ㅡㅡ': 1, '멀더': 1, '바': 1, '래': 1, '잘': 1, '찍었네요': 1, '..': 1, '하지만': 1, '흥행': 1, '과는': 1, '
무': 1, '관할': 1, '듯': 1, '게': 1, '8': 1, '야': 1, '최소': 1, '9': 1, '는': 1, '돼야지': 1, '빌리': 1, '귀엽다': 1,
'잉': 1, 'ㅋㅋㅋㅋㅋㅋㅋㅋ': 1, '캐릭터': 1, '귀요미': 1, '허큘리스': 1, '개': 1, '재밌음': 1, '학교': 1, '봤는데': 1, '
걸': 1, '극장': 1, '못': 1, '봤는지': 1, '의문': 1, '임': 1, '그리고': 1, '평점': 1, '은': 1, '또': 1, '이렇게': 1, '낮
음': 1, 'ㅋㅋ': 1, "']": 1})
```