

1. 소나기 분석(불용어 처리X)

```
In [2]: #소나기

from konlpy.tag import Kkma
from collections import Counter

infile = open("소나기.txt", encoding='utf-8')
data=infile.read()

kkma = Kkma()
words=kkma.morphs(data)

vocab=Counter(words)

#딕셔너리의 items -> 튜플로 반환을 해주고 그것을 다시 리스트로 반환함. value값을 잘 활용하기 위해서.
final_result = []
tu_voca = list(vocab.items())
#print(tu_voca)

#튜플이 들어간 리스트에서 value값이 5보다 크거나 같으면 final_result에 해당 튜플을 다시 넣어줌.
for i in tu_voca:
    if i[1] >= 5:
        final_result.append(i)

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

print(final_result)

[(('.', 311), ('다', 226), ('엇', 168), ('이', 137), ('', 119), ('"', 106), ('을', 106), ('가', 96), ('어', 94), ('는', 86), ('ㄴ', 78), ('하', 77), ('소녀', 72), ('은', 71), ('에', 70), ('고', 54), ('를', 54), ('있', 54), ('소년', 49), ('아', 45), ('ㄴ다', 43), ('았', 41), ('지', 38), ('것', 35), ('의', 35), ('도', 32), ('으', 27), ('게', 26), ('나', 25), ('보', 25), ('그러', 23), ('않', 22), ('한', 22), ('르', 20), ('에서', 20), ('더', 17), ('?', 16), ('갈', 16), ('로', 16), ('기', 15), ('며', 15), ('속', 15), ('그', 14), ('꽃', 14), ('날', 14), ('저', 14), ('생각', 13), ('서', 13), ('"', 12), ('만', 12), ('말', 12), ('물', 12), ('오', 12), ('손', 11), ('네', 10), ('튼', 10), ('았', 10), ('니', 9), ('모르', 9), ('어서', 9), ('전', 9), ('개울가', 8), ('고는', 8), ('길', 8), ('내', 8), ('달리', 8), ('수', 8), ('안', 8), ('알', 8), ('얼굴', 8), ('없', 8), ('이것', 8), ('참', 8), ('하나', 8), ('갈꽃', 7), ('그리', 7), ('데', 7), ('돌아오', 7), ('들', 7), ('바보', 7), ('벼', 7), ('비', 7), ('이번', 7), ('집', 7), ('개울', 6), ('곳', 6), ('까지', 6), ('나오', 6), ('눈', 6), ('는지', 6), ('다가', 6), ('더니', 6), ('두', 6), ('듯이', 6), ('머리', 6), ('못', 6), ('밀', 6), ('산', 6), ('소리', 6), ('송아지', 6), ('이사', 6), ('일', 6), ('자기', 6), ('잡', 6), ('짐검다리', 6), ('허수아비', 6), ('호두', 6), ('치', 5), (')', 5), ('ㄴ다는', 5), ('ㅁ', 5), ('가을', 5), ('가지', 5), ('갈밭', 5), ('건너', 5), ('그렇', 5), ('느냐', 5), ('다고', 5), ('달려가', 5), ('대', 5), ('덜', 5), ('라도', 5), ('덜', 5), ('면', 5), ('버리', 5), ('번', 5), ('보이', 5), ('뿐', 5), ('수숫단', 5), ('아버지', 5), ('없이', 5), ('오늘', 5), ('우리', 5), ('옴키', 5), ('윤', 5), ('조약돌', 5), ('주', 5), ('주머니', 5), ('쪽', 5), ('초시', 5), ('크', 5)]
```

2. 메밀꽃 분석(불용어 처리X)

```
In [3]: #메밀꽃

from konlpy.tag import Kkma
from collections import Counter

infile = open("메밀꽃.txt", encoding='utf-8')
data=infile.read()

kkma = Kkma()
words=kkma.morphs(data)

vocab=Counter(words)

#딕셔너리의 items -> 튜플로 반환을 해주고 그것을 다시 리스트로 반환함. value값을 잘 활용하기 위해서.
final_result = []
tu_voca = list(vocab.items())
#print(tu_voca)

#튜플이 들어간 리스트에서 value값이 5보다 크거나 같으면 final_result에 해당 튜플을 다시 넣어줌.
for i in tu_voca:
    if i[1] >= 5:
        final_result.append(i)

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

print(final_result)

[(('.', 233), ('이', 169), ('엇', 140), ('다', 132), ('는', 128), ('"', 112), ('을', 111), ('하', 110), ('ㄴ', 107), ('에', 102), ('은', 96), ('고', 89), ('어', 65), ('도', 60), ('지', 60), ('가', 50), ('를', 44), ('', 40), ('르', 39), ('의', 38), ('았', 38), ('들', 34), ('게', 32), ('것', 31), ('나', 27), ('아', 27), ('생원', 25), ('않', 25), ('으냐', 25), ('았', 23), ('없', 23), ('기', 20), ('수', 20), ('허', 20), ('둥이', 19), ('되', 19), ('보', 19), ('고', 18), ('어서', 17), ('?', 16), ('때', 16), ('야', 16), ('나귀', 15), ('로', 15), ('생각', 15), ('에서', 15), ('만', 14), ('꽃', 13), ('네', 13), ('두', 13), ('물', 13), ('장', 13), ('계집', 12), ('과', 12), ('말', 12), ('면', 12), ('였', 12), ('이나', 12), ('젼', 11), ('며', 11), ('.....', 10), ('같이', 10), ('니', 10), ('달', 10), ('더', 10), ('밤', 10), ('시작', 10), ('으로', 10), ('짐승', 10), ('ㄴ다', 9), ('그렇', 9), ('까지', 9), ('녀석', 9), ('뽕', 9), ('서', 9), ('소리', 9), ('조선', 9), ('갈', 8), ('걸', 8), ('튀', 8), ('스럼', 8), ('여야', 8), ('구', 7), ('벌써', 7), ('부터', 7), ('사람', 7), ('세', 7), ('속', 7), ('아이', 7), ('이야기', 7), ('일', 7), ('젼', 7), ('제천', 7), ('지금', 7), ('집', 7), ('ㄴ지', 6), ('길', 6), ('내', 6), ('눈', 6), ('다고', 6), ('라고', 6), ('러', 6), ('버리', 6), ('요', 6), ('울', 6), ('장판', 6), ('젼', 6), ('쵸', 6), ('처녀', 6), ('치', 6), ('평', 6), ('하나', 6), ('군', 5), ('그것', 5), ('까', 5), ('눔', 5), ('달이', 5), ('대화', 5), ('듯이', 5), ('란', 5), ('마음', 5), ('면서', 5), ('아니', 5), ('알', 5), ('앞', 5), ('어머니', 5), ('오', 5), ('으면', 5), ('을까', 5), ('이란', 5), ('저', 5), ('차', 5), ('축', 5), ('충춇집', 5)]
```

3. 소나기 분석(불용어 제거O)

```
In [3]: #불용어 제거

from nltk.corpus import stopwords
from konlpy.tag import Kkma
from collections import Counter

#불용어 리스트
stop_words = """"? ..... -- ", ( ) ' - 이 었 었 은 다 을 를 가 에 에서 로 와 과 보다 아 야 랑 며 나 는 만 밖에 뿐 도 조차 마저 까지 든지 나마 라도 의 한테 께서 있 하 것 들 그 도
infile = open("소나기.txt", encoding='utf-8')
data=infile.read()

#형태소 분리 및 불용어 리스트에서 공백을 기준으로 분리하여 리스트로 반환
kkma = Kkma()
words = kkma.morphs(data)
stop_words = stop_words.split(' ')

#형태소로 분리된 것이 불용어 리스트에 없다면 result 리스트에 넣어줌.
result = []
for i in words:
    if i not in stop_words:
        result.append(i)

#Counter로 몇 개가 나왔는지 세줌.
voca = Counter(result)
#print(voca)

#딕셔너리의 items -> 튜플로 반환을 해주고 그것을 다시 리스트로 반환함. value값을 잘 활용하기 위해서.
final_result = []
tu_voca = list(voca.items())
#print(tu_voca)

#튜플이 들어간 리스트에서 value값이 5보다 크거나 같으면 final_result에 해당 튜플을 다시 넣어줌.
for i in tu_voca:
    if i[1] >= 5:
        final_result.append(i)

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

print(final_result)

[('어', 94), ('ㄴ', 78), ('소녀', 72), ('고', 54), ('소년', 49), ('ㄴ다', 43), ('았', 41), ('지', 38), ('으로', 27), ('게', 26), ('그러', 23), ('한', 22), ('르', 20), ('더', 17), ('갈', 16), ('기', 15), ('속', 15), ('꽃', 14), ('날', 14), ('저', 14), ('생각', 13), ('서', 13), ('말', 12), ('물', 12), ('오', 12), ('손', 11), ('네', 10), ('튼', 10), ('았', 10), ('니', 9), ('모르', 9), ('어서', 9), ('전', 9), ('개울가', 8), ('고는', 8), ('길', 8), ('내', 8), ('달리', 8), ('안', 8), ('알', 8), ('얼굴', 8), ('이것', 8), ('참', 8), ('하나', 8), ('갈꽃', 7), ('그리', 7), ('데', 7), ('돌아오', 7), ('벼', 7), ('비', 7), ('이번', 7), ('집', 7), ('개울', 6), ('곳', 6), ('나오', 6), ('눈', 6), ('는지', 6), ('다가', 6), ('더니', 6), ('두', 6), ('듯이', 6), ('머리', 6), ('못', 6), ('밀', 6), ('산', 6), ('소리', 6), ('송아지', 6), ('이사', 6), ('일', 6), ('자기', 6), ('잡', 6), ('짐검다리', 6), ('허수아비', 6), ('호두', 6), ('ㄴ다는', 5), ('ㅁ', 5), ('가을', 5), ('가지', 5), ('갈밭', 5), ('건너', 5), ('그렇', 5), ('느냐', 5), ('다고', 5), ('달려가', 5), ('대', 5), ('덜', 5), ('면', 5), ('버리', 5), ('번', 5), ('보이', 5), ('수숫단', 5), ('아버지', 5), ('없이', 5), ('오늘', 5), ('우리', 5), ('옴키', 5), ('윤', 5), ('조약돌', 5), ('주', 5), ('주머니', 5), ('쪽', 5), ('초시', 5), ('크', 5)]
```

4 메밀꽃 분석(불용어 제거O)

```
In [4]: #불용어 제거

from nltk.corpus import stopwords
from konlpy.tag import Kkma
from collections import Counter

#불용어 리스트
stop_words = """"? ..... -- ", ( ) ' - 이 은 었 었 다 을 를 가 에 에서 로 와 과 보다 아 야 랑 며 나 는 만 밖에 뿐 도 조차 마저 까지 든지 나마 라도 의 한테 께서 있 하 것 들 그 도
infile = open("메밀꽃.txt", encoding='utf-8')
data=infile.read()

#형태소 분리 및 불용어 리스트에서 공백을 기준으로 분리하여 리스트로 반환
kkma = Kkma()
words = kkma.morphs(data)
stop_words = stop_words.split(' ')

#형태소로 분리된 것이 불용어 리스트에 없다면 result 리스트에 넣어줌.
result = []
for i in words:
    if i not in stop_words:
        result.append(i)

#Counter로 몇 개가 나왔는지 세줌.
voca = Counter(result)
#print(voca)

#딕셔너리의 items -> 튜플로 반환을 해주고 그것을 다시 리스트로 반환함. value값을 잘 활용하기 위해서.
final_result = []
tu_voca = list(voca.items())
#print(tu_voca)

#튜플이 들어간 리스트에서 value값이 5보다 크거나 같으면 final_result에 해당 튜플을 다시 넣어줌.
for i in tu_voca:
    if i[1] >= 5:
        final_result.append(i)

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

print(final_result)

[('ㄴ', 107), ('고', 89), ('어', 65), ('지', 60), ('르', 39), ('게', 32), ('생원', 25), ('으냐', 25), ('았', 23), ('기', 20), ('하', 20), ('둥이', 19), ('어서', 17), ('때', 16), ('나귀', 15), ('생각', 15), ('꽃', 13), ('네', 13), ('두', 13), ('물', 13), ('장', 13), ('계집', 12), ('말', 12), ('면', 12), ('이나', 12), ('젼', 11), ('같이', 10), ('니', 10), ('달', 10), ('더', 10), ('밤', 10), ('시작', 10), ('으로', 10), ('짐승', 10), ('ㄴ다', 9), ('그렇', 9), ('녀석', 9), ('뽕', 9), ('서', 9), ('소리', 9), ('조선', 9), ('갈', 8), ('걸', 8), ('튀', 8), ('스럼', 8), ('여야', 8), ('구', 7), ('벌써', 7), ('부터', 7), ('사람', 7), ('세', 7), ('속', 7), ('아이', 7), ('이야기', 7), ('일', 7), ('젼', 7), ('ㄴ지', 6), ('집', 7), ('내', 6), ('눈', 6), ('다고', 6), ('라고', 6), ('러', 6), ('버리', 6), ('요', 6), ('울', 6), ('장판', 6), ('젼', 6), ('쵸', 6), ('처녀', 6), ('치', 6), ('평', 6), ('하나', 6), ('군', 5), ('그것', 5), ('까', 5), ('눔', 5), ('달이', 5), ('대화', 5), ('듯이', 5), ('란', 5), ('마음', 5), ('면서', 5), ('아니', 5), ('알', 5), ('앞', 5), ('어머니', 5), ('오', 5), ('으면', 5), ('을까', 5), ('이란', 5), ('저', 5), ('차', 5), ('축', 5), ('충춇집', 5)]
```