자연어처리 – 과제 #1

202204495 홍창희

1. 7쪽의 프로그램을 수정하여 '삼포 가는 길.txt' 파일을 분석함. 이 문서에서 나온 횟수가 1번 이상인 명사 단어들을 출력함. 화면에 나온 데이터를 복사해서 제출해도 됨. 1번 이상 나타난 명사의 숫자를 구함.

코드

```
In [1]: from konlpy.tag import Kkma

In [2]: from collections import Counter

In [3]: infile = open("삼포 가는 길.txt", encoding='utf-8')

In [4]: data = infile.read()

In [5]: kkma = Kkma()

In [6]: words = kkma.nouns(data)

In [7]: vocab = Counter(words)

In [8]: print(vocab)

In [9]: print('1번 이상 나타난 명사의 숫자:', len(vocab))
```

출력

```
Counter({'내': 3, '만': 3, '리': 3, '구': 3, '여': 2, '달': 2, '때': 2, '수': 2, '천': 2, '개': 2, '야': 2, '외': 2, '시': 2, '적': 2, '줄': 2, '도': 2, '참': 2, '일': 2, '원': 2, '육': 2, '우리': 2, '오': 2, '오만': 2, '간': 2, '하나': 2, '대': 2, '등': 2, '코우트': 2, '이래': 2, '여덟': 2, '명': 2, '영달': 1, '어디': 1, '궁리': 1, '새벽': 1, '겨울': 1, '바람': 1, '아침': 1, '햇볕': 1, '아래': 1, '들판': 1, '곳곳': 1, '시냇물': 1, '웅덩이': 1, '반사': 1, '빛': 1, '소리': 1, '데': 1, '그': 1, '창공': 1, '남': 1, '나무': 1, '수십': 1, '수십여': 1, '그루': 1, '들판가': 1, '전': 1, '이곳': 1, '한참': 1, '추수기': 1, '공사': 1, '막판': 1, '봄': 1, '연기': 1, '터': 1, '진작': 1, '예상': 1, '현장': 1, '사무소': 1, '사흘': 1, '문': 1, '영달이': 1, '밥집': 1, '기회': 1, '누군가': 1, '밭고랑': 1, '걸어오고': 1, '음지': 1, '양지': 1, '구분': 1, '언덕': 1, '그림자': 1, '숲': 1, '그늘': 1, '곳': 1, '흙': 1, '해': 1, '시작': 1, '사람': 1, '신발': 1, '끝': 1, '진흙': 1, '뭉치': 1, '뒤': 1, '점': 1, '길가': 1, '담배': 1, '쪽': 1, '키': 1, '맹꽁이': 1, '배낭': 1, '어깨': 1, '히': 1, '머리': 1, '개털': 1, '모자': 1, '귀': 1, '야전': 1, '잠바': 1, '깃': 1, '속': 1, '턱': 1, '반': 1, '누': 1, '누군지': 1, '군지': 1, '쌍': 1, '쌍통': 1, '통': 1, '도리': 1, '걸음': 1, '털모자': 1, '챙': 1, '이마빡': 1, '천씨네': 1, '씨네': 1, '집': 1, '양반': 1, '낯': 1, '서른': 1, '댓': 1, '사내': 1, '공사장': 1, '마을': 1, '어귀': 1, '주막': 1, '얼굴': 1, '아까': 1, '존': 1, '구경': 1, '했시': 1, '단추': 1, '나': 1, '비': 1, '비행사': 1, '행사': 1, '양쪽': 1, '뺨': 1, '귀가리개': 1, '가리개': 1,. . .(생략). . . , '천지': 1,'공사판': 1, '사람들': 1, '나룻배': 1, '변': 1, '변고지': 1, '고지': 1, '풍문': 1, '발걸음': 1, '정처': 1, '결': 1, '입장': 1})

1 번 이상 나타난 명사의 숫자: 1164
```

2. 토지 1~2 권 파일에 나타난 단어 숫자를 두 가지 방식으로 계산

a. 어절 단위로 나누었을 때 나타나는 서로 다른 어절의 숫자를 구함.

코드

```
In [1]: from konlpy.tag import Kkma

In [2]: from collections import Counter

In [3]: from nltk import word_tokenize

In [4]: infile = open("토지1.txt", encoding='utf-8')

In [5]: data = infile.read()

In [6]: kkma = Kkma()

In [7]: words = word_tokenize(data)

In [8]: vocab = Counter(words)

In [9]: print(vocab)

In [10]: print('서로 다른 어절의 숫자:', len(vocab))
```

출력

```
Counter({'.': 8282, '``': 3626, "''": 3602, ',': 2135, '?': 1228, '!': 712, '그': 506, '안': 406, '...': 383, '있었다': 343, '한': 278, '것': 264, '다': 260, '이': 213, '수': 196, "'": 193, '있는': 188, '것이다': 187, '못': 184, '하고': 176, '같은': 172, '용이는': 168, '와': 165, '없는': 152, '내': 136, '내가': 134, '어디': 133, '평산은': 132, '말을': 124, '그러나': 124, '할': 123, '그는': 121, '했다': 119, '하는': 111, '두': 110, '그런': 109, '다시': 109, '일이': 106, '때': 103, '것을': 103, '더': 101, '좀': 100, '것도': 100, '그리': 94, '용이': 94, '거': 93, '가서': 92, '제': 91, '하며': 91, '말이': 91, '무슨': 90, '없이': 90, '아': 89, '눈을': 89, '말': 88, '머': 87, '는': 85, '그래': 85, '또': 85, '없었다': 84, '강청댁은': 84, '지': 83, '말했다': 82, '없다': 81, '얼굴을': 81, '나': 81, '눈이': 80, '듯': 79, '고': 78, '서': 75, '예': 75, '하나': 74, '않았다': 74, '가': 74, '누가': 73, '줄': 72, '것은': 72, '치수는': 72, '일': 71, '사람': 71, '될': 70, '나는': 70, '보고': 69, '니': 69, '소리가': 67, '있다': 66, '우리': 66, '을': 65, '긴데': 64, '야': 63, '준구는': 63, '무신': 62, '뒤': 61, '한다': 60, '있던': 60, '물었다': 60, '어느': 59, '없고': 59, '참': 58, '것이': 57, '그의': 57, '우찌': 57, '않고': 57, '기이': 56, '소리를': 56, '들고': 56, '살': 54, '서희는': 53, '게': 53, '흥': 53, '그럴': 52, '날': 52, '웃는다': 52, '역시': 52, '같았다': 52, '칠성이는': 52, '왜': 51, '아니': 50, '그라믄': 50, '평산이': 50, '있을': 49, '기요': 49, '마을': 49, '에': 49, '평산의': 49, '집에': 48, '봉순네는': 48, '죽은': 48, '따라': 47, '칠성이': 47, '채': 47, '눈에': 47, '보였다': 47, '나도': 47, '소리': 46, '함께': 46, '많이': 46, '아니가': 46, '생각이': 46, '강포수는': 46, '하지': 45, '월선이는': 45, '문의원은': 45, '것처럼': 44, '가는': 44, '년': 44, '몇': 44, '임이네는': 44, '와서': 43, '때문에': 43, '된': 43, '알고': 43, '얼굴이': 43, '온': 43, '길상은': 43, '일을': 42, '하는데': 42, '사램이': 42, '큰': 41, '귀녀는': 41, '향해': 41, '몸을': 41, '앉아': 41, '잘': 41, '묵고': 41, '되는': 41, '허': 41, '기다': 40, '나서': 40, '손을': 40, '사람이': 40, . . .(생략). . . , '하나면': 1, '그만이라는': 1, '죽겠다는': 1, '으름장이다': 1, '입이네가': 1, '침소봉대해서': 1, '곧이들은': 1, '쫓겨날': 1, '2 권에서': 1})

서로 다른 어절의 숫자: 30480
```

b. 문장들에 대해 형태소 분석기를 적용하여 단어들을 분리한 다음, a와 같은 방식으로 서로 다른 단어의 갯수를 구함.

코드

```
In [1]: from konlpy.tag import Kkma

In [2]: from collections import Counter

In [3]: infile = open("토지2.txt", encoding='utf-8')

In [4]: data = infile.read()

In [5]: kkma = Kkma()

In [6]: words = kkma.morphs(data)

In [7]: vocab = Counter(words)

In [8]: print(vocab)

In [9]: print('서로 다른 형태소의 숫자:', len(vocab))
```

출력

```
Counter({'.': 8736, '이': 7922, '"': 6025, '는': 5300, '을': 5134, '다': 5095, 'ㄴ': 4746,
'었': 4640, '하': 4623, '고': 3941, '은': 2988, '에': 2973, '어': 2950, '아': 2335, '가': 2245,
',': 2069, '를': 1987, 'ㄹ': 1934, '의': 1790, '것': 1604, '그': 1507, '도': 1501, '있': 1497,
'지': 1435, '나': 1184, '?': 1167, 'ㄴ다': 1112, '았': 1085, '들': 1031, '말': 876, '게': 862,
'오': 845, '더': 816, '없': 777, '!': 748, '에서': 730, '보': 694, '며': 691, '네': 640, '으로':
621, '되': 609, '기': 598, '않': 591, '로': 566, '늘': 565, '만': 555, '니': 552, '일': 552,
'어서': 546, '알': 542, '소': 499, '서': 499, '사람': 473, '라': 472, '그러': 460, '안': 435,
'믄': 433, '아니': 418, "'": 402, '내': 401, '김': 392, '두': 384, '같': 376, '수': 370, '아서':
358, '씨': 346, '눈': 345, '강': 344, '한': 343, '얼굴': 342, '댁': 341, '주': 332, '포수': 329,
'생각': 325, '는데': 317, '윤': 311, '죽': 310, '시': 305, '용': 298, '때': 294, '서방': 281,
'길': 278, '에게': 276, '소리': 264, '면': 262, '봉': 260, '순': 259, '저': 257, '마': 255, '임':
253, '요': 251, '겠': 244, '구': 244, '놓': 241, '과': 240, '살': 237, '듣': 228, 'ㄴ데': 225,
'여': 225, '집': 224, '아이': 220, '자': 216, '으나': 210, '같이': 209, '모르': 209, '못하': 208,
'으며': 207, '치': 206, '께': 205, '그렇': 204, '다가': 196, '못': 193, '스': 189, '면서': 187,
'부인': 186, '듯': 185, '야': 183, '머': 178, '와': 177, '우': 171, '속': 170, '다는': 167, '귀':
167, '놈': 166, '거': 165, '카': 165, '웃': 164, '사': 164, '마을': 164, '이네': 164, '치수':
162, '또': 162, '“': 162, '”': 161, '그것': 160, '어디': 159, '방': 158, '귀녀': 157, '다시':
157, '없이': 155, '입': 154, '물': 153, '까지': 153, '남': 151, '그리하': 151, '최': 150, '줄':
150, 'ㄴ가': 149, '좋': 147, '으': 146, '손': 145, '갈': 144, '묵': 144, '몸': 144, '울': 144,
'장': 143, '나오': 143, '길상': 143, '앉': 142, '러': 141, '잡': 140, 'ㅁ': 138, '말하': 137,
'팔': 137, '까': 136, '리': 134, '겄': 134, '날': 134, '문': 133, '데': 133, '최치수': 131, '뒤':
129, '산': 129, '보이': 128, '하나': 128, '앞': 126, '이나': 125, '가지': 125, '...': 124, '노':
124, '허': 124, '자신': 122, '강청': 122, '수동이': 121, '돌': 119, '평산': 119, 'ㄹ지': 118,
'여자': 118, '였': 117, '녀': 117, '영': 116, '년': 115, '칠성': 114, '좀': 113, '마음': 111,
'보다': 111, '겉': 110, '때문': 110, . . .(생략). . . , '모하고': 1, '당했이믄서': 1, '부끄러움': 1,
'선발을': 1, '집었': 1, '토란': 1, '두근두근': 1, '이처럼': 1, '벌받': 1, '가꺼': 1, '찮겄': 1})

서로 다른 형태소의 숫자: 11564
```