[숙제06] 파일 처리 연습

언어와 컴퓨터

2018년 10월 16일 화요일 13시까지

- 소스코드 스크립트 `count.py`, 결과물 `counts_00000.pkl` 및 보고서 `hw06_00000.pdf` 파일을 `hw06_00000.zip` 파일로 압축하여 제출하라.
- 데이터 `proper_nouns.txt`가 소스코드 `count.py`와 같은 폴더 안에 있다고 가정하라.
- `count.py` 파일은 과제와 함께 업로드된 것을 수정하여 사용하라.
- 보고서는 문제 해결 방법, 코드 설명, 테스트 실행 결과 등을 포함하여 작성하라.

1 현대 국어 사용 빈도 조사 결과 처리하기

국립국어원에서 2002년에 발표한 현대 국어 사용 빈도 조사 결과 파일[1]을 파이썬에서 읽고 처리하려고 한다. 과제에 첨부된 `proper_nouns.txt`는 이 파일의 첫 행을 삭제한 것으로, 다음과 같은 형식으로 이루어져 있다(`\t`는 탭, `\n`는 줄바꿈 문자):

```
'차례\t항목\t빈도\t개수\t교재\t교과\t교양\t문학\t신문\t잡지\t대본\t구어\t기타\n'
```

예를 들어 `'161\t경복궁\t50\t10\t1\t16\t19\t9\t2\t3\t0\t0\t0\n'`은 '경복궁'이라는 161번째 단어가 자료에 나타난 총 빈도가 50건이고 개수는 10개이며, 장르별로 교재 1건, 교과 16건, 교양 19건, 문학 9건, 신문 2건, 잡지 3건, 대본, 구어, 기타에는 0건 출현했다는 의미이다.

이러한 형식의 파일을 처리하여 아래와 같은 구조의 딕셔너리를 얻을 수 있도록 코드를 작성하라.

```
>>> pprint(counts)
{
    ...,
    '경복궁': {'빈도': 50, '개수': 10, '교재': 1, ..., '기타': 0},
    '경부': {'빈도': 2, '개수': 2, '교재': 0, ..., '기타': 0},
    ...
}
```

1.1 `numeral(string)` (3점)

`'3,456'`이나 `'$789'` 등을 인자로 받아 대응하는 정수 값을 반환하는 함수. 숫자가 포함되지 않은 문자열을 받으면 아무 것도 반환하지 않으며, 이 문제에서는 인자에 소수점이 포함된 경우가 없다고 가정한다.

대화형 모드에서 아래와 같이 작동하면 된다.

```
>>> numeral('12')
12
>>> numeral('3,456')
3456
>>> numeral('$789')
789
>>> numeral('')
>>> numeral('??')
```

[1] http://www.korean.go.kr/front/etcData/etcDataView.do?mn_id=46&etc_seq=61

1.2 count_by_genre(filename) (4점)

현대 국어 사용 빈도 조사 결과 파일 이름을 인자로 받아 앞에서 제시된 구조의 딕셔너리를 반환하는 함수.

힌트 zip() 함수를 활용하여 딕셔너리를 만드는 방법[2]을 참조하면 좋다.

이 함수에서는 반환되는 딕셔너리의 크기가 크므로, 함수 호출 결과를 대화형 모드에서 직접 확인하기는 어렵다. 아래와 같은 테스트 코드로 결과를 확인해 보자.

테스트 코드

```
if __name__ == '__main__':
    filename = 'proper_nouns.txt'
    counts = count_by_genre(filename)
    print('"가르시아"의 장르별 빈도 딕셔너리:: ', counts['가르시아'])
    print('"고구려"의 교재 출현 빈도:: ', counts['고구려']['교재'])
    print()
```

결과

```
"가르시아"의 장르별 빈도 딕셔너리:: {'빈도': 184, '개수': 1, '교재': 0, '교과': 0,
'교양': 0, '문학': 184, '신문': 0, '잡지': 0, '대본': 0, '구어': 0, '기타': 0}
"고구려"의 교재 출현 빈도:: 1
```

1.3 출력 및 저장 (4점)

위의 테스트 코드에서 저장한 딕셔너리 counts에서 잡지에 50회 넘게 출현한 단어만 모아서 잡지 출현 빈도와 함께 출력하라. 화면에 아래와 같은 결과가 나오면 된다.

```
김 358
김영삼 76
김정원 96
노 58
독일 59
미국 202
북한 96
서울 161
여의주 86
이 112
일본 148
중국 73
청와대 117
평화의댐 53
한국 243
```

마지막으로 pickle 모듈을 사용하여 딕셔너리 counts를 counts_00000.pkl 파일로 저장하라.

1.4 보너스 (추가 +1점)

numeral(string) 함수에서 소수점이 포함된 경우를 무시하기 싫다면, 인자에 소수점이 포함된 경우 부동소수점 값이 반환되도록 함수를 수정하라.

[2] https://dojang.io/mod/page/view.php?id=1003