[숙제10심화] 트위터 사용자 데이터 분석

언어와 컴퓨터

2018년 11월 23일 금요일 15시까지

- 소스코드 스크립트 `hungry.py` 및 보고서 `hw10adv_00000.pdf` 파일을 `hw10adv_00000.zip` 파일로 압축하여 제출하라.
- 보고서는 문제 해결 방법, 코드 설명, 테스트 실행 결과 등을 포함하여 작성하라.

1 트위터 사용자 데이터 분석

첨부된 데이터 파일 `tweets.txt`는 2018년 11월 11일에 수집된 한국어 트위터 텍스트의 일부이다.[1] 이 데이터를 통해 사람들이 무엇을 먹고 싶어하는지를 알아보자.

1.1 프로그램 작성

데이터 파일을 읽은 뒤 "◯◯ 먹고 싶어" 꼴의 표현에서 ◯◯에 해당하는 것을 모두 찾아서 빈도를 계산하라.

힌트 `help(re)` 도움말에서 아래의 부분을 참조하면 좋다.

```
(?=...)  Matches if ... matches next, but doesn't consume the string.
(?!...)  Matches if ... doesn't match next.
(?<=...) Matches if preceded by ... (must be fixed length).
(?<!...) Matches if not preceded by ... (must be fixed length).
```

빈도 계산 결과를 `collections.Counter` 자료형으로 저장하면, `Counters.most_common()` 메서드로 상위 n 개를 볼 수 있다. 결과가 아래의 예시와 일치할 필요는 없다.

```
>>> type(food)
<class 'collections.Counter'>
>>> food.most_common(50)
[('너무', 43), ('빼빼로', 27), ('뭐', 24), ('거', 18), ('안', 17), ('떡볶이',
17), ('저도', 16), ('또', 16), ('더', 15), ('피자', 15), ('치킨', 13), ('라면',
12), ('같이', 12), ('췌장을', 12), ('밥', 12), ('게', 10), ('고기', 10), ('귤',
9), ('뭐가', 8), ('떡볶이는', 8), ('아이스크림', 8), ('나도', 8), ('내가', 7),
('닭발', 7), ('넘', 7), ('그거', 6), ('간식', 6), ('따', 6), ('불닭', 6), ('다',
5), ('저거', 5), ('과자', 5), ('빵', 5), ('맛있는거', 5), ('갑자기', 5),
('케이크가', 5), ('존나', 5), ('딸기', 5), ('사', 4), ('술', 4), ('써', 4),
('욕', 4), ('신전', 4), ('카레', 4), ('초콜릿', 4), ('마카롱', 4), ('카레가', 4),
('걸', 4), ('도넛', 4), ('초밥', 4)]
```

1.2 결과 평가 및 분석

위의 결과에서 무엇을 알 수 있는지, 어떤 점을 어떻게 개선하면 좋을지를 자유롭게 서술하라.

참조

배고파 고로케(–2014, `say.coroke.net/hungry`)[2]

[1] ID 등 개인을 식별할 수 있는 정보는 삭제하였다.

[2] 현재는 서비스가 종료되었으나 몇몇 웹 페이지에서 흔적을 찾을 수 있다.

- `https://www.slideshare.net/cbs15min/ss-37519125`
- `http://www.nocutnews.co.kr/news/4072530`