NLP 개인 프로젝트 ①

7. Tokenizing/ POS 태깅/ Word extraction

○ Data
- 네이버 뉴스기사 '12개 문장'

○ 형태소 분석기
- 꼬꼬마, 트위터 분석기(open-korean-text), 코모란 사용 및 비교

○ 비교 분석 항목
- 형태소 분석기별 태그셋
- 형태소 분석기별 시간측정(%%time)
- 형태소 분석기별 단어 빈도수 계산 (OOV[out of vocabulary problem]에 대한 처리능력)
- 형태소 분석기별 시간측정(%%time)
- OOV 처리를 위한 트릭 사용하기

○ Environment
- Python 3.73 with `jupyter==1.0.0`
- pandas, numpy, matplotlib, seaborn
- scikit-learn

➢ Perpose

꼬꼬마, 트위터 분석기(open-korean-text), 코모란 형태소 분석기를 사용, 비교하고 저의 상황에서 가장 사용하기에 편리한 분석기가 무엇인지 찾아보았습니다.

➢ 시간측정

```
import time

tokens = []

for name, tagger in zip(names, taggers):

    t = time.time()
    tokens.append(
        [pos for sent in sents for pos in tagger.pos(sent)]
    )
    t = time.time() - t

    print('{:8}: {:.3f} secs'.format(name, t))
```

```
kkma    : 2.806 secs
twitter : 0.440 secs
komoran : 0.128 secs
```

한 구문의 실행 시간을 측정 : %% time
→ 현재 시각을 측정하고, 이전의 시간 t 를 빼면 중간의 함수가 실행되었던 시간이 초 단위로 측정하였습니다.

→ 결과물은 tokens 안에 넣고 형태소 분석기 마다 결과를 비교 했습니다.

➢ OOV 처리를 위한 트릭 사용하기 (komoran 사용자사전 추가)

```
pprint(tokens[0][:15])
```

```
[('최', 'NNP'),
 ('순', 'NNG'),
 ('실', 'NNG'),
 ('씨', 'NNB'),
```

'최순실'이라는 이름이 꼬꼬마 형태소 분석기에서 '최', '순', '실'로 나누어 지는 문제가 발생했습니다.

이는 '최순실'이라는 명사가 학습데이터에 없었기 때문입니다. 이를 out of vocabulary라 합니다.

- **Komoran 에 사용자 사전을 추가함으로써 OOV 문제를 해결했습니다.**

```
komoran_userdic = Komoran(userdic='./userdic.txt')
komoran_userdic.pos(sent)
```