형태소분석기를 이용한 사전구축

MeCab Komoran

최연식

형태소분석기를 이용한 사전구축

INDEX

Step1

- 형태소분석기 종류 & 선택

분석 시간 & 성능 비교
Mecab, Komoran

Step2

- 분석 과정

Jupyter notebook

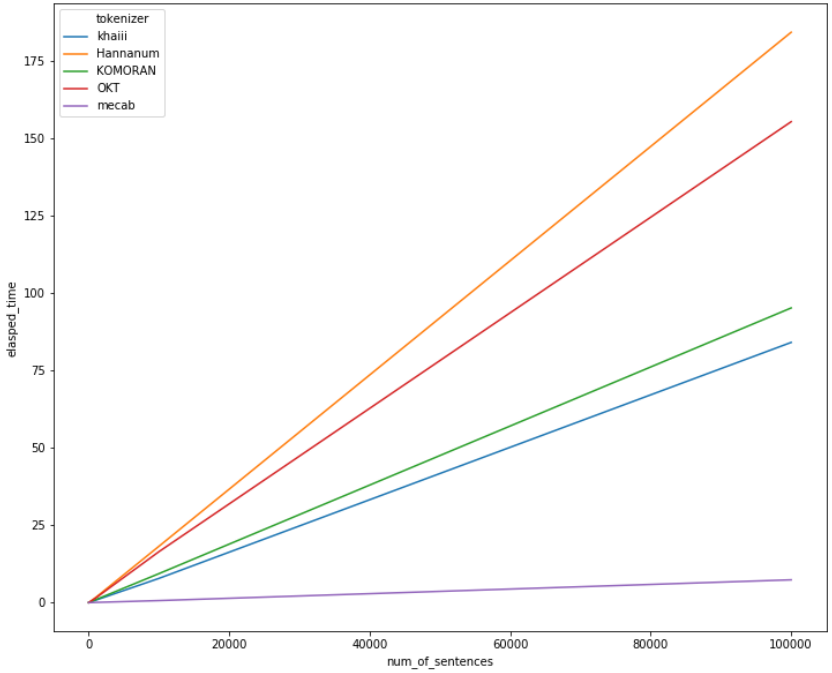
Step3

- 분석 결과

Corpus.txt

Step1

형태소분석기 종류 & 선택



실험 문장: 개봉했을때부터 지금까지 마음이답답하거나 힘들때 이영화 보고있어요 그때마다 심적인 위로를 받을수있는영화같아요 장면 하나하나가 너무예쁘고 마음에 남아서 진한 여운까지 주는영화 감사합니다

khaiii	한나눔	꼬꼬마	KOMORAN	OKT	mecab
개봉/NNG	개봉/N	개봉/NNG	개봉/NNG	개봉/Noun	개봉/NNG
하/XSV	하/X	하/XSV	하/XSV	했을/Verb	했/XSV+EP
였/EP	었을/E	었/EPT	았/EP	때/Noun	을/ETM
을/ETM	때/N	을/ETD	을/ETM	부터/Josa	때/NNG
때/NNG	부터/J	때/NNG	때/NNG	지금/Noun	부터/JX
부터/JX	지금/M	부터/JX	부터/JX	까지/Josa	지금/NNG
지금/NNG	까지/J	지금/NNG	지금/NNG	마음/Noun	까지/JX
					마음/NNG

Windows 지원 X

~~Windows 지원 X~~
Windows 가능

분석시간 : Hannanum 〉 Okt 〉 Komoran 〉 khaiii 〉 MeCab

https://iostream.tistory.com/144

Step2

분석 과정

```
import pandas as pd
data = pd.read_excel(pd.ExcelFile('과제정보(2013-2017)_미래부,과기부.xlsx'))
```

데이터 불러오기

```
df = data[['과제명', '요약문_연구목표', '요약문_연구내용', '요약문_기대효과',
           '요약문_한글키워드', '요약문_영문키워드']]
```

변수 선택

```
df.describe()
```

변수별 요약값 확인

	과제명	요약문_연구목표	요약문_연구내용	요약문_기대효과	요약문_한글키워드	요약문_영문키워드
count	10674	10328	10326	10328	10309	10267
unique	10531	10245	10256	10224	10184	10122
top	보안과제정보	보안과제정보	보안과제정보	보안과제정보	보안과제정보	보안과제정보
freq	22	22	22	22	22	22

결측값 제거

```
df = df[df.과제명 != '보안과제정보']
df = df[df.과제명 != '시설비']
df = df[df.요약문_연구목표 != '평가관리비']
df = df[df.요약문_기대효과 != '보안상 생략']
```

Jupyter notebook

Step2

분석 과정

MeCab을 활용한 사전구축

Konlpy.tag 내 Mecab 모듈 불러오기

```
from konlpy.tag import Mecab
```

collections 모듈 불러오기

```
import collections
```

Mecab 모듈 안의 하드코딩된 경로를 나의 실제 경로로 변경

```
mecab = Mecab('c:₩mecab₩mecab-ko-dic')
```

분석에 사용할 전체 데이터 셀을 하나의 문자열로 더함

```
text = ''
for i in range(df.shape[0]):
    for j in range(df.shape[1]):
        text = text + ' ' + str(df.iloc[i][j])
```

형태소분석 후, 명사만 추출

```
nouns = mecab.nouns(text)
```

명사 별 Count 후, Count값을 기준으로 내림차순으로 정렬

```
mecab_dic = sorted(collections.Counter(nouns).items(), key = lambda t:t[1], reverse = True)
```

Jupyter notebook

Step3

분석 결과

MeCab

1	연구	61370
2	개발	61059
3	기술	56686
4	수	25193
5	분석	24168
6	시스템	19729
7	이용	19662
8	기반	18500
9	것	17188
10	등	16657
11	나노	15894
12	세포	15264
13	구조	14094

Komoran

1	연구	59008
2	개발	58540
3	기술	53160
4	수	26436
5	분석	24295
6	시스템	19072
7	기반	18425
8	것	17204
9	등	16368
10	세포	14768
11	구조	14572
12	평가	12391
13	특성	12164

Corpus.txt

THANK YOU FOR YOUR
T I M E
감 사 합 니 다

형태소분석기를 이용한 사전구축