Python을 활용한 텍스트 마이닝 3.텍스트 분석-데이터 전처리

```
okja = []
file = open('okja.txt','r',encoding='utf-8')
lines=file.readlines()

for line in lines:
okja.append(line)
file.close()
```

먼저 저번 포스팅에서 수집한 데이터를 불러와서 옥자라는 객체에 저장합니다.

```
from konlpy.tag import Twitter

twitter = Twitter()

sentences_tag=[]
for sentence in okja:
sentences_tag.append(twitter.pos(sentence))
```

그다음 저번에 설치한 KoNLPy 모듈을 불러와서 sentences_tag라는 리스트 형태로 저장합니다.

```
print(sentences_tag)
print('-'*30)
print(len(sentences_tag))
```


Search

검색

Recently posted

Recommendation System 1. 5월 13, 2018

Regression with Machine Learning 4. Regularization for spartsity(희소 학습) 4월 8, 2018

경사 하강법(Gradient Descent) 3월 27, 2018

Regression with Machine Learning 3. Constrained Least Squares(제약 최소제곱) 3월 18, 2018

Regression with Machine Learning 2. Stochastic Gradient Descent(확률적 경사법) 3월 18, 2018

Posts

2018년 5월 (1)

2018년 4월 (1)

를 입력하시면,

```
[[('웃기', 'Noun'), ('고', 'Josa'), ('재밌지'.....
.......
```

```
Eomi'), ('?', 'Punctuation')]]
------------------------------
300
```

이와 같은 형태로 추출된 token과 token의 형태소 분석 결과를 출력합니다. 300개의 token이 출력된 것을 볼 수 있습니다.

이제 자주 사용되는 단어를 추출하겠습니다.

```
noun_adj_list = []

for sentence in sentences_tag:
for word, tag in sentence:
if tag in ['Noun','Adjective']:
noun_adj_list.append(word)
```

먼저 분석에 중요한 품사로 판단되는 명사와 형용사 token만 추출하여 noun_adj_list에 담았습니다.

```
from collections import Counter
counts = Counter(noun_adj_list)
print(counts.most_common(10))
```

collections모듈을 이용해서 most frequent word 10가지를 출력한 결과입니다.

```
[('영화', 151), ('봉준호', 56), ('옥자', 56), ('좋', 55), ('
```

kis0403 7월 5, 2017 Text Mining 댓글 없음

File failed to load: /extensions/MathZoom.js

2018년 3월 (5)
2018년 1월 (1)
2017년 9월 (2)
2017년 8월 (6)
2017년 7월 (12)
2017년 6월 (7)

Etc

사이트 관리
로그아웃
글 RSS
댓글 RSS
WordPress.org

category

misc (2)
Python (17)
coding with python (1)
installation (5)
Neural Network (1)
Text Mining (10)
R (14)
2017 Weather Contest (4)
machine learning (5)
Packages & Base (2)
Recommendation System (2)
Visualization (2)

← Python을 활용한 텍스트 마이닝 2.텍스트 분석-데이터 수집

Python을 활용한 텍스트 마이닝 4.텍스트 분석-데이터 시각화 →

답글 남기기

kis0403로(으로) 로그인 함. 로그아웃?

댓글

댓글 달기