Python을 활용한 텍스트 마이닝 2.텍스트 분석-데이터 수집

텍스트를 분석 역시 일반적인 데이터 분석처럼 데이터 수집 -> 데이터 전처리 -> EDA -> 데이터 분석 -> knowledge 도출의 단계를 거칩니다. 하지만 unstructured data인 관계로 데이터의 수집과 전처리 과정에서 큰 차이가 나타납니다.

이번 포스팅에서는 movie.daum.net의 영화 리뷰 댓글을 이용해서 데이터를 수집하려고 합니다.

Jupyter notebook에서 새로운 스크립트를 만들고 제일 윗 부분에

```
#-*- coding: utf-8 -*-
```

를 입력합니다 기본적으로 윈도우 환경의 Python은 ANSI 형식을 따르기 때문에 한글이 깨지는 경우가 발생합니다. 그래서 기본 인코딩 형식을 utf-8로 바꾸셔야 텍스트를 불러왔을 때 깨지지 않게 됩니다.

```
from urllib.request import urlopen

from bs4 import BeautifulSoup
```

전에 설치한 크롤링 모듈을 불러옵니다.

먼저 한 페이지의 리뷰들을 불러오는 과정을 살펴보겠습니다.

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)

```
url='http://movie.daum.net/moviedb/grade?movieId=97728&type=netizen&page=2'
```

```
webpage=urlopen(url)
```

먼저 urlopen을 이용해 url주소의 리뷰 데이터를 객체에 저장합니다.

```
source = BeautifulSoup(webpage,'html.parser',from_encoding='utf-8')
```

HTML형식으로 된 데이터를 parsing하여 텍스트 형태로 변환합니다.

```
reviews = source.findAll('p',{'class': 'desc_review'})

for review in reviews:

print(review.get_text().strip())
```

HTML 소스 코드를 보면 내용 형태로 데이터가 저장되어 있습니다. 그래서 reviews라는 객체에 이와 같은 양식을 모두 찾아 list형식으로 변환했습니다. 또한 get_text()를 이용해 텍스트 데이터만 추출하고 strip()을 통해 공백을 제거했습니다.

2페이지의 리뷰들을 모두 추출했는데 for문을 통해 url을 리스트 형식으로 만들어 여러 페이지의 데이터를 추출할 수 있습니다.

```
review_list=[]
for n in range(10):
url = 'http://movie.daum.net/moviedb/grade?movieId=97728&type=netizen&page={}'.format(n+1)
webpage = urlopen(url)
```


Etc

category

```
source = BeautifulSoup(webpage,'html.parser',from_encoding='utf-8')
reviews = source.findAll('p',{'class': 'desc_review'})
for review in reviews:
review_list.append(review.get_text().strip().replace('₩n',' ').replace('₩t',' ').replace('₩r',' '))
```

그 다음 추출한 텍스트 데이터를 txt파일로 저장합니다.

```
file = open('okja.txt','w',encoding='utf-8')

for review in review_list:
file.write(review+'₩n')
file.close()
```

kis0403 7월 4, 2017 Text Mining 댓글 2개
편집

← Python을 활용한 텍스트 마이닝 1.모듈설치

Python을 활용한 텍스트 마이닝 3.텍스트 분석-데이터 전처리 →

"Python을 활용한 텍스트 마이닝 2.텍스트 분석-데이터 수집"에 대한 2개의 생각

ender info by CleanTalk
abouts@naver.com | 1.210.105.172 | Mark as spa

이충열
11월 13, 2017 9:27 오전
고유주소
고치기

안녕하세요. 좋은 글 잘 보고 있습니다.
다름이 아니라 올려주신 코드 보면서 따라해보고 있는데요.
오류가 나서 혹시 답변을 얻을수 있을까 싶어서요.

for review in reviews:
riview_list.append(review.get_text().strip().replace('₩n','').replace('₩t','').replace('₩r','')) 이 부분을 입력하면,

NameError: name 'riview_list' is not defined
와 같은 에러가 뜨더라구요.

응답

이충열
11월 14, 2017 8:52 오전
고유주소
고치기

에러 해결했는데 댓글 지우기가 안되네요 ㅎㅎ

응답

답글 남기기

kis0403로(으로) 로그인 함. 로그아웃?

댓글

댓글 달기