[python] 영화 리뷰에 대한 자연어 처리분석/ 감성분석하기 feat. 스크래핑 — 나무늘보의 개발 블로그

노트북:	첫 번째 노트북
만든 날짜:	2020-10-25 오후 5:13
URL:	https://continuous-development.tistory.com/107?category=736681

Python

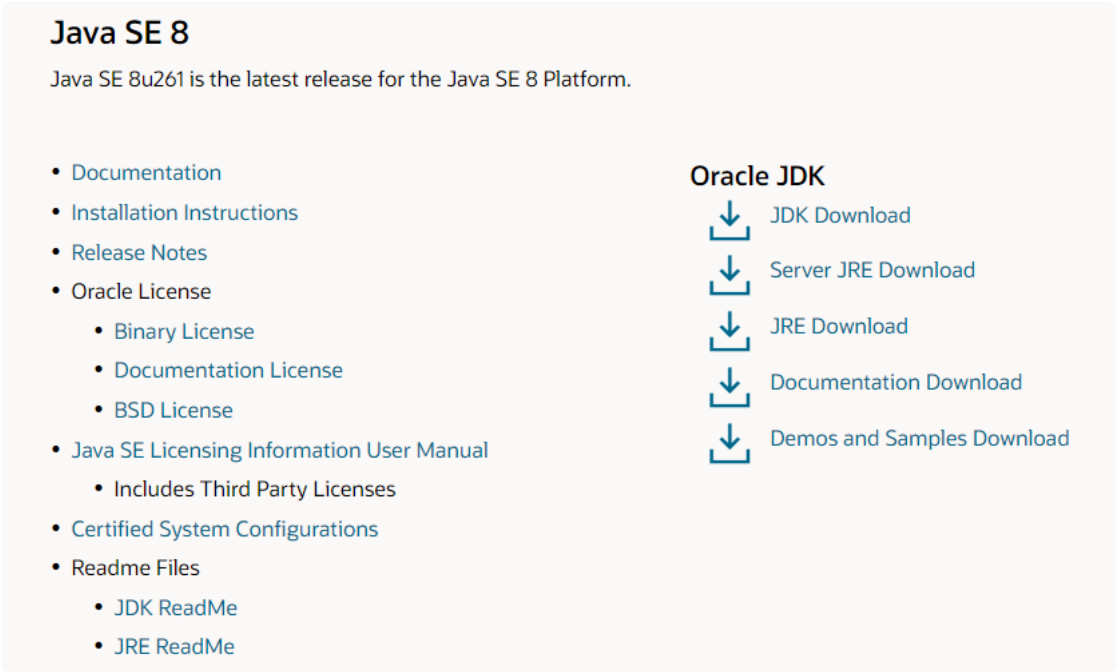
[python] 영화 리뷰에 대한 자연어 처리분석/ 감성분석하기 feat. 스크래핑

2020. 10. 7. 11:11 수정 삭제 공개

※아나콘다가 깔려있는 환경에서 실행했습니다.

자연어 처리를 하기 위해서는 jdk 가 필요하다. 그 이유는 자바 가상 머신 위에서 돌아가기 때문이다. 그래서 jdk를 다운로드한다.

https://www.oracle.com/kr/java/technologies/javase/javase-downloads.html



버전은 무난하게 8 정도를 받는다.

그다음 anaconda prompt를 들어가 아래와 같이 konlpy를 받는다.

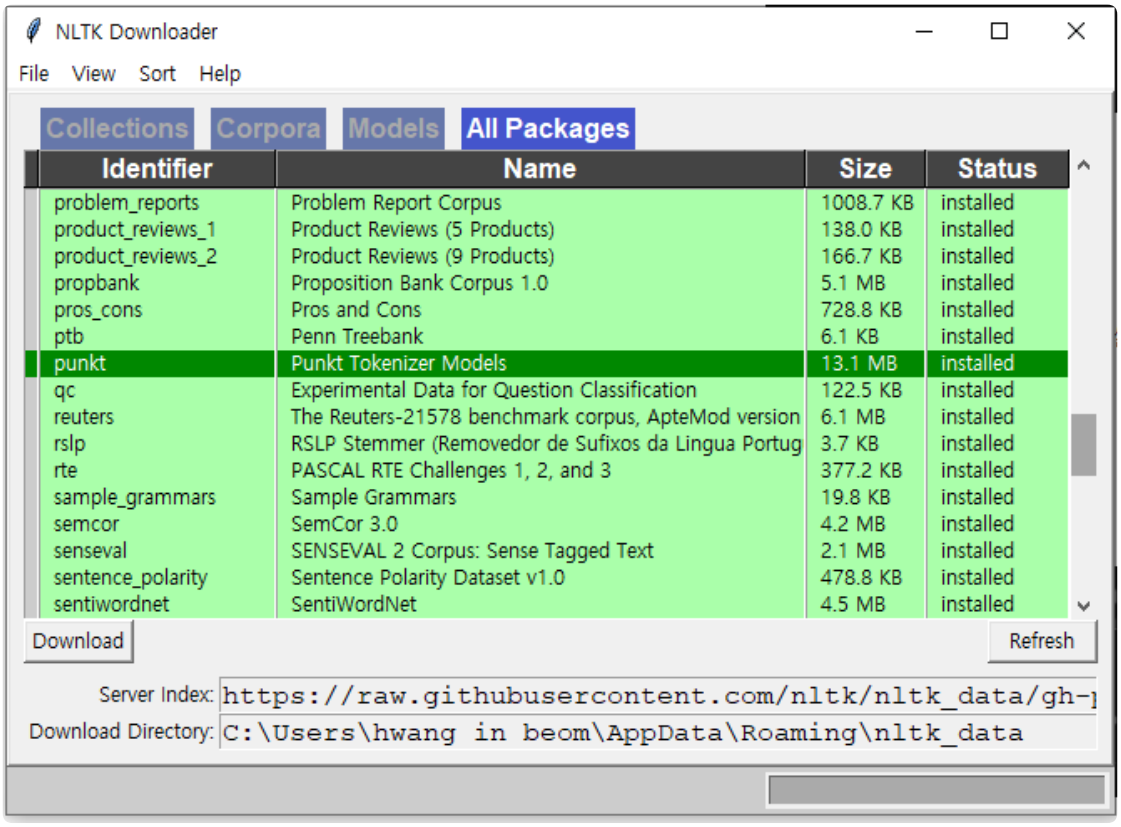
```
conda install konlpy
```

```
(base) C:₩Users₩hwang in beom>conda install konlpy
Collecting package metadata (repodata.json): done
Solving environment: failed with initial frozen solve. Retrying with flexible solve.

PackagesNotFoundError: The following packages are not available from current channels:

  - konlpy

Current channels:

  - https://repo.anaconda.com/pkgs/main/win-64
  - https://repo.anaconda.com/pkgs/main/noarch
  - https://repo.anaconda.com/pkgs/r/win-64
  - https://repo.anaconda.com/pkgs/r/noarch
  - https://repo.anaconda.com/pkgs/msys2/win-64
  - https://repo.anaconda.com/pkgs/msys2/noarch

To search for alternate channels that may provide the conda package you're
looking for, navigate to

    https://anaconda.org

and use the search bar at the top of the page.
```

설치 한 다음 python으로 들어간 다음
import nltk를 실행해본다. 정상적으로 설치가 되면 아래와 같이 뜨고
nltk.download()
라고 칠 경우 하나의 창이 뜬다.

```
(base) C:₩Users₩hwang in beom>PYTHON
Python 3.6.5 |Anaconda, Inc.| (default, Mar 29 2018, 13:32:41) [MSC v.1900 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> nltk.download
<bound method Downloader.download of <nltk.downloader.Downloader object at 0x0000018099A4DE10>>
>>> nltk.download()
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
True
```

All packages에서 punkt를 받고

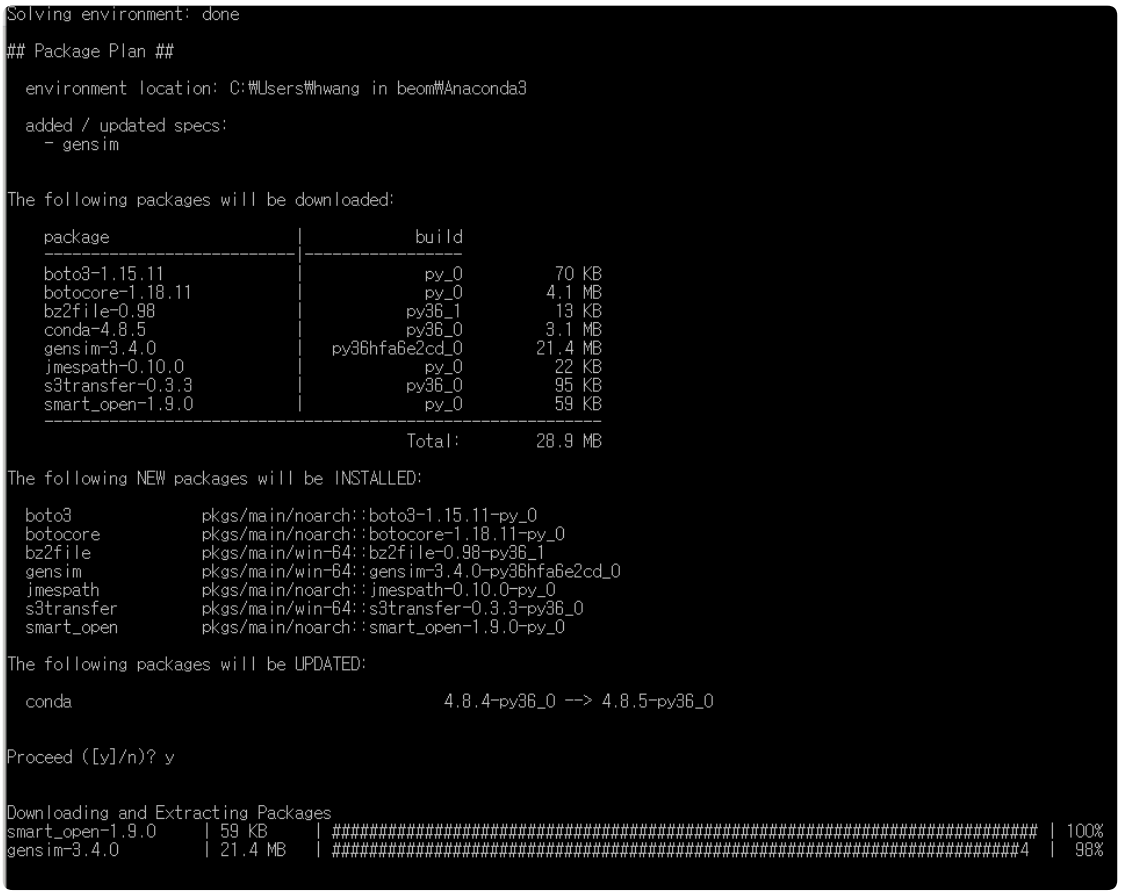


punkt 은 받고 사용하면서 필요한 부분은 또 받도록 하자

그다음

```
conda install gensim
```

워드를 벡터로 만들어주는 라이브러리다. 요새 많이 쓴다고 한다.



중간에 proceed가 뜨면 y를 눌러 다운로드를 이어간다.

그다음은

```
conda install -c conda-forge jpype1
```

자바와 파이썬이 통신하기 위해서 다운로드한다.

```
## Package Plan ##

  environment location: C:₩Users₩hwang in beom₩Anaconda3

  added / updated specs:
    - jpype1


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    ca-certificates-2020.6.20  |       hecda079_0         184 KB  conda-forge
    certifi-2020.6.20          |   py36h9f0ad1d_0         152 KB  conda-forge
    conda-4.8.5                |   py36h9f0ad1d_1         3.1 MB  conda-forge
    jpype1-0.7.2               |   py36he980bc4_0         1.1 MB  conda-forge
    openssl-1.1.1f             |       hfa6e2cd_0         4.7 MB  conda-forge
    python_abi-3.6             |          1_cp36m           4 KB  conda-forge
    ------------------------------------------------------------
                                           Total:         9.2 MB

The following NEW packages will be INSTALLED:

  jpype1             conda-forge/win-64::jpype1-0.7.2-py36he980bc4_0
  python_abi         conda-forge/win-64::python_abi-3.6-1_cp36m

The following packages will be UPDATED:

  conda                      pkgs/main::conda-4.8.5-py36_0 --> conda-forge::conda-4.8.5-py36h9f0ad1d_1
  openssl              pkgs/main::openssl-1.1.1c-he774522_1 --> conda-forge::openssl-1.1.1f-hfa6e2cd_0

The following packages will be SUPERSEDED by a higher-priority channel:

  ca-certificates    pkgs/main::ca-certificates-2020.7.22-0 --> conda-forge::ca-certificates-2020.6.20-hecda079_0
  certifi               pkgs/main::certifi-2020.6.20-py36_0 --> conda-forge::certifi-2020.6.20-py36h9f0ad1d_0


Proceed ([y]/n)? y
```

이것 또한 y를 입력해 다운을 받는다.

이제 기본적인 자연어 처리를 해보자

자연어 처리 기초

- 꼬꼬마
-

In [27]:
```
from konlpy.tag import Kkma
```

In [28]:
```
kkoma = Kkma()
```

```
C:₩Users₩hwang in beom₩Anaconda3₩lib₩site-packages₩jpype₩_core.py:217: UserWarning:
-------------------------------------------------------------------------------
Deprecated: convertStrings was not specified when starting the JVM. The default
behavior in JPype will be False starting in JPype 0.8. The recommended setting
for new code is convertStrings=False.  The legacy value of True was assumed for
this session. If you are a user of an application that reported this warning,
please file a ticket with the developer.
-------------------------------------------------------------------------------

  """)
```

In [29]:
```
kkoma.nouns('한국어 문장 분석을 시작합니다. 재미있어요~~')
```

Out[29]:
```
['한국어', '문장', '분석']
```

In [30]:
```
kkoma.pos('한국어 문장 분석을 시작합니다. 재미있어요~~')
```

Out[30]:
```
[('한국어', 'NNG'),
 ('문장', 'NNG'),
 ('분석', 'NNG'),
 ('을', 'JKO'),
 ('시작하', 'VV'),
 ('ㅂ니다', 'EFN'),
 ('.', 'SF'),
 ('재미있', 'VA'),
 ('어요', 'EFN'),
 ('~~', 'SW')]
```

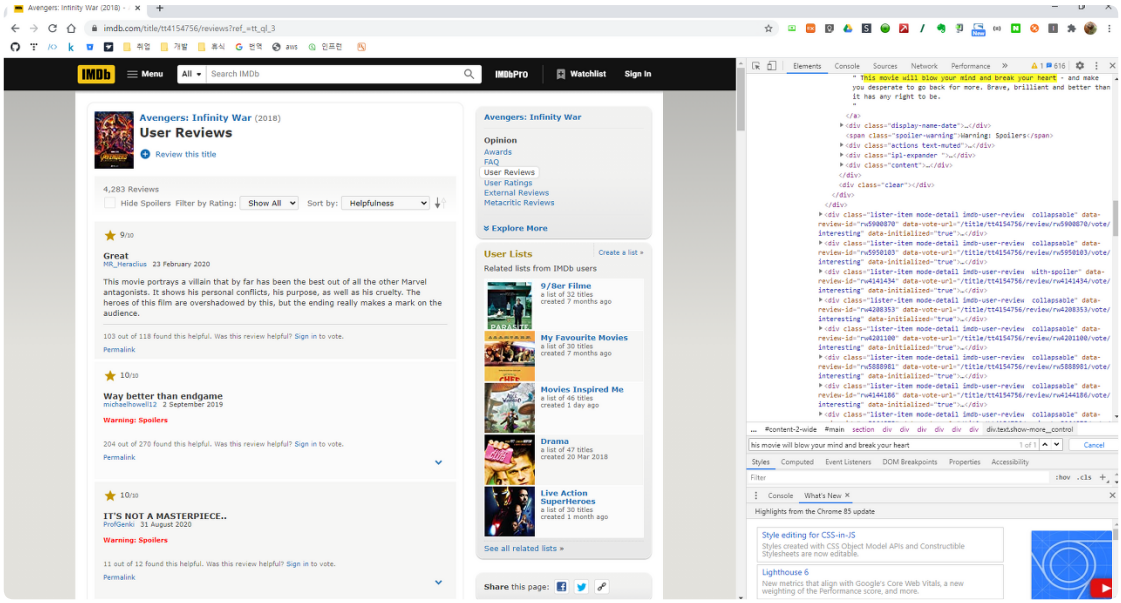
nouns는 명사를 추출하고 pos는 문장 내 단어들의 품사를 식별을 한다.
멀쩡히 되는 걸 확인한다.

감성 분석 하기

```
- 점수(별점), 리뷰제목, 작성자 닉네임, 작성날짜, 리뷰내용
- 감정분석(VADER) - NLTK
- good+0.1 , awful -0.1, perfect+0.2
- 문장에서 저런 단어가 추출되면 나올 때마다 점수를 더하고 빼서 점수 긍정, 점수 부정
```

In [57]:
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from nltk import tokenize
import nltk
```

감성 분석을 하기 위해 위와 같이 import를 한다.
여기서는 영화 리뷰에 대한 감성 분석이다.



이 부분은 영화 리뷰에 대해 스크래핑을 하기 위한 작업이다.

In [1]:
```
from urllib.request import urlopen
from bs4 import BeautifulSoup
from urllib.error import HTTPError
from urllib.error import URLError
import pandas as pd
```

In [16]:
```
url = 'https://www.imdb.com/title/tt4154756/reviews?ref_=tt_ql_3'
try:
    html = urlopen(url)
except HTTPError as he:
    print('http error')
except URLError as us:
    print('url error')
else:
    soup = BeautifulSoup(html.read(),'html.parser', from_encoding='UTF-8')
```

- 점수(별점), 리뷰제목, 작성자 닉네임, 작성날짜, 리뷰내용

In [24]:
```
reivew_list= soup.find_all('div',{'class':'imdb-user-review'})
```

In [56]:
```
sid = SentimentIntensityAnalyzer()
```

감성 분석을 하기 위해 위 함수를 넣는다.

In [136]:
```
sum_review=''
for review in reivew_list:
    score = review.find('span').get_text().replace('₩n','')
    title = review.find('a').get_text().replace('₩n','')
    writer = review.find('span',{'class':'display-name-link'}).get_text()
    date = review.find('span',{'class':'review-date'}).get_text()
    content = review.find('div',{'class':'text show-more__control'}).get_text()
    sum_review =sum_review + content

    lines_list =tokenize.sent_tokenize(content)
    sum = 0
    # polarity_scores() : 문장을 단어별로 분석해서 긍정, 부정, 중립에 대한 점수를 계산해주고 종합 점수를 반환
    for sent in lines_list:
        ss = sid.polarity_scores(sent)
        sum = sum + ss['compound']
    sum1 = (sum/len(lines_list))

    data.append([score, title, writer, date, content, sum1])
```

이렇게 해당 값들을 빼오고
tokenize.sent_tokenize(content)를 통해 content 값을 토큰화 한다.
그다음 토큰화 한 값을 문장을 단어별로 분석해서 긍정, 부정, 중립에 대한 점수를 계산하고 반환하는 로직을 구현한다.
sid에 있는 함수 polarity_scores라는 함수를 통해 토큰화 한 sent를 넣어 점수를 계산하고 sum에서 점수를 합한다.
그렇게 총점을 구한 뒤 길이로 나눠 점수를 매긴다.

In [160]:
```
import pandas as pd

df = pd.DataFrame(data)
df.columns = ['score', 'title', 'writer', 'date', 'content', 'sum']
df.to_csv('./service_imdb_wordcloud.csv')
```

우리가 만들었던 리스트를 데이터 프레임 형태로 만들고 그것을 to_csv로 csv 파일로 만든다.

	A	B	C	D	E	F	G	H
5	3	10월 10일	Somehow	Jesper280	#######	I consider	0.222762	
6	4	10월 10일	Unlike ar	kjames-26	#######	This movie	-0.01282	
7	5	10월 10일	This movi	shawneoft	#######	Over the p	0.123665	
8	6	10월 10일	Best movi	udit-mehr	12-Jul-20	Best movie	0.29388	
9	7	10월 10일	Worth the	ubtgkse	29-Jul-20	Avengers i	0.49345	
	8	10월 10일	A Summe	garethvk	#######	Summer movies often hype themselves as spectacular events not to be missed and their ad campaigns use words	0.290167	

csv 파일의 결과는 위와 같다. 이렇게 해당 리뷰에 대한 감성 분석을 할 수 있다.

'Python' 카테고리의 다른 글

[Python] BeautifulSoup을 통한 이미지 블로그 스크래핑하기

[Python] BeautifulSoup을 통한 이미지 스크래핑 하기

[python] 영화 리뷰에 대한 자연어 처리분석/ 감성분석하기 feat. 스크래핑

[python] BeautifulSoup를 통한 영화리뷰 scraping 하기

[Python] 파이썬 기초 14 - 아주 기초적인 pandas 사용법과 예제

[Python] 파이썬 기초 13 - 파이썬을 통한 파일 입출력 사용법

movie 감성분석 감성 분석 영화 리뷰 감성분석

나무늘보스

혼자 끄적끄적하는 블로그 입니다.