nytimesNewsFetch

March 29, 2018

Jayant Solanki

0.1 Finding trending words from a news topic in NYTimes website.

0.1.1 Using nytimes articles API to fetch articles for a particular search topic

In [2]:
```
#Background: The nytimes api doesn't returns full article body, just the snippet of it
# so, I have to write the code to fetch the weburl for each article, then perfrom furth
#fetch the story body of each article
import requests # for performing html request to nytimes API
from datetime import datetime
from time import sleep
import lxml.html as html # for scrapping the content of the article URL of nytimes
import json
```

In [3]:
```
# Each API keys can provide 10000 response hits
topic = "south+china+sea"#topic to be looked for
apikey = "527c18ffc4e648cb936f582a0e264ff1"
fl = "snippet,web_url"#selective attributes of json response
pageNo = "0"#initial page is 0, articles fetched using api are grouped in 10 per page
dateRange = ["20180321", "20180322", "20180323", "20180324", "20180325", "20180326", "2
```

In [4]:
```
#this function parse the json response from nytimes and create new dictionary using tw
def parse_articles(articles):
    '''
    This function takes in a response to the NYT api and parses
    the articles into a list of dictionaries
    '''
    news = []
    fetch = articles['response']['docs']
    for i in range(0,len(fetch)):
        dic = {}
#         print(fetch[i])
        dic['web_url'] = fetch[i]['web_url']
        if fetch[i]['snippet'] is not None:
            dic['snippet'] = fetch[i]['snippet']
#         dic['url'] = i['web_url']
        news.append(dic)
    return(news)
```

In [5]:
```
# this function perfrom request to nytimes api using the paramters passed and returns
def get_articles(topic, begin_date, end_date, fl, apikey):
    all_articles = []#stores all articles for a particular day
    page = 0
    while(page<100):
        sleep(1)
#     for page in range(0,100): #NYT limits pages to first 100 pages starting page 0,
        try:

            url = "http://api.nytimes.com/svc/search/v2/articlesearch.json?q="+topic+"&
            print(url)
            requestArticles = requests.get(url)
            data = requestArticles.json()
            if len(data["response"]["docs"])>0:
                all_articles.append(parse_articles(data))
#                 print(data)
            else:# checks if further pages have no articles to show, if yes then break
                print(parse_articles(data))
                break
        except:
            print("You called the api way to fast, Dude, trying again")
            print(data)
#             page = page - 1
            sleep(1)
            continue#try again
        print("Page: "+str(page))
#         break
        page=page+1
    return(all_articles)
```

In [7]:
```
#caller function
processArticles = []
for i in range(0,7):
    datetimeobject = datetime.strptime(dateRange[i],'%Y%m%d')
    beginDate = datetimeobject.strftime('%m-%d-%Y')
    datetimeobject = datetime.strptime(dateRange[i+1],'%Y%m%d')
    endDate = datetimeobject.strftime('%m-%d-%Y')
    print("Fetching articles for Data period: " + beginDate + " - "+ endDate)
    processArticles = get_articles(topic, dateRange[i], dateRange[i+1],fl, apikey)
    if(len(processArticles)>0):
#         try:
#             dataToWrite = processArticles
# #             print(dataToWrite)
#         except:
#             print(len(processArticles))
#             print(processArticles)
#             print(processArticles[0])
#             break
```

```
                with open("textcorpus/"+topic+dateRange[i]+".txt", 'w') as outfile:#used for s
                    for item in processArticles:
                        for articles in item:
            #                 print(articles)
                            outfile.write(articles["snippet"])
                            outfile.write("\n")
                with open("textcorpus/"+topic+dateRange[i]+"-full.txt", 'w') as outfile:#used
                    for item in processArticles:
                        for articles in item:
                            fullpage = requests.get(articles["web_url"])
                            htmlbody = html.fromstring(requests.get(articles["web_url"]).conte
                            output = "".join(htmlbody.xpath('//p[contains(@class,"story-body-te
            #                 print(output+"\n\n")
            #                 output = output.decode('utf8').encode('latin1').decode('utf8')
                            output = str(output.encode("ascii", "ignore"))#since the body has
                # utf-8 response into ascii using ignore flag to bypass those escape characters
                            outfile.write(output[2:-1])
                            outfile.write("\n")
                else:
                    print("Insufficient data for date: "+beginDate+" to save")
            #     break
            # print(processArticles[0][0:-1])
```

```
Fetching articles for Data period: 03-21-2018 - 03-22-2018
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180321&
Page: 0
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180321&
Page: 1
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180321&
Page: 2
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180321&
[]
Fetching articles for Data period: 03-22-2018 - 03-23-2018
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180322&
Page: 0
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180322&
Page: 1
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180322&
Page: 2
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180322&
Page: 3
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180322&
Page: 4
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180322&
[]
Fetching articles for Data period: 03-23-2018 - 03-24-2018
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180323&
Page: 0
```

```
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180323&e
Page: 1
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180323&e
Page: 2
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180323&e
Page: 3
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180323&e
Page: 4
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180323&e
Page: 5
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180323&e
[]
Fetching articles for Data period: 03-24-2018 - 03-25-2018
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180324&e
Page: 0
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180324&e
Page: 1
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180324&e
[]
Fetching articles for Data period: 03-25-2018 - 03-26-2018
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180325&e
Page: 0
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180325&e
Page: 1
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180325&e
[]
Fetching articles for Data period: 03-26-2018 - 03-27-2018
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180326&e
Page: 0
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180326&e
Page: 1
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180326&e
Page: 2
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180326&e
Page: 3
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180326&e
[]
Fetching articles for Data period: 03-27-2018 - 03-28-2018
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180327&e
Page: 0
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180327&e
Page: 1
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180327&e
Page: 2
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180327&e
Page: 3
http://api.nytimes.com/svc/search/v2/articlesearch.json?q=south+china+sea&begin_date=20180327&e
[]
```

```
In [ ]: # /usr/local/hadoop/bin/hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-str
```