Sathish_Kumar_Rajendiran_Week9.2_Text_Tokenization

August 26, 2020

```
Name: Sathish Kumar Rajendiran
Task: Week 9: 9.2 Text Tokenization
Date: 8/26/2020
```

[ ]:
```
#pip install pymongo library
# !pip install pymongo
# !pip install tweepy
```

[53]:
```
# nltk.download('punkt')
# nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/sathishrajendiran/nltk_data…
[nltk_data]   Unzipping corpora/stopwords.zip.
```

[53]:
```
True
```

[42]:
```
#import libraries

# standard library
import os
import sys
from datetime import datetime
import time


# text tokenization
import nltk
import re

#MongoDB libraries
import pymongo
from pymongo import MongoClient



os.getcwd()
```

[42]:
```
'/Users/sathishrajendiran/ist652-python'
```

[10]:
```
#  sample text
text = '''I'll never fly Delta again!! My flight was
supposed to leave MCO for ATL at 6:25pm on Saturday,
March 26, however, due to severe weather, it was
delayed until 8:12pm - no problem. At approx. 8pm we
started boarding. We just sat there at the gate. No
explanation etc. until about 9:30pm when the pilot said
we were pushing away from the gate but wouldn't take
off. '''
```

[11]:
```
#split text into words
words = text.split()
# words
```

[33]:
```
# Connecting to the database
# Connection to Mongo DB
try:
    client = MongoClient('localhost', 27017)
    print ('Authentication OK - You''re now connected to the MongoDB.\n')
# use database named fbusers or create it if not there already
    db = client.bball
    # create collection named delta or create it if not there already
    coll = db.bbcoll
    print('MongoDB database: ' + str(db))
    print('MongoDB collection:' + str(coll))

except pymongo.errors.ConnectionFailure as e:
    print ('Could not connect to MongoDB: %s' % e )
```

```
Authentication OK - Youre now connected to the MongoDB.

MongoDB database: Database(MongoClient(host=['localhost:27017'],
document_class=dict, tz_aware=False, connect=True), 'bball')
MongoDB collection:Collection(Database(MongoClient(host=['localhost:27017'],
document_class=dict, tz_aware=False, connect=True), 'bball'), 'bbcoll')
```

[34]:
```
#list database defined
client.database_names()
```

```
/opt/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:2:
DeprecationWarning: database_names is deprecated. Use list_database_names
instead.
```

[34]:
```
['admin',
 'bball',
 'config',
```

```
 'covid',
 'disney',
 'fbusers',
 'local',
 'peopledb',
 'pytweetsdb',
 'tweetsdb',
 'usgs']
```

[35]:
```
db.collection_names()
```

```
/opt/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:1:
DeprecationWarning: collection_names is deprecated. Use list_collection_names
instead.
  """Entry point for launching an IPython kernel.
```

[35]:
```
['bbcoll']
```

[36]:
```
#search the first item from the collection
coll.find_one()
```

[36]:
```
{'_id': ObjectId('58d992de64a4f3e56d2db3c7'),
 'user': {'profile_background_tile': True,
  'friends_count': 217,
  'profile_sidebar_fill_color': 'EFEFEF',
  'id_str': '513196438',
  'is_translation_enabled': False,
  'profile_link_color': 'F70808',
  'followers_count': 256,
  'location': '',
  'protected': False,
  'default_profile_image': False,
  'contributors_enabled': False,
  'favourites_count': 2187,
  'profile_background_color': 'BFD0D9',
  'statuses_count': 1104,
  'id': 513196438,
  'profile_banner_url':
'https://pbs.twimg.com/profile_banners/513196438/1437359097',
  'created_at': 'Sat Mar 03 13:54:37 +0000 2012',
  'profile_image_url_https':
'https://pbs.twimg.com/profile_images/732514428051181568/OkOVa8Ia_normal.jpg',
  'time_zone': 'Eastern Time (US & Canada)',
  'follow_request_sent': None,
  'listed_count': 1,
  'utc_offset': -14400,
  'lang': 'en',
```

```
  'is_translator': False,
  'name': 'Will',
  'description': "Do not worry about tomorrow, for tomorrow will worry about
itself.-Matthew 6:34 HSC'20",
  'profile_use_background_image': True,
  'verified': False,
  'geo_enabled': True,
  'profile_text_color': '333333',
  'profile_image_url':
'http://pbs.twimg.com/profile_images/732514428051181568/OkOVa8Ia_normal.jpg',
  'entities': {'description': {'urls': []}},
  'notifications': None,
  'url': None,
  'translator_type': 'none',
  'has_extended_profile': True,
  'default_profile': False,
  'screen_name': 'frenchythe1st',
  'following': None,
  'profile_background_image_url': 'http://pbs.twimg.com/profile_background_image
s/850934042/26e0a43c1f821ac098571fb3de80944c.jpeg',
  'profile_sidebar_border_color': 'FFFFFF',
  'profile_background_image_url_https': 'https://pbs.twimg.com/profile_backgroun
d_images/850934042/26e0a43c1f821ac098571fb3de80944c.jpeg'},
 'metadata': {'iso_language_code': 'en', 'result_type': 'recent'},
 'favorited': False,
 'in_reply_to_user_id': None,
 'text': 'RT @HowardWKYT: The final seconds of the Kentucky-North Carolina game
was an emotional roller coaster for Big Blue fans. #marchmadness #WKY…',
 'retweet_count': 9739,
 'id': 8.464896333870899e+17,
 'in_reply_to_screen_name': None,
 'created_at': 'Mon Mar 27 22:30:30 +0000 2017',
 'is_quote_status': False,
 'retweeted_status': {'user': {'profile_background_tile': False,
   'friends_count': 573,
   'profile_sidebar_fill_color': 'C0DFEC',
   'id_str': '27059989',
   'is_translation_enabled': False,
   'profile_link_color': '0084B4',
   'followers_count': 3330,
   'location': 'Lexington, KY',
   'protected': False,
   'default_profile_image': False,
   'contributors_enabled': False,
   'favourites_count': 1231,
   'profile_background_color': '000000',
   'statuses_count': 9463,
```

```
   'id': 27059989,
   'profile_banner_url':
'https://pbs.twimg.com/profile_banners/27059989/1431999643',
   'created_at': 'Fri Mar 27 18:09:16 +0000 2009',
   'profile_image_url_https':
'https://pbs.twimg.com/profile_images/575376840304386049/atONJG3G_normal.jpeg',
   'time_zone': 'Eastern Time (US & Canada)',
   'follow_request_sent': None,
   'listed_count': 77,
   'utc_offset': -14400,
   'lang': 'en',
   'is_translator': False,
   'name': 'Lee K. Howard',
   'description': 'Sports Anchor/Reporter for CBS/FOX in Lexington Kentucky,
providing sports news and my random rants and chants!',
   'profile_use_background_image': True,
   'verified': True,
   'geo_enabled': True,
   'profile_text_color': '333333',
   'profile_image_url':
'http://pbs.twimg.com/profile_images/575376840304386049/atONJG3G_normal.jpeg',
   'entities': {'description': {'urls': []},
    'url': {'urls': [{'expanded_url':
'http://www.facebook.com/profile.php?id=100003241678454',
       'display_url': 'facebook.com/profile.php?id…',
       'indices': [0, 22],
       'url': 'http://t.co/C2UFfLkjB1'}]}},
   'notifications': None,
   'url': 'http://t.co/C2UFfLkjB1',
   'translator_type': 'none',
   'has_extended_profile': False,
   'default_profile': False,
   'screen_name': 'HowardWKYT',
   'following': None,
   'profile_background_image_url': 'http://pbs.twimg.com/profile_background_imag
es/596008542/6qfqqg0jndpp3su53ay5.jpeg',
   'profile_sidebar_border_color': 'A8C7F7',
   'profile_background_image_url_https': 'https://pbs.twimg.com/profile_backgrou
nd_images/596008542/6qfqqg0jndpp3su53ay5.jpeg'},
  'metadata': {'iso_language_code': 'en', 'result_type': 'recent'},
  'favorited': False,
  'in_reply_to_user_id': None,
  'text': 'The final seconds of the Kentucky-North Carolina game was an
emotional roller coaster for Big Blue fans.… https://t.co/TPZ6PuXHxH',
  'possibly_sensitive': False,
  'retweet_count': 9739,
  'id': 8.462142300261786e+17,
```

```
  'in_reply_to_screen_name': None,
  'created_at': 'Mon Mar 27 04:16:08 +0000 2017',
  'is_quote_status': False,
  'favorite_count': 12590,
  'contributors': None,
  'lang': 'en',
  'in_reply_to_status_id_str': None,
  'source': '<a href="http://twitter.com" rel="nofollow">Twitter Web
Client</a>',
  'in_reply_to_status_id': None,
  'place': {'country': 'United States',
   'contained_within': [],
   'id': '6ffcf3b0b904bbcb',
   'country_code': 'US',
   'bounding_box': {'type': 'Polygon',
    'coordinates': [[[-89.57151, 36.497129],
      [-81.964971, 36.497129],
      [-81.964971, 39.147359],
      [-89.57151, 39.147359]]]},
   'place_type': 'admin',
   'full_name': 'Kentucky, USA',
   'url': 'https://api.twitter.com/1.1/geo/id/6ffcf3b0b904bbcb.json',
   'name': 'Kentucky',
   'attributes': {}},
  'entities': {'hashtags': [],
   'urls': [{'expanded_url':
'https://twitter.com/i/web/status/846214230026178564',
     'display_url': 'twitter.com/i/web/status/8…',
     'indices': [106, 129],
     'url': 'https://t.co/TPZ6PuXHxH'}],
   'symbols': [],
   'user_mentions': []},
  'geo': None,
  'truncated': True,
  'coordinates': None,
  'in_reply_to_user_id_str': None,
  'retweeted': False,
  'id_str': '846214230026178564'},
 'favorite_count': 0,
 'contributors': None,
 'lang': 'en',
 'in_reply_to_status_id_str': None,
 'source': '<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter
for iPhone</a>',
 'in_reply_to_status_id': None,
 'place': None,
 'entities': {'hashtags': [{'indices': [121, 134], 'text': 'marchmadness'}],
```

```
 'urls': [],
 'symbols': [],
 'user_mentions': [{'indices': [3, 14],
   'name': 'Lee K. Howard',
   'id': 27059989,
   'id_str': '27059989',
   'screen_name': 'HowardWKYT'}]},
'geo': None,
'truncated': False,
'coordinates': None,
'in_reply_to_user_id_str': None,
'retweeted': False,
'id_str': '846489633387089920'}
```

[39]:
```
#read through collections for tweets
docs = coll.find()
doclist = list(docs)
msglist =[doc['text'] for doc in doclist if 'text' in doc.keys()]
print('Number of docs: ',len(msglist))
```

```
Number of docs:  2000
```

[50]:
```
all_tokens = [tok for msg in msglist for tok in nltk.word_tokenize(msg)]
print('Number of tokens: ',len(all_tokens))
#print top 5 tokens
all_tokens[:50]
```

```
Number of tokens:  45701
```

[50]:
```
['RT',
 '@',
 'HowardWKYT',
 ':',
 'The',
 'final',
 'seconds',
 'of',
 'the',
 'Kentucky-North',
 'Carolina',
 'game',
 'was',
 'an',
 'emotional',
 'roller',
 'coaster',
 'for',
```

```
 'Big',
 'Blue',
 'fans',
 '.',
 '#',
 'marchmadness',
 '#',
 'WKY…',
 'RT',
 '@',
 'WhistleSports',
 ':',
 'When',
 'you',
 'perfectly',
 'time',
 'the',
 '#',
 'UNC',
 'buzzer',
 'beater',
 ' ',
 '#',
 'MarchMadness',
 '(',
 'via',
 ':',
 '@',
 'SamuelGrubbs1',
 ')',
 'https',
 ':']
```

[47]:
```
msgFD = nltk.FreqDist(all_tokens)
msgFD.most_common(30)
```

[47]:
```
[('#', 4245),
 (':', 2900),
 ('@', 2154),
 ('RT', 1543),
 ('https', 1224),
 ('marchmadness', 1124),
 ('the', 1092),
 ('.', 1049),
 ('MarchMadness', 928),
 ('of', 901),
 ('game', 669),
```

```
 ('Carolina', 668),
 ('for', 659),
 ('was', 645),
 ('fans', 644),
 ('The', 582),
 ('final', 526),
 ('an', 526),
 ('seconds', 515),
 ('Big', 510),
 ('HowardWKYT', 507),
 ('Kentucky-North', 507),
 ('emotional', 507),
 ('roller', 507),
 ('coaster', 507),
 ('Blue', 507),
 ('WKY…', 507),
 ('!', 492),
 ('to', 290),
 (',', 279)]
```

[49]:
```
#all tokens to lowercase
all_tokens = [tok.lower() for msg in msglist for tok in nltk.word_tokenize(msg)]
all_tokens[:30]
```

[49]:
```
['rt',
 '@',
 'howardwkyt',
 ':',
 'the',
 'final',
 'seconds',
 'of',
 'the',
 'kentucky-north',
 'carolina',
 'game',
 'was',
 'an',
 'emotional',
 'roller',
 'coaster',
 'for',
 'big',
 'blue',
 'fans',
 '.',
 '#',
```

```
 'marchmadness',
 '#',
 'wky…',
 'rt',
 '@',
 'whistlesports',
 ':']
```

[55]:
```
nltk_stopwords = nltk.corpus.stopwords.words('english')
print('Number of stopwords: ',len(nltk_stopwords))
nltk_stopwords
```

```
Number of stopwords:  179
```

[55]:
```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
 'his',
 'himself',
 'she',
 "she's",
 'her',
 'hers',
 'herself',
 'it',
 "it's",
 'its',
 'itself',
 'they',
 'them',
 'their',
```

```
'theirs',
'themselves',
'what',
'which',
'who',
'whom',
'this',
'that',
"that'll",
'these',
'those',
'am',
'is',
'are',
'was',
'were',
'be',
'been',
'being',
'have',
'has',
'had',
'having',
'do',
'does',
'did',
'doing',
'a',
'an',
'the',
'and',
'but',
'if',
'or',
'because',
'as',
'until',
'while',
'of',
'at',
'by',
'for',
'with',
'about',
'against',
'between',
'into',
```

```
'through',
'during',
'before',
'after',
'above',
'below',
'to',
'from',
'up',
'down',
'in',
'out',
'on',
'off',
'over',
'under',
'again',
'further',
'then',
'once',
'here',
'there',
'when',
'where',
'why',
'how',
'all',
'any',
'both',
'each',
'few',
'more',
'most',
'other',
'some',
'such',
'no',
'nor',
'not',
'only',
'own',
'same',
'so',
'than',
'too',
'very',
's',
```

```
't',
'can',
'will',
'just',
'don',
"don't",
'should',
"should've",
'now',
'd',
'll',
'm',
'o',
're',
've',
'y',
'ain',
'aren',
"aren't",
'couldn',
"couldn't",
'didn',
"didn't",
'doesn',
"doesn't",
'hadn',
"hadn't",
'hasn',
"hasn't",
'haven',
"haven't",
'isn',
"isn't",
'ma',
'mightn',
"mightn't",
'mustn',
"mustn't",
'needn',
"needn't",
'shan',
"shan't",
'shouldn',
"shouldn't",
'wasn',
"wasn't",
'weren',
```

```
 "weren't",
 'won',
 "won't",
 'wouldn',
 "wouldn't"]
```

[57]:
```
def alpha_filter(w):
    pattern = re.compile('^[^a-z]+$')
    if (pattern.match(w)):
        return True
    else:
        return False
```

[59]:
```
token_list = [tok for tok in all_tokens if not alpha_filter(tok)]
print('Number of tokens: ',len(token_list))
token_list[:30]
```

```
Number of tokens:  29855
```

[59]:
```
['HowardWKYT',
 'The',
 'final',
 'seconds',
 'of',
 'the',
 'Kentucky-North',
 'Carolina',
 'game',
 'was',
 'an',
 'emotional',
 'roller',
 'coaster',
 'for',
 'Big',
 'Blue',
 'fans',
 'marchmadness',
 'WhistleSports',
 'When',
 'you',
 'perfectly',
 'time',
 'the',
 'buzzer',
 'beater',
 'MarchMadness',
```

```
 'via',
 'SamuelGrubbs1']
```

[62]:
```
# Top 30 words by frequency

msgFD = nltk.FreqDist(token_list)
top_words = msgFD.most_common(30)
for word, freq in top_words:
    print(word, freq)
```

```
https 1224
marchmadness 1124
the 1092
MarchMadness 928
of 901
game 669
Carolina 668
for 659
was 645
fans 644
The 582
final 526
an 526
seconds 515
Big 510
HowardWKYT 507
Kentucky-North 507
emotional 507
roller 507
coaster 507
Blue 507
to 290
a 265
's 235
and 208
you 184
BleacherReport 180
And 175
is 175
four 173
```