FIT5196 Assignment 1

Task 1

Student name: Ayush Sharma

Due date: 13 Sept 2020

Environment: Python 3.6.0 and Anaconda 4.3.0 (64-bit)

Library used:

- re: This library is used to work on the document for pattern matching using regular expression
- os: os library is used to parse between directory and the files in it to load them for reading the data
- langid: langid is a python library the help identify the langague passed into the string and identify the language of the text

1.) INTRODUCTION:

This assignment works on text processing and parsing between mupltiples files. Looking at the task it focuses on analyzing textual data which includes extracting data which can be in form of semi-structured data as well. Provides by the examiner we are given data of 2020 tweets bases on COVID-19/ CoronaVirus. The task is to extract the data and transform the data into the XML format with the following elements:

1. id: is a 19-digit number.
2. text: is the actual tweet.
3. Created_at: is the date and time that the tweet was created

2.) Importing Library:

In [1]:

```
# Importing Libraries
import re
import os
import langid
```

3.) Loading Data and Storing it:

First work is to load the data from the folder that has more then 2000+ text files that contains the tweet, date, id I have contain the directory in a list and parse on it to read each file and work on it. It is loaded in a variable named `data` .

- each tweet have a date id and tweet text this can be in any order
- I have iterated the list of files and opened them one by one and appended it to a single string

In [2]:

```
%%time
file_list=[]
path = '30823293_Task1/'
for root, dirs, files in os.walk(path):
    file_list = files
data=str()
for i in file_list:
    textdata=open(path+i,'r',encoding='UTF-8')
    data=data+textdata.read()
```

```
Wall time: 53.3 s
```

4.) Processing on the Data

In this step we will read the whole text data and analyse it according to the required items that are need and will work on understaning the data after loading the tweet data in the `data` variable I printed it and understood the form of data and after that I understood that it is a semi-structed data.

- to make the data in a better format and work on it in a better way I decided to create a list
- using the 'split' function of a list type data we have converted the `text string` into a list from `{` because each tweet begined at `{` .
- Created empty lists to append the required data of date, tweet, and id.
- I have applied 3 regex to extract the date, tweet Id, and the tweet text

used RegEx:

- Tweet Text: `(?:(?:"text":"(.*?)(?:(?:"})|(?:","created_at")|(?:","id"))))` which extracts the text by finding the pattern
- Tweet ID : `(?:"id":"(.*?)")`
- Tweet Date: `((?:20[1-2][0-9])-(?:0[0-9]|[1][0-2])-(?:[0-2][0-9]|[3][0-1]))` look for the pattern of a specified date only
- these three Regex have returned me three list and now I have matched each list and checked if there are tweets against each id
- then I have created a tuple with first element as the date and followed by id and text of tweet
- Replaced all the special characters with text and changed the encoding of different characters

In [8]:

```
%%time
# Splitting data at {and creating list for each tweet
splitted_tweets=data.split("{")
# Empty list for tweet
list_tup=[]
# empty lisr of ids
list_of_ids=[]
for tweet in splitted_tweets:
    # Date regEx
    date_list=re.findall("((?:20[1-2][0-9])-(?:0[0-9]|[1][0-2])-(?:[0-2][0-9]|[3][0-1]))",t
    #Id RegEx
    id_list=re.findall('(?:"id":"(.*?)")',tweet)
    # List RegEx
    text_list=re.findall('(?:(?:"text":"(.*?)(?:(?:"})|(?:","created_at")|(?:","id"))))',tw
    #checking if there are all the elemeent of the tweet present
    if(len(date_list)!=0 and len(id_list)!=0 and len(text_list)!=0):
        if(id_list[0] not in list_of_ids):
            list_of_ids.append(id_list[0])
            list_tup.append((date_list[0],id_list[0],(eval('"'+text_list[0]+'"').encode('ut
                                                      .decode('utf-16')).replace("<","&lt;"
                             .replace("&","&amp;").replace("'","&apos;").replace('"',"&quot
```

```
Wall time: 7min 39s
```

5.) Filtering English Tweets:

In this step I have Filtered all the tweets that were not of english language as required by the specifications all the tweets are removed using the `langid classifier` function from the langid libarary which we imported previously for this purpose Finally I have appended all the english tweet to a list of tuples that have the Date, Id, tweet text.

Following that I have converted it into the dictionary that has the date as the key to which we need our tweets in i.e for one date all the tweets that has that date. now we have date as key and a List of tuples where each tuple has a tweet id and text at 0 and 1 position.

In [4]:

```
#creating tuple from the above tuple and running langid classifier
# That Classifies the data only on en language and keep it
tup2=list_tup
tt=[]
for i in tup2:
    if langid.classify(i[2])[0]=='en':
        tt.append(i)
```

```
Wall time: 12min 7s
```

In [5]:

```
# Creating dictionary as date as key and value as list of items with id and text
dict_1=dict()

for date,ids,text  in tt:
    dict_1.setdefault(date, []).append((ids,text))
```

```
Wall time: 111 ms
```

6.) Creating XML:

Now using the Dictionary I have made a string which is in the `XML` format by iterating over the dictiary and extracting its key for date and values using the value which are in a list of tuple format and accesing it using the index . appending each value within the string I have created the text format for the for the XML and write the file using the write function.

In [6]:

```
# Formating the string
strin='<?xml version="1.0" encoding="UTF-8"?>'+'\n'+'<data>'
for i,j in dict_1.items():
    strin=strin+'\n'+'<tweets date='+'"'+i+'"'+'>'+'\n'
    for j in dict_1[i]:
        strin=strin+'<tweet id='+'"' +(j[0])+'"'+'>'+j[1]+'</tweet>'+'\n'
strin=strin+'</tweets>'+'\n'+'</data>'
```

```
Wall time: 49min 25s
```

In [7]:

```

# Writing the file
finalxml = open("30823293.xml", "w",encoding='UTF-8')
finalxml.write(strin)
finalxml.close()
```

```
Wall time: 154 ms
```

Conclusion:

After the completion of the task I was able to create an XML file that has all the tweet date, tweets text, tweet id in the required format. From this Task I learnt how to handle unprocessed semi- structured data by exploring it and analysing the pattern using regular expression library that help me identify the pattern of the data and how to extract them with help of regular expression, dictionary in python, and langid package to filter out non english tweets and repeted tweets I was able to exttract and store the data in a structred format and generate the XML. This Task has helped me understand how regEx works and how to extract data and apply exploration on it to get desired format.

Thank You

In [ ]:

```
```