CONFIGURATION MANUAL

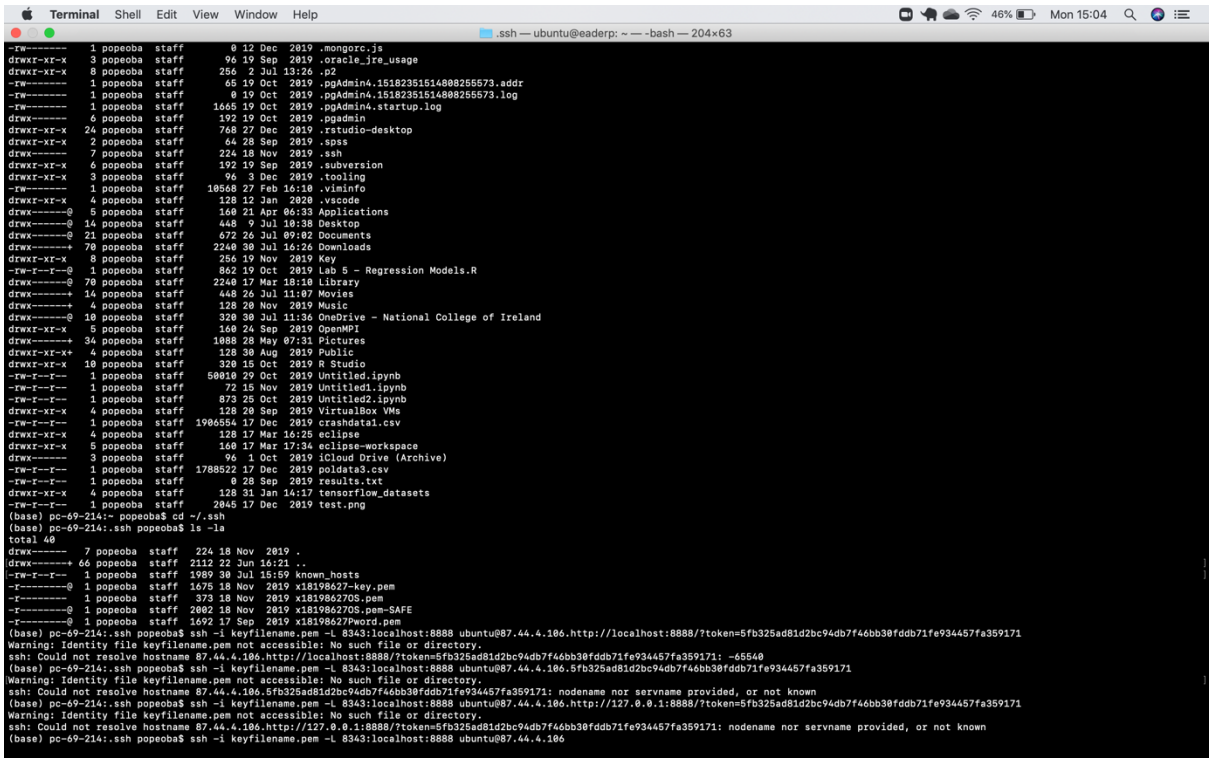
Infectious Disease Surveillance with GLEPI: A Natural Language Processing and Deep Learning System

Adekola Emmanuel
18198627

SOFTWARE SETUP	•OPENSTACK INSTALLATIONS •UBUNTU 18.04 •PYTHON INSTALLATION •ANACONDA NAVIGATOR INSTALLATION
DATA PRE-PROCESSING	•DOWNLOAD NCITR DATASET •FEATURE ENGINEERING (0,1)
MODEL IMPLEMENTATION	•LSTM •CNN •BI-DIRECTIONAL LSTM
TWEET PRE-PROCESSING	•TWEET STREAMING •STOP WORDS REMOVAL AND OTHER TEXTUAL DATA CLEANING •TOKENIZATION, LEMMATIZATION AND STEMMING
PAGE RANK IMPLEMENTATION	•WEB SEARCH FOR RELEVANT NEWS ARTICLES •STOP WORDS REMOVAL AND OTHER TEXTUAL CLEANING •TOKENIZATION, LEMMATIZATION AND STEMMING
MODEL DEPLOYMENT	•BI-LSTM MODELING ON STREAMED TWEETS AND NEWS ARTICLES

STEP 1: SOFTWARE SETUP

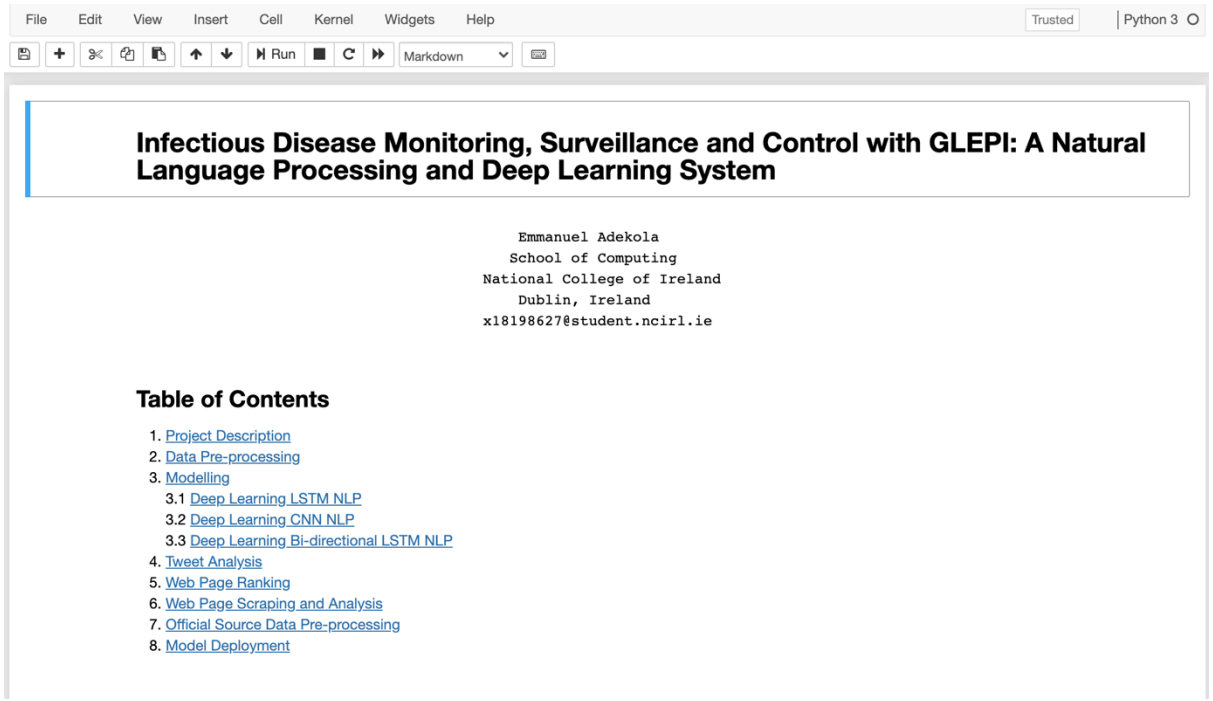
Follow the steps documented on https://youpple.com/dataclergy/2020/08/16/my-project-installations/



STEP 2: DATA PRE-PROCESSING

Download NCITR MedWeb Dataset from http://mednlp.jp/medweb/NTCIR-13/ and place in the same directory with EAde18198627-Codebook.ipynb

Run all blocks of code under section 2 (Data Pre-processing) of EAde18198627-Codebook.ipynb



2. Data Pre-processing

In [116]:
```
#Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [117]:
```
#Load symptoms sentiment dataset and print top 10 rows for verification
MedWebDF = pd.read_csv('Datasets/NTCIR-13_MedWeb_en_training.csv')
MedWebDF.head(10)
```

Out[117]:

	ID	Tweet	Influenza	Diarrhea	Hayfever	Cough	Headache	Fever	Runnynose	Cold
0	1en	The cold makes my whole body weak.	n	n	n	n	n	n	n	p
1	2en	It's been a while since I've had allergy symptoms.	n	n	p	n	n	n	p	n
2	3en	I'm so feverish and out of it because of my allergies. I'm so sleepy.	n	n	p	n	n	p	p	n
3	4en	I took some medicine for my runny nose, but it won't stop.	n	n	n	n	n	n	p	n
4	5en	I had a bad case of diarrhea when I traveled to Nepal.	n	n	n	n	n	n	n	n
5	6en	It takes a millennial wimp to call in sick just because they're coughing. It's always important to go to work, no matter what.	n	n	n	p	n	n	n	n
6	7en	I'm not going today, because my stuffy nose is killing me.	n	n	n	n	n	n	p	n
7	8en	I never thought I would have allergies.	n	n	p	n	n	n	p	n
8	9en	I have a fever but I don't think it's the kind of cold that will make it to my stomach.	n	n	n	n	n	p	n	p
9	10en	My phlegm has blood in it and it's really gross.	n	n	n	p	n	n	n	n

In [118]:
```
#Replace 'n' values with nothing and 'p' values with the column header value
for col in MedWebDF.columns:
    MedWebDF[col].replace({'n':'','p':col},inplace=True)
```

STEP 3: MODEL IMPLEMENTATION

LSTM

Run all blocks of code under section 3.1 (Deep Learning LSTM NLP) of EAde18198627-Codebook.ipynb

3. Modelling

3A. Deep Learning LSTM NLP

In [50]:
```
#Import necessary libraries
import pandas as pd
from termcolor import colored
from sklearn.model_selection import train_test_split

#Define variables
COLUMNS = ['ID', 'Tweet', 'Influenza', 'Diarrhea', 'Hayfever', 'Cough', 'Headache', 'Fever', 'Runnynose', 'Cold', 'Sent

#Read dataset
MedWebDFClean = pd.read_csv('MedWebDF2.csv', names = COLUMNS, encoding = 'latin-1')
print(colored("Columns: {}".format(', '.join(COLUMNS)), "yellow"))

MedWebDFClean = MedWebDFClean.drop(MedWebDFClean.index[MedWebDFClean.Sentiment == 'Sentiment'])
#Remove extra columns
print(colored("Useful columns: Sentiment and Tweet", "yellow"))
print(colored("Removing other columns", "red"))
MedWebDFClean.drop(['ID', 'Influenza', 'Diarrhea', 'Hayfever', 'Cough', 'Headache', 'Fever', 'Runnynose', 'Cold'], axis
print(colored("Columns removed", "red"))

#Train test split
print(colored("Splitting train and test dataset into 80:20", "yellow"))
X_train, X_test, y_train, y_test = train_test_split(MedWebDFClean['Tweet'], MedWebDFClean['Sentiment'], test_size = 0.2
train_MedWebDFClean = pd.DataFrame({
    'Tweet': X_train,
    'Sentiment': y_train
    })
print(colored("Train data distribution:", "yellow"))
print(train_MedWebDFClean['Sentiment'].value_counts())
test_MedWebDFClean = pd.DataFrame({
    'Tweet': X_test,
    'Sentiment': y_test
    })
```

CNN

Download pre-trained GloVe embedding

Run all blocks of code under section 3.2 (Deep Learning CNN NLP) of EAde18198627-Codebook.ipynb

3B. Deep Learning CNN NLP

In [101]:
```
essary libraries
as pd
lor import colored
.model_selection import train_test_split

iables
'ID', 'Tweet', 'Influenza', 'Diarrhea', 'Hayfever', 'Cough', 'Headache', 'Fever', 'Runnynose', 'Cold', 'Sentiment']

et
an = pd.read_csv('MedWebDF2.csv', names = COLUMNS, encoding = 'latin-1')
ed("Columns: {}".format(', '.join(COLUMNS)), "yellow"))


a columns
ed("Useful columns: Sentiment and Tweet", "yellow"))
ed("Removing other columns", "red"))
an.drop(['ID', 'Influenza', 'Diarrhea', 'Hayfever', 'Cough', 'Headache', 'Fever', 'Runnynose', 'Cold'], axis = 1, inplac
an = MedWebDFClean.drop(MedWebDFClean.index[MedWebDFClean.Sentiment == 'Sentiment'])
ed("Columns removed", "red"))
an
```

```
Columns: ID, Tweet, Influenza, Diarrhea, Hayfever, Cough, Headache, Fever, Runnynose, Cold, Sentiment
Useful columns: Sentiment and Tweet
Removing other columns
Columns removed
```

Out[101]:

	Tweet	Sentiment
0.0	The cold makes my whole body weak.	1
1.0	It's been a while since I've had allergy symptoms.	1
2.0	I'm so feverish and out of it because of my allergies. I'm so sleepy.	1
3.0	I took some medicine for my runny nose, but it won't stop.	1
4.0	I had a bad case of diarrhea when I traveled to Nepal	0

Bi-directional LSTM

Run all blocks of code under section 3.3 (Deep Learning Bi-directional LSTM NLP) of EAde18198627-Codebook.ipynb

3C. Deep Learning Bi-directional LSTM NLP

In [4]:
```
#Import necessary libraries
import tensorflow as tf
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from termcolor import colored
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
```

In [5]:
```
et', 'Influenza', 'Diarrhea', 'Hayfever', 'Cough', 'Headache', 'Fever', 'Runnynose', 'Cold', 'Sentiment']


ad_csv('MedWebDF2.csv', names = COLUMNS, encoding = 'latin-1', skiprows=1)
s: {}".format(', '.join(COLUMNS)), "yellow"))


 columns: Sentiment and Tweet", "yellow"))
ng other columns", "red"))
ID', 'Influenza', 'Diarrhea', 'Hayfever', 'Cough', 'Headache', 'Fever', 'Runnynose', 'Cold'], axis = 1, inplace = True)
bDFClean.drop(MedWebDFClean.index[MedWebDFClean.Sentiment == 'Sentiment'])
s removed", "red"))
```

```
Columns: ID, Tweet, Influenza, Diarrhea, Hayfever, Cough, Headache, Fever, Runnynose, Cold, Sentiment
Useful columns: Sentiment and Tweet
Removing other columns
Columns removed

//anaconda3/lib/python3.7/site-packages/pandas/core/ops.py:1649: FutureWarning: elementwise comparison failed; return
ing scalar instead, but in the future will perform elementwise comparison
  result = method(y)
```

Out[5]:

	Tweet	Sentiment

STEP 4: TWEET PRE-PROCESSING

Complete all variables in access.properties file

Run all blocks of code under section 4 (Tweet Analysis) of EAde18198627-Codebook.ipynb

4. Tweet Analysis

In [3]:
```
pip install ConfigParser
```

```
Collecting ConfigParser
  Downloading configparser-5.0.0-py3-none-any.whl (22 kB)
Installing collected packages: ConfigParser
Successfully installed ConfigParser-5.0.0
WARNING: You are using pip version 20.0.2; however, version 20.2.2 is available.
You should consider upgrading via the '//anaconda3/bin/python -m pip install --upgrade pip' command.
Note: you may need to restart the kernel to use updated packages.
```

In [5]:
```
pip install tweepy
```

```
Collecting tweepy
  Downloading tweepy-3.9.0-py2.py3-none-any.whl (30 kB)
Requirement already satisfied: requests[socks]>=2.11.1 in /anaconda3/lib/python3.7/site-packages (from tweepy) (2.22.
0)
Requirement already satisfied: six>=1.10.0 in /anaconda3/lib/python3.7/site-packages (from tweepy) (1.12.0)
Collecting requests-oauthlib>=0.7.0
  Downloading requests_oauthlib-1.3.0-py2.py3-none-any.whl (23 kB)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in /anaconda3/lib/python3.7/site-packages (from requests[socks]>
=2.11.1->tweepy) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in /anaconda3/lib/python3.7/site-packages (from requests[socks]>=2.
11.1->tweepy) (2019.11.28)
Requirement already satisfied: idna<2.9,>=2.5 in /anaconda3/lib/python3.7/site-packages (from requests[socks]>=2.11.1
->tweepy) (2.8)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /anaconda3/lib/python3.7/site-packages (fro
m requests[socks]>=2.11.1->tweepy) (1.24.2)
Requirement already satisfied: PySocks!=1.5.7,>=1.5.6; extra == "socks" in /anaconda3/lib/python3.7/site-packages (fr
om requests[socks]>=2.11.1->tweepy) (1.7.0)
Collecting oauthlib>=3.0.0
  Downloading oauthlib-3.1.0-py2.py3-none-any.whl (147 kB)
     |████████████████████████████████| 147 kB 5.9 MB/s eta 0:00:01
Installing collected packages: oauthlib, requests-oauthlib, tweepy
Successfully installed oauthlib-3.1.0 requests-oauthlib-1.3.0 tweepy-3.9.0
WARNING: You are using pip version 20.0.2; however, version 20.2.2 is available.
```

STEP 5: PAGE RANK IMPLEMENTATION

Run all blocks of code under section 5 and 6 (Web Page Ranking and Web Page Scraping and Analysis) of EAde18198627-Codebook.ipynb

5. Web Page Ranking using Google Search

In [45]:
```
pip install beautifulsoup4
```

```
Requirement already satisfied: beautifulsoup4 in /anaconda3/lib/python3.7/site-packages (4.7.1)
Requirement already satisfied: soupsieve>=1.2 in /anaconda3/lib/python3.7/site-packages (from beautifulsoup4) (1.8)
Note: you may need to restart the kernel to use updated packages.
```

In [46]:
```
pip install google
```

```
Collecting google
  Downloading google-3.0.0-py2.py3-none-any.whl (45 kB)
     |████████████████████████████████| 45 kB 1.6 MB/s eta 0:00:011
Requirement already satisfied: beautifulsoup4 in /anaconda3/lib/python3.7/site-packages (from google) (4.7.1)
Requirement already satisfied: soupsieve>=1.2 in /anaconda3/lib/python3.7/site-packages (from beautifulsoup4->google)
(1.8)
Installing collected packages: google
Successfully installed google-3.0.0
Note: you may need to restart the kernel to use updated packages.
```

In [698]:
```
try:
    from googlesearch import search
except ImportError:
    print("No module named 'google' found")

query = "fever, chills, cough, fatigue, shortness of breathe, muscle ache, body ache, headache, loss of taste, loss of

for j in search(query, tld="com", num=10, stop=10, pause=2):
    print(j)
```

```
https://www.hollandhospital.org/healthylife/healthy-life-blogs/is_it_covid19_the_flu_or_just_a_cold_211
https://www.uwhealth.org/flu/know-the-difference-between-a-cold-and-the-flu/10376
https://www2.hse.ie/conditions/coronavirus/symptoms.html
https://www2.hse.ie/conditions/shortness-of-breath.html
https://www2.hse.ie/conditions/hay-fever.html
```

6. Web Page Scraping and Analysis

In [92]:
```
from datetime import date
today = date.today()

d = today.strftime("%m-%d-%y")
#d = today.strftime("%Y/%m/%d")
print("date =", d)
```

```
date = 09-03-20
```

In [93]:
```
from datetime import date
from datetime import timedelta
today = date.today()
yesterday = today - timedelta(days=1)
d = yesterday.strftime("%m-%d-%y")

print("date =", d)
```

```
date = 09-02-20
```

In [94]:
```
cnn_url="https://www.cnn.com/world/live-news/coronavirus-pandemic-{}-intl/index.html".format(d)
#euc_url="https://covidnews.eurocities.eu/{}/".format(d)
```

In [95]:
```
#print(euc_url)
print(cnn_url)
```

```
https://www.cnn.com/world/live-news/coronavirus-pandemic-09-02-20-intl/index.html
```

In [96]:
```
from bs4 import BeautifulSoup
import requests
```

In [97]:
```
html = requests.get(cnn_url).text
```

In [135]:
```
soup = BeautifulSoup(html)
```

STEP 7: MODEL DEPLOYMENT

Run all blocks of code under section 8 (Model Deployment) of EAde18198627-Codebook.ipynb

8. Model Deployment

In [114]:
```
# Use the model to predict text sentiment
combined_corpus = pd.read_csv('FinalTweets.csv')
corpus = combined_corpus['text']
print(corpus)
```

```
0                                United Kingdom Daily Coronavirus (COVID-19) Report. #coronavirus #UK #Corona #covid
19 #Covid19UK
1                                United Kingdom Daily Coronavirus (COVID-19) Report. #coronavirus #UK #Corona #covid
19 #Covid19UK
2               We're using CDC Data to visualize the impact of COVID at a national and local level. On the national
level, it se…
3                The Latest Coronavirus statistics are as follows:  Tweeted: 2020-08-15 15:07:22.920608 Confirmed: 21
162299 Death…
4                                          #EDUCATION: When can #kids return to #sports after recovering from #COVID1
9?#parenting…
5       Researchers say high rates of obesity in the United States could make a COVID-19 vaccine less effective here.
#coronavirus…
6                                The government is scrambling to find much-needed beds for the infected, but is gett
ing no help  …
7             Daily dose of #COVID19 reality from #Nebraska. The media, Lancaster County Health Dept and the mayor of
#Lincoln co…
8                                The government is scrambling to find much-needed beds for the infected, but is gett
ing no help  …
9                                United Kingdom Daily Coronavirus (COVID-19) Report. #coronavirus #UK #Corona #covid
```

In [136]:
```
#Create the sequences
padding_type='post'
sample_sequences = tokenizer.texts_to_sequences(corpus)
corpus_padded = pad_sequences(sample_sequences, padding=padding_type, maxlen=max_length)
```

In [137]:
```
print(corpus_padded)
```

```
[[ 717    0    0 ...    0    0    0]
```