

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/>

```
Terminal Shell Edit View Window Help
ssh - ubuntu@eaderp: ~ - bash - 204x63

-rw-r--r-- 1 popeoba staff  812 Dec 2019 .mongorc.js
drwxr-xr-x 3 popeoba staff  96 19 Sep 2019 .oracle_jre_usage
-rw-r--r-- 1 popeoba staff  266 2 Jul 13:26 .p2
-rw-r--r-- 1 popeoba staff  65 19 Oct 2019 .pgAdmin4.15182351514808265573_addr
-rw-r--r-- 1 popeoba staff  9 19 Oct 2019 .pgAdmin4.15182351514808265573_log
-rw-r--r-- 1 popeoba staff 1665 19 Oct 2019 .pgAdmin4.startup.log
drwxr-xr-x 6 popeoba staff 192 19 Oct 2019 .pgadmin
drwxr-xr-x 24 popeoba staff 768 27 Dec 2019 .studio-desktop
drwxr-xr-x 2 popeoba staff  64 28 Sep 2019 .spss
drwxr-xr-x 7 popeoba staff 224 18 Nov 2019 .ssh
drwxr-xr-x 6 popeoba staff 192 19 Sep 2019 .subversion
drwxr-xr-x 3 popeoba staff  96 3 Dec 2019 .tooling
-rw-r--r-- 1 popeoba staff 18568 27 Feb 16:10 .viminfo
drwxr-xr-x 4 popeoba staff 128 12 Jan 2020 .vscode
drwxr-xr-x 5 popeoba staff 168 21 Apr 06:33 Applications
drwxr-xr-x 14 popeoba staff 448 9 Jul 10:38 Desktop
drwxr-xr-x 21 popeoba staff 672 26 Jul 09:02 Documents
drwxr-xr-x 78 popeoba staff 2240 30 Jul 16:26 Downloads
drwxr-xr-x 8 popeoba staff 256 19 Nov 2019 Key
-rw-r--r-- 1 popeoba staff 862 19 Oct 2019 Lab 5 - Regression Models.R
drwxr-xr-x 78 popeoba staff 2240 17 Mar 18:10 Library
drwxr-xr-x 14 popeoba staff 448 26 Jul 11:07 Movies
drwxr-xr-x 4 popeoba staff 128 20 Nov 2019 Music
drwxr-xr-x 10 popeoba staff 320 30 Jul 11:36 OneDrive - National College of Ireland
drwxr-xr-x 5 popeoba staff 160 24 Sep 2019 OpenMPI
drwxr-xr-x 34 popeoba staff 1088 23 May 07:31 Pictures
drwxr-xr-x 4 popeoba staff 128 30 Aug 2019 Public
drwxr-xr-x 10 popeoba staff 320 15 Oct 2019 R Studio
-rw-r--r-- 1 popeoba staff 58018 29 Oct 2019 Untitled.ipynb
-rw-r--r-- 1 popeoba staff 119 29 Nov 2019 Untitled1.ipynb
-rw-r--r-- 1 popeoba staff 873 25 Oct 2019 Untitled2.ipynb
drwxr-xr-x 4 popeoba staff 128 20 Sep 2019 VirtualBox VMs
-rw-r--r-- 1 popeoba staff 1906504 17 Dec 2019 crashdata1.csv
drwxr-xr-x 4 popeoba staff 128 17 Mar 16:25 eclipse
drwxr-xr-x 5 popeoba staff 160 17 Mar 17:34 eclipse-workspace
drwxr-xr-x 3 popeoba staff 96 1 Oct 2019 iCloud Drive (Archive)
-rw-r--r-- 1 popeoba staff 1788522 17 Dec 2019 polysta3.csv
-rw-r--r-- 1 popeoba staff  9 28 Sep 2019 results.txt
drwxr-xr-x 4 popeoba staff 128 31 Jan 14:17 tensorflow_datasets
-rw-r--r-- 1 popeoba staff 2045 17 Dec 2019 test.png
(base) pc-69-214:~$ popoeba$ cd ~/ssh
(base) pc-69-214:~/ssh$ popoeba$ ls -la
total 40
drwxr-xr-x 7 popeoba staff 224 18 Nov 2019 .
drwxr-xr-x 60 popeoba staff 2112 12 Jun 14:21 ..
-rw-r--r-- 1 popeoba staff 1989 30 Jul 15:59 known_hosts
-rw-r--r-- 1 popeoba staff 1675 18 Nov 2019 x18198627-key.pem
-rw-r--r-- 1 popeoba staff 873 18 Nov 2019 x1819862703.pem
-rw-r--r-- 1 popeoba staff 2892 18 Nov 2019 x1819862705.pem-SAFE
-rw-r--r-- 1 popeoba staff 1692 17 Sep 2019 x18198627word.pem
(base) pc-69-214:~/ssh$ popoeba$ ssh -l keyfilename.pem -L 8343:localhost:8888 ubuntu@87.44.4.106.http://localhost:8888/?token=5fb325ad81d2bc94db7f46bb38fddb71fe934457fa359171
Warning: Identity file keyfilename.pem not accessible: No such file or directory.
ssh: Could not resolve hostname 87.44.4.106.http://localhost:8888/?token=5fb325ad81d2bc94db7f46bb38fddb71fe934457fa359171: -65548
(base) pc-69-214:~/ssh$ popoeba$ ssh -l keyfilename.pem -L 8343:localhost:8888 ubuntu@87.44.4.106.5fb325ad81d2bc94db7f46bb38fddb71fe934457fa359171
Warning: Identity file keyfilename.pem not accessible: No such file or directory.
ssh: Could not resolve hostname 87.44.4.106.http://127.0.0.1:8888/?token=5fb325ad81d2bc94db7f46bb38fddb71fe934457fa359171: nodename nor servname provided, or not known
Warning: Identity file keyfilename.pem not accessible: No such file or directory.
ssh: Could not resolve hostname 87.44.4.106.http://127.0.0.1:8888/?token=5fb325ad81d2bc94db7f46bb38fddb71fe934457fa359171: nodename nor servname provided, or not known
(base) pc-69-214:~/ssh$ popoeba$ ssh -l keyfilename.pem -L 8343:localhost:8888 ubuntu@87.44.4.106
```

```
Last login: Fri Aug 14 20:23:19 on ttys008
~/anaconda3/bin/jupyter_mac_command ; exit;
(base) Emmanuel1s-MacBook-Pro:~$ popoeba$ /anaconda3/bin/jupyter_mac_command ; exit;
[ I 21:16:07.173 NotebookApp] Jupyterlab extension loaded from //anaconda3/lib/python3.7/site-packages/jupyterlab
[ I 21:16:07.173 NotebookApp] Jupyterlab application directory is //anaconda3/share/jupyter/lab
[ I 21:16:07.176 NotebookApp] Serving notebooks from local directory: /Users/popeoba
[ I 21:16:07.176 NotebookApp] The Jupyter Notebook is running at:
[ I 21:16:07.176 NotebookApp] http://localhost:8888/?token=97c79f011add212be064f974f99b3e1ebfba2e0dda2650f
[ I 21:16:07.176 NotebookApp] or http://127.0.0.1:8888/?token=97c79f011add212be064f974f99b3e1ebfba2e0dda2650f
[ I 21:16:07.176 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 21:16:07.198 NotebookApp]

To access the notebook, open this file in a browser:
file:///Users/popeoba/Library/Jupyter/runtime/nbserver-29881-open.html
Or copy and paste one of these URLs:
http://localhost:8888/?token=97c79f011add212be064f974f99b3e1ebfba2e0dda2650f
or http://127.0.0.1:8888/?token=97c79f011add212be064f974f99b3e1ebfba2e0dda2650f
[ E 21:16:08.567 NotebookApp] Could not open static file
[W 21:16:08.659 NotebookApp] 404 GET /static/components/react/react-dom.production.min.js (:::1) 1.22ms referer=http://localhost:8888/tree?token=97c79f011add212be064f974f99b3e1ebfba2e0dda2650f
[W 21:16:08.788 NotebookApp] 404 GET /static/components/react/react-dom.production.min.js (:::1) 1.83ms referer=http://localhost:8888/tree?token=97c79f011add212be064f974f99b3e1ebfba2e0dda2650f
[W 21:16:54.728 NotebookApp] 404 GET /static/components/react/react-dom.production.min.js (:::1) 1.93ms referer=http://localhost:8888/notebooks/Desktop/NCI/Projects/ResearchM20Project/Codek20Files/EAd1819827-Codebook.ipynb
[W 21:16:54.788 NotebookApp] 404 GET /static/components/react/react-dom.production.min.js (:::1) 1.96ms referer=http://localhost:8888/notebooks/Desktop/NCI/Projects/ResearchM20Project/Codek20Files/EAd1819827-Codebook.ipynb
[ I 21:17:01.795 NotebookApp] Kernel started: a7d83c9d-2327-4a7a-9a00-d997c3039801
[ I 21:17:02.763 NotebookApp] Adapting from protocol version 5.1 (kernel a7d83c9d-2327-4a7a-9a00-d997c3039801) to 5.3 (client).
```

STEP 2: DATA PRE-PROCESSING

Download NCITR MedWeb Dataset from <http://mednlp.jp/medweb/NTCIR-13/> and place in the same directory with EAd18198627-Codebook.ipynb

Run all blocks of code under section 2 (Data Pre-processing) of EAd18198627-Codebook.ipynb

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

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

Emmanuel Adekola
School of Computing
National College of Ireland
Dublin, Ireland
x18198627@student.ncirl.ie

Table of Contents

- [Project Description](#)
- [Data Pre-processing](#)
- [Modelling](#)
 - [Deep Learning LSTM NLP](#)
 - [Deep Learning CNN NLP](#)
 - [Deep Learning Bi-directional LSTM NLP](#)
- [Tweet Analysis](#)
- [Web Page Ranking](#)
- [Web Page Scraping and Analysis](#)
- [Official Source Data Pre-processing](#)
- [Model Deployment](#)

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 EAd18198627-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', 'Runny nose', 'Cold', 'Sentiment']

#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', 'Runny nose', 'Cold'], axis=1, inplace=True)
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 EAd18198627-Codebook.ipynb

3B. Deep Learning CNN NLP

```
In [101]: #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', 'Runny nose', 'Cold', 'Sentiment']

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

#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', 'Runny nose', 'Cold'], axis=1, inplace=True)
print(colored("Columns removed", "red"))
```

```
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 EAd18198627-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.euocities.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 EAd18198627-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
10 #Covid19UK
```

```
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]
```