**Project Closure Report**

Luffa Political Database and Analysis – backend development project

Date: Jul, 05, 2019



**Table of Contents**

[1. Project Closure Summary 3](#_Toc13232137)

[2.1 Project Functionality Overview 3](#_Toc13232140)

[2.3 Project Closure Synopsis 3](#_Toc13232141)

[2. Project Artifacts 4](#_Toc13232142)

[3.1 Project logs 4](#_Toc13232143)

# Project Closure Summary



## 1.1 Project Functionality Overview

The project delivers political databases including ones for articles, keywords and others to support Luffa’s consulting work on Myanmar’s Peace process. Various scripts and algorithms perform web scraping, extract data from a few websites using automated artificial intelligence models. There after apply tagging and topic model to help categorization of information from media and social media articles and feed the information in the database automatically using the model to create a few visualizations.

### Database details

1. Articles\_database\_unique1: This database consists of all the web scrapped data scrapped from the given websites.
2. Keyword\_db: This database consists of all the keyword extracted with model
3. Other\_database: This database consists of all the stakeholder person name, organization name, location, topics, keywords, topic score, tags, weblink and full text.

## 1.2 Project Closure Synopsis

Following is a list of scripts used for delivering the above stated project functionality:

1. Model for this project has been deployed on Luffa’s developer server (IP Address - 34.93.121.44) under "/home/developer/Luffa\_final\_03\_07\_2019" path. Scripts listed below are stored in the same location.
2. Command to execution the model : "python3.4 wrapper.py"

|  |  |  |
| --- | --- | --- |
| **Sr. no.** | **Python Script/ File Name** | **Description/ details** |
| 1 | article\_db | Uploading extracted data in article\_title,publication\_date,author,full\_text,weblink,source,news\_category,media\_links, input\_time and media\_topic\_tag in article\_database\_unique1. |
| 2 | bbc\_main | Integrated articles, categories and web scrapping data extraction scripts. |
| 3 | bnionline\_articles | Extracted all the articles from the web-source |
| 4 | bnionline\_data | Extracted all the required data from the articles |
| 5 | bnionline\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 6 | browser\_setting | Open selenium browser for web scrapping |
| 7 | burmairrawaddy\_articles | Extracted all the articles from the web-source |
| 8 | burmairrawaddy\_categories | Extracted all the categories from the web-source |
| 9 | burmairrawaddy\_data | Extracted all the required data from the articles |
| 10 | burmairrawaddy\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 11 | burmese\_dvb | Integrated articles , categories and web scrapping data extraction scripts. |
| 12 | burmese\_monnews | Integrated articles , categories and web scrapping data extraction scripts. |
| 13 | burmese\_narinjara | Integrated articles , categories and web scrapping data extraction scripts. |
| 14 | burmese\_shannews | Integrated articles , categories and web scrapping data extraction scripts. |
| 15 | chunkers | Chunking model to extract relevant tokens in data. |
| 16 | cincds\_articles | Extracted all the articles from the web-source |
| 17 | cincds\_data | Extracted all the required data from the articles |
| 18 | cincds\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 19 | data\_cleaning\_model | Apply all preprocessing steps for the text and apply this model before topic extraction , stakeholder name , position , location and sentiment analysis. |
| 20 | daydaily\_articles | Extracted all the articles from the web-source |
| 21 | daydaily\_categories | Extracted all the categories from the web-source |
| 22 | daydaily\_data | Extracted all the required data from the articles |
| 23 | daydaily\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 24 | elevenmyanmar\_articles | Extracted all the articles from the web-source |
| 25 | elevenmyanmar\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 26 | Facebook\_login | Login access for Facebook page. |
| 27 | Facebook\_mba\_gov\_main | Integrated Facebook posts and web scrapping data extraction scripts. |
| 28 | fb\_hlamaung | Integrated Facebook posts and web scrapping data extraction scripts. |
| 29 | fb\_JointCeasefireMonitoringCommittee | Integrated Facebook posts and web scrapping data extraction scripts. |
| 30 | fb\_mba | Integrated Facebook posts and web scrapping data extraction scripts. |
| 31 | fb\_ministry | Integrated Facebook posts and web scrapping data extraction scripts. |
| 32 | fb\_myanmarpresidentoffice | Integrated Facebook posts and web scrapping data extraction scripts. |
| 33 | fb\_nca\_s\_eao | Integrated Facebook posts and web scrapping data extraction scripts. |
| 34 | fb\_nrpc | Integrated Facebook posts and web scrapping data extraction scripts. |
| 35 | fb\_state\_cousellor | Integrated Facebook posts and web scrapping data extraction scripts. |
| 36 | fb\_TaiFreedomBurmese | Integrated Facebook posts and web scrapping data extraction scripts. |
| 37 | frontiermyanmar\_articles | Extracted all the articles from the web-source |
| 38 | frontiermyanmar\_categories | Extracted all the categories from the web-source |
| 39 | frontiermyanmar\_data | Extracted all the required data from the articles |
| 40 | frontiermyanmar\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 41 | google\_news | Integrated articles , categories and web scrapping data extraction scripts. |
| 42 | hintharmedia | Integrated articles , categories and web scrapping data extraction scripts. |
| 43 | irrawaddy\_articles | Extracted all the articles from the web-source |
| 44 | irrawaddy\_categories | Extracted all the categories from the web-source |
| 45 | irrawaddy\_data | Extracted all the required data from the articles |
| 46 | irrawaddy\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 47 | kachinlandnews | Integrated articles , categories and web scrapping data extraction scripts. |
| 48 | keyword\_list | Apply lemmatization , stemming and data cleaning model. Apply pos tagging method with spacy to identify and extract all keywords. |
| 49 | kicnews\_articles | Extracted all the articles from the web-source |
| 50 | kicnews\_categories | Extracted all the categories from the web-source |
| 51 | kicnews\_data | Extracted all the required data from the articles |
| 52 | kicnews\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 53 | location | Extract location from full-text with part-of-speech tagging and geo text and geography libraries. |
| 54 | location\_stakeholder\_name | Extract stakeholder person position and group name by LDA tagging model |
| 55 | mizzima\_articles | Extracted all the articles from the web-source |
| 56 | mizzima\_categories | Extracted all the categories from the web-source |
| 57 | mizzima\_data | Extracted all the required data from the articles |
| 58 | mizzima\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 59 | mmtimes\_articles | Extracted all the articles from the web-source |
| 60 | mmtimes\_articles\_mm | Extracted all the articles from the web-source |
| 61 | mmtimes\_categories\_mm | Extracted all the categories from the web-source |
| 62 | mmtimes\_data | Extracted all the required data from the articles |
| 63 | mmtimes\_data\_mm | Extracted all the required data from the articles |
| 64 | mmtimes\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 65 | mmtimes\_main\_mm | Integrated articles , categories and web scrapping data extraction scripts. |
| 66 | model\_for\_political\_database\_elevenmyanmar | Integrated articles , categories and web scrapping data extraction scripts. |
| 67 | model\_for\_political\_database\_globalnewlightofmynamar | Integrated articles , categories and web scrapping data extraction scripts. |
| 68 | model\_for\_political\_database\_twitter | Integrated articles , categories and web scrapping data extraction scripts. |
| 69 | news-eleven | Integrated articles , categories and web scrapping data extraction scripts. |
| 70 | nmg\_articles | Extracted all the articles from the web-source |
| 71 | nmg\_categories | Extracted all the categories from the web-source |
| 72 | nmg\_data | Extracted all the required data from the articles |
| 73 | nmg\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 74 | rfa\_articles | Extracted all the articles from the web-source |
| 75 | rfa\_articles\_mm | Extracted all the articles from the web-source |
| 76 | rfa\_categories | Extracted all the categories from the web-source |
| 77 | rfa\_categories\_mm | Extracted all the categories from the web-source |
| 78 | rfa\_data | Extracted all the required data from the articles |
| 79 | rfa\_data\_mm | Extracted all the required data from the articles |
| 80 | rfa\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 81 | rfa\_main\_mm | Integrated articles , categories and web scrapping data extraction scripts. |
| 82 | seniorgeneralminaunghlaing\_articles | Extracted all the articles from the web-source |
| 83 | seniorgeneralminaunghlaing\_categories | Extracted all the categories from the web-source |
| 84 | seniorgeneralminaunghlaing\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 85 | sentiment\_sklearn | Extract sentiments and categories with positive and negative through tfidf algorithm and LDA model |
| 86 | sentiment\_spacy | Extract sentiments and categories with positive and negative through tfidf algorithm and spacy model |
| 87 | sentiments | Extract sentiments and categories with positive and negative through tfidf algorithm and pos model |
| 88 | stakeholder\_name | Extract stakeholder person name and group name by LDA tagging model |
| 89 | stakeholder\_position | Extract stakeholder person position and group name position by LDA tagging model |
| 90 | stakeholders | Extract stakeholder group name by LDA tagging model |
| 91 | tf-idf | Apply tfidf algorithm with preprocessing steps |
| 92 | tf-idf\_basic | Apply basic tfidf algorithm |
| 93 | tf-idf\_feature\_extraction | Apply tfidf for features extraction model |
| 94 | thanlwintimes | Integrated articles , categories and web scrapping data extraction scripts. |
| 95 | topic\_tags | Apply LDA, scikitlearn and data preprocessing, genism, keras, tensorflow and word embedding matrix , word cloud model for topics |
| 96 | upload\_data\_others | To upload extracted data in article\_title, stakeholder\_person\_name, stakeholder\_org\_name, location, location\_country, location\_cities, topics, keywords, topic\_and\_score, tags, news\_category, weblink, full\_text, sentiments in article\_database\_unique1. |
| 97 | voa\_news\_burmese\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 98 | voa\_news\_myanmar\_main | Integrated articles , categories and web scrapping data extraction scripts. |
| 99 | voanews\_articles | Extracted all the articles from the web-source |
| 100 | voanews\_categories | Extracted all the categories from the web-source |
| 101 | voanews\_data | Extracted all the required data from the articles |
| 102 | wrapper | Integrate all the scripts in this wrapper |

**Models Applied:**

1. Applied web scrapping model for all the given websites.
2. Applied model for Stakeholder name with their respective position and Stakeholder organization name.
3. Applied model for location with their respective city and country.
4. Applied topic modelling for finding popular topics, keywords and topic score.
5. Applied tagging model for finding relevant tags.
6. Applied sentiment analysis for news articles.

**Python Libraries Used:**

* Pandas
* Numpy
* Spacy
* Nltk
* Genism
* Beautifulsoup
* Selenium
* Re

**List of web source details:**

1. 'http://burmese.dvb.no/'
2. 'http://burmese.monnews.org/'
3. 'http://cincds.gov.mm/recentest-news'
4. 'http://kicnews.org/'
5. 'http://thanlwintimes.com/'
6. 'http://www.7daydaily.com/'
7. 'http://www.globalnewlightofmyanmar.com'
8. 'http://www.kachinlandnews.com'
9. 'http://www.mizzima.com/'
10. 'http://www.mizzimaburmese.com/'
11. 'http://www.mmtimes.com/'
12. 'http://www.nmg-news.com/'
13. 'https://burma.irrawaddy.com'
14. 'https://burmese.narinjara.com/news/'
15. 'https://burmese.shannews.org/'
16. 'https://burmese.voanews.com'
17. 'https://elevenmyanmar.com'
18. 'https://frontiermyanmar.net/en/'
19. 'https://hintharmedia.com'
20. 'https://khonumthung.org/'
21. 'https://myanmar.mmtimes.com'
22. 'https://news-eleven.com/'
23. 'https://news.google.com'
24. 'https://twitter.com/arsa\_official?lang=en'
25. 'https://www.bbc.com/burmese/'
26. 'https://www.bnionline.net/en/news'
27. 'https://www.facebook.com/hlamaung.shwe'
28. 'https://www.facebook.com/JointCeasefireMonitoringCommittee/'
29. 'https://www.facebook.com/myanmarpresidentoffice.gov.mm/'
30. 'https://www.facebook.com/ncasignatoryeaoofficial/'
31. 'https://www.facebook.com/nrpc.myanmar/'
32. 'https://www.facebook.com/pages/category/Government-Organization/Ministry-of-Ethnic-Affairs-Myanmar-822764397855686/'
33. 'https://www.facebook.com/state.counsellor/'
34. 'https://www.facebook.com/TaiFreedomBurmese/'
35. 'https://www.facebook.com/www.mba.gov.mm/'
36. 'https://www.irrawaddy.com'
37. 'https://www.kantarawaddytimes.org/'
38. 'https://www.rfa.org/burmese/'
39. 'https://www.rfa.org/english/news/myanmar'
40. 'https://www.seniorgeneralminaunghlaing.com.mm/en/Home/news/'
41. 'https://www.seniorgeneralminaunghlaing.com.mm/Home/news/'
42. 'https://www.voanews.com/search?search\_api\_fulltext=burma'
43. 'https://www.voanews.com/search?search\_api\_fulltext=myanmar'

# Project Artifacts

## Project logs

1. Political Database Project - Issue Tracker Jul 05, 2019.xls 🡪 
2. Political Database Project - Progress Tracker - Jul 05, 2019.xls 🡪 

**3.instruction to run the scheduler every day the all the script including scraper, others\_database and Visualization**

1. Install following package in python3.4

“pip install python-crontab”

1. Execute following command **once** to set the schedule job for 8 AM every day

0 8 \* \* \* python3.4 Luffa\_final\_03\_07\_2019/wrapper.py

1. To run the others\_database separetely

“python3.4 Luffa\_final\_03\_07\_2019/upload\_data\_others.py”

1. To run the visualization separetely

“python3.4 Luffa\_final\_03\_07\_2019/visualization\_model.py”

4. **To export a database to JSON format**

1. Login to MySQL workbench

2. execute the following query

“select \* FROM political\_database.articles\_database\_unique1”

1. Click on export and select JSON file format option and save the file