**INSTRUCTIONS**

1. **APPROACHING THE PROBLEM**

The problem assigned can be broken down into two separate parts:

1. **EXTRACTING THE TEXT FROM URLS**

The text through articles from given urls was extracted from web through python library beautifulsoup , specifically designed for the purpose of web scraping.

The combination of reuqests and beautifulsoup module was applied to extract data from all the given urls iteratively.

The data scraped was stored in files with same name as url\_id.

1. **TEXT ANALYSIS**
2. **PRE-PROCESSING AND CREATION OF CLEAN FILES**

The objective here was achieved through tokenizing the text of each file with help of nltk module and creating a set of stop words from each file iteratively and creating clean files through matching tokens from element in the set.

Each clean word was subsequently written to a separate clean file with same url\_id.

1. **CREATION OF POSITIVE AND NEGATIVE WORDS DICTIONARIES**

Positive and negative words dictionaries were created using the words present in the given files through application of set data structures.

1. **TEXT ANALYSIS VARIABLES**

Each variable provided for analysis was processed through the logic given according to the instructed logic in Input.docx

1. **STORING THE RESULTS**

The results for each variable was stored in a pandas dataframe which was later converted into an excel file for output and processed according to given data structure.

1. **HOW TO EXECUTE THE SCRIPT**

The code is divided into two parts : scrape.py and text\_analysis.py

The codes can be executed in vs code terminal as well as google colab or jupyter environment.

The scrape.py code extracts and creates text files for article data while text\_analysis.py generates

results for the analysis variables.

text\_analysis.py was written in colab environment, hence the files it uses are all uploaded though google drive.

Changing file paths as required is necessary to execute the program as the files used in the program are uploaded through google drive.

1. **REQUIREMENTS**
2. **MODULES/LIBRARIES**
3. For Web Scraping

Requests

bs4(BeautifulSoup)

1. For text analysis and file management

pandas

os

nltk

textblob

codecs (for file encoding)

string

re (for regex)

1. **PACKAGES**

nltk.download('punkt')

nltk.download('words')

nltk.download('stopwords')