Instructions

Phase – 1

Extracting Data from given URLs.

- Data extraction code are written in python.
- In code all lines are explained with comment out.
- Used python library is pandas, requests, BeautifulSoup.
- Textual data are stored in their URL name with txt extension, but I removed the '/' and ':' from URL because txt file name does not contain the '/'.
- All txt files are stored in one file which name is **articles**.

Phase – 2

Perform textual analysis and compute variables.

- Textual analysis and compute variables code are written in python.
- In code all lines are explained with comment out.
- Used python library is os, json, nltk, textblob.
- All data is stored in json format.
- File name is **Out_put.json**

Phase – 3

Data formatting as required like Output Data Structure.

- Used Jupyter Notebook to open the Out_put.json file.
- Perform some transformation on json file and save it to xlsx format.
- Jupyter Notebook file are provided which name is **Data_formatting.**
- At last in xlsx file in column URL I removed the extensions txt and add the symbols '/' and ':' which is make it again URL.

Files are provided –

1- articles (File folder) 2- Data_formatting (jupyter notebook) 3- Error (PNG file)
4- Out_put.json 5- Output_Data_Structure.xlsx 6- url_text_extraction_code.py
7- text_analysis_code.py