# How I approached the solution:

# 1. I analysed the Input.xlsx as mentioned in the Objective and Text Analysis documents.

# 2. I used openpyxl for fetching data from Input.xlsx.

# 3. I used requests to fetch webpage and beautifulsoup4 to scrap the data. I found out that some webpages has different tag instead entry-title tag to retrieve blog title. I have handled the situation.

# 4. I carefully followed the Text Analysis documnent and completed text analysis using nltk and given files in the MasterDictionary and Stopwords directories using glob to fetch files.

# 5. Again with openpyxl I iteratively stored the data into a new file Output\_AnuraagaNath.xlsx with the recommended format in Output Data Structure.xlsx

# How to run the .py file:

# 1. I used conda for setting up environment

# 2. This file may be used to create an environment using:

# $ conda create --name <env> --file <this file>

# platform: linux-64

\_libgcc\_mutex=0.1=main

\_openmp\_mutex=5.1=1\_gnu

beautifulsoup4=4.12.3=pypi\_0

bzip2=1.0.8=h7b6447c\_0

ca-certificates=2023.12.12=h06a4308\_0

certifi=2023.11.17=pypi\_0

charset-normalizer=3.3.2=pypi\_0

click=8.1.7=pypi\_0

et-xmlfile=1.1.0=pypi\_0

expat=2.5.0=h6a678d5\_0

idna=3.6=pypi\_0

joblib=1.3.2=pypi\_0

ld\_impl\_linux-64=2.38=h1181459\_1

libffi=3.4.4=h6a678d5\_0

libgcc-ng=11.2.0=h1234567\_1

libgomp=11.2.0=h1234567\_1

libstdcxx-ng=11.2.0=h1234567\_1

libuuid=1.41.5=h5eee18b\_0

ncurses=6.4=h6a678d5\_0

nltk=3.8.1=pypi\_0

openpyxl=3.1.2=pypi\_0

openssl=3.0.12=h7f8727e\_0

pip=23.3.1=py312h06a4308\_0

python=3.12.0=h996f2a0\_0

readline=8.2=h5eee18b\_0

regex=2023.12.25=pypi\_0

requests=2.31.0=pypi\_0

setuptools=68.2.2=py312h06a4308\_0

soupsieve=2.5=pypi\_0

sqlite=3.41.2=h5eee18b\_0

tk=8.6.12=h1ccaba5\_0

tqdm=4.66.1=pypi\_0

tzdata=2023d=h04d1e81\_0

urllib3=2.1.0=pypi\_0

wheel=0.41.2=py312h06a4308\_0

xz=5.4.5=h5eee18b\_0

zlib=1.2.13=h5eee18b\_0

# 3. Activate the conda environment: conda activate <env>

# 4. Run the .py file: python3 -m assignment\_AnuraagaNath

# 5. Remember: Input.xlsx file should be in the same path, that of the .py file to execute. The .py file will automatically create and fill the values creating an output file Output\_AnuraagaNath.xlsx