**Tejas Hirurkar**

**Data Extraction and NLP**

**Test Assignment**

# **How I Approached the solution Step-by-Step**

1. **Understanding the requirements:**

I first gone through all the files especially ‘Objective.docx’ so that I can understand the requirements fully. And also explore other files as well to get complete clarity of all the files. When I understood the requirements of this assignment and what includes in ‘MasterDictionary’, ‘StopWords’ folders and Input.xlsx, Output Data Structure.xlsx files then I started thinking the technical approach that I should follow

1. **Choosing Software and Libraries:**

As I have to complete this assignment in ‘Python’ Language, I decided to go with ‘Jupyter Notebook’.

I choosed some libraries that were necessary like ‘BeautifulSoup’ for web scraping. I choosed ‘BeautifulSoup’ and not others because of the familiarity with the library.

Then:

‘requests’ – for making HTTP requests to fetch web pages

‘nltk’ – for text processing such as tokenization, stopwords removal, and sentiment analysis

‘pandas’ – for data manipulation and handling structured data like DataFrames

‘openpyxl – for reading and writing Excel files

1. **Now its time for Data Extraction:**

I Implemented a function called ‘extract\_text’ to fetch and extract article content from URLs using ‘requests’ and ‘BeautifulSoup’ and saved each article’s content into separate text files named after their URL\_IDs for further analysis

1. **Text Analysis:**

Then I implemented a function called ‘analyze\_text’ to compute various metrics such as positive\_score, negative\_score and so on with respective metrics that were in ‘Output Data Structure.xlsx’ and then I applied NLP techniques to tokenize text, filter stopwords, and calculate specified metrics

1. **Data Aggregation:**

Iterated through the extracted text files and applied the ‘analyze\_text’ function to each file’s content and then aggregated the computed metrics into a structured DataFrame ‘output\_df’

1. **Output File Generation:**

Created an Excel file ‘Output Data Structure.xlsx’ to store the aggregated results in a specified format with columns corresponding to each metric and rows representing each URL\_ID

# **How to run ‘NLP\_Assignment\_code.ipynb’ file**

1. ‘NLP\_Assignment\_code.ipynb’ is a jupyter notebook file
2. Open jupyter notebook, if installed already
3. To open Jupyter Notebook – Go to ‘cmd’ or ‘command prompt’ and Type ‘jupyter notebook’ and press enter
4. Open jupyter notebook is opened
5. Now, Upload all the files that I provided in google drive
6. Now, open run ‘NLP\_Assignment\_code.ipynb’ file as it is
7. Wait till all the cells gets executed properly
8. Now, after the successful execution of file, go to home page of jupyter notebook and look for ‘Output Data Structure.xlsx’ file, you can see all the mathematical values in their respective columns
9. To see the output, download ‘Output Data Structure.xlsx’ file by right clicking on it and then open in MS Excel Software