Instructions:

The submission folder contains 3 files namely main.py, scrape.py, alter.py

To automate the URL feeding to our scraping program only the URL column of the input.xlsx is stored in a CSV format named intern.csv next

We scrape the data using python file scrape.py to obtain the title of the webpage and the article’s texts and all of the contents scraped were stored in text files

Note: only those enclosed in paragraph tag were scraped but those within li tag and:: marker were scraped manually

Since the code does not take each of the text scraped in utf8 encoding there may be an error output at some point of the URL list(which is named intern.csv) For that an alternate python script is written called alter.py in this script you have to enter the URL manually by checking the already scraped file index position. After we enter the URL that posed the error we get the scraped data which is stored in another text document.

At the end of these processes, we get the complete scraped data. Now the text files that contain the scraped data of each of these webpages are fed to the main.py python script.

In main.py we clean the scraped data by removing stopwords and punctuations. Then a separate list of both positive and negative words from the master dictionary named nu.csv is downloaded from the link <https://sraf.nd.edu/loughranmcdonald-master-dictionary/> after that all the main computations are done for all the URLs provided

Computations: Done as per Text Analysis.docx provided by the company

1. POSITIVE SCORE
2. NEGATIVE SCORE
3. POLARITY SCORE
4. SUBJECTIVITY SCORE
5. AVG SENTENCE LENGTH
6. PERCENTAGE OF COMPLEX WORDS
7. FOG INDEX
8. AVG NUMBER OF WORDS PER SENTENCE
9. COMPLEX WORD COUNT
10. WORD COUNT
11. SYLLABLE PER WORD
12. PERSONAL PRONOUNS
13. AVG WORD LENGTH