By Shrey Srivastava

\*\*\*INSTRUCTIONS\*\*\*

To extract the article text from the articles in the input.xlsx file using Scrapy, you can follow these steps:

* Install Scrapy by running the following command:
  + pip install scrapy
* Create a new Scrapy project by running the following command:
  + scrapy startproject article\_extractor
* Create a new Scrapy spider by running the following command:
  + scrapy genspider article\_spider insights.blackcoffer.com
* Inside article\_spider.py, Define the parse method to read the ‘Input.xlsx’ file and obtain URL\_IDs and URLs and then handle the response from each URL. This method should extract the article title and text using the appropriate HTML tags and CSS selectors.
* To extract only the article title and text, you can exclude certain elements such as the website header and footer. You can do this by using the extract\_first() method or by using the xpath() method and specifying the desired elements.
* Save id to url conversion dictionary as json for future access.
* Once you have extracted the article title and text, you can save them to a file using the with open statement and the .write() method. The file name should be the URL\_ID from the input.xlsx file.
* Run the spider by creating a CrawlerProcess object to run the spider class.
* This should extract the article text from the articles in the input.xlsx file and save them to separate text files with the URL\_ID as the file name in the ‘extracted\_articles’ subfolder.

To perform data analysis we follow these steps:

* Import necessary modules and nltk packages.
* Store stopwords, positive words, negative words, cmu dictionary.
* Convert id\_to\_url json file to dictionary for usage.
* Define compute\_variables function to compute each variable store them together in a list.
* Create workbook object to insert column names.
* Open the text files from folder ‘extracted\_articles’ and compute the variables for each file.
* Obtain url\_id from file name and url from id\_to\_url dictionary.
* Insert the computed variables for each text file in 2d list.
* Sort the list based in URL\_ID column.
* Save the changed list in ‘Output Data Structure.xlsx’.