RESEARCH INTERNSHIP REPORT

HIGHLIGHTING THE DIFFERENCES BETWEEN THE WEB-SCRAPING SCRIPTS:

Differences	Web Script 1	Web Script 2
Website Information	URL and Name	URL, Company Name, Document Type, Document Format
Scraped Data Segregation	Not segregated	Segregated based on Document Type and Format
Document Type and Format Determination	Not determined	Determined based on patterns/keywords
CSV and Pickle File Naming	Uses current date and time for file name	Uses current date for file names
CSV and Pickle File Structure	Simple structure	Organized structure with headers and sections
Additional Data Saving	None	Saves data as a pickle file in addition to CSV

KEY INSIGHTS

- These differences highlight the additional functionality and organization introduced in Web Script 2 compared to Web Script 1. Web Script 2 provides detailed information about each website, segregates the scraped data based on document type and format, determines document type and format using specific patterns/keywords, and saves the data in CSV and pickle formats.
- Considering both scripts I did use the second script in my final report but not the full implementation because while going through all the websites that I scraped for, each site have its own HTML structure which was becoming very difficult for me to create a generic web scraping script and just simply add the URL of websites in the code.
- I made sure I adjusted the HTML structure of the given sites but some of them were tricky in the company that uses dynamic mapping which means they changed their HTML structure of the page dynamically through Java script.
- I referred to the old web scraping previously provided to me by Don, although almost 80% of those codes had a lot of issues while running and needed modifications. So according to the requirement I have modified and also made sure the scraped data for a particular company get saved with the date the trigger is provided to the script and description type/company name.
- Previously I had issues running the topic modeller that was created by Don and Rebecca, because of this reason I decided to create my topic model although it is not of that large scale enterprise level it does work when I ran a sample CSV scraped file through the model.
- I have made two types of topic modellers: One is the basic topic modeller where we just have pre-processing, vectorization and LDA. The other topic model I created is using the Gensim Model where I did train the model (LDA) with the corpus and dictionary.