Reflection on Tools for Data Analysis Semester Project

Name: Nana Adwoa Nkrumah

NUID: 14561847

Spring 2023

I. Introduction

The Human Rights Violators & War Crimes Center (HRVWCC) of Homeland Security Investigations was interested in data mining tools to gather information on the Wagner Private Military Company's operations in Africa. Their main objective was to create a network analysis, identifying funding sources, link between Wagner operatives and host governments, key operatives, and human rights violations committed by the group and host nations. The HRVWCC representatives mentioned that as a center they primarily conduct research to support various teams to identify, locate, investigate, prosecute, and remove violators from the United States.

To address the needs presented, I employed tools introduced during the semester to collect and clean data, as well as conduct some fundamental analysis. The final product is a custom web scraping tool. This tool was designed to extract news articles from the Bing news aggregator related to the Wagner Group's operations in Africa and then compile the relevant information into a text document for further analysis. This provides an efficient approach to access the most recent news articles about the group and facilitates the identification of key individuals as well as the establishment of links between host governments or countries.

Through this project, I learned that data mining can be very useful specifically, that web scrapping can be an effective method for gathering relevant web articles and extracting valuable information. Furthermore, I realized that text analysis requires a lot more effort and creative thinking than quantitative analysis. Looking back, I realize that I could have dedicated more time to fine tune the text analysis rather than focusing on the development of the web scrapping tool because the Natural Language Processing (NLP) library spaCy has several functionalities that could have been used to do so much more analyses-wise.

II. Lessons Learned

I would like start by acknowledging that things do not always go as planned and that it is important to be comfortable with changing direction when necessary. At the initial stage of the project, I put together a proposal to investigate the Wagner Group's operations over the years, with a focus on understanding its tactics or strategies, and objectives in each location. This objective was however hindered by the Bing News API's limitation of extracting news article within a month timeframe and nothing beyond that. As I was not aware of this, I spent a lot of time devising different ways to get information from the year 2018 to 2023, so I could have 5 years of data to collect and analyze for insights into the Wagner's Group overall strategy, tactics, and objectives while providing a timeline of network commonalities and differences in operations. The findings of the project provide valuable insights into the operations of the Wagner Group in Africa but not as extensive as I hoped for.

As I was determined to build a web scrapping tool to be used by HRVWCC to aid in their research, I explored a variety of methods to handle errors and exceptions that may arise when

working with Web APIs. For example, to address the strict rate limits, I included the sleep function to pause the function for a while before sending additional requests. Through this process, I gained a lot of experience in building reproducible parts of code and found a lot more on automated error-handling I could attempt if given a similar project in the future.

Another valuable takeaway was the significance of having a wide selection of tools available and understanding how and when to use them. It was interesting how regular expressions helped in the data cleaning process and to ensure that user inputs are given in the correct format to run the code. In contrast to that success, attempts at implementing multiprocessing failed horribly. Nonetheless, I am motivated to improve this skill over time.

III. Conclusion

In summary, this project presented both challenges and opportunities for growth. I have gained a lot of experience applying the knowledge and skills acquired in the class to address a real-word proud. While I am excited about the tool I developed, I admit that there is room for improvement in terms of its robustness.

Upon reflecting, I realized I could have extended the `scrape_website` function to work on wiki tables to provide more functionality to the users (HRVWCC). Further text analysis could have provided more insights into the Wagner Group's activities at a more in-depth level than I currently have. If I were to re do this project again, I would allocate more time devising tools to clean the names and countries data collected to address duplicates. In addition to that, I would provide additional information such as affiliation of the people, their country of origin, any recent conflicts they may have been involved with. For countries, I would aim to collect information on current conflicts, number of Wagner operatives involved, number of casualties, objective of the group in that country and other details that the Center could use to aid their mission.

Overall, I am grateful for the experience and practicality of the project and class. I look forward to applying what I have learned about web scrapping, data processing, adaptability and tool proficiency to future projects. I know for sure that I would be using Web APIs in my job, so I am very appreciative for the exposure.

References

[1] Analytics Vidhya. (2021, May 5). *How to Build Word Cloud in Python?* Retrieved May 5, 2023, from https://www.analyticsvidhya.com/blog/2021/05/how-to-build-word-cloud-in-python/

[2] Analytics Vidhya. (2021, November 1). *Web Scraping a News Article and Performing Sentiment Analysis using NLP.* Retrieved May 5, 2023, from https://www.analyticsvidhya.com/blog/2021/11/web-scraping-a-news-article-and-performing-sentiment-analysis-using-nlp/

[3] spaCy 101: *Everything you need to know.* Retrieved April 20, 2023, from https://spacy.io/usage/spacy-101

[4] Medium. (2020, July 3). *Generate Meaningful Word Clouds in Python*. Retrieved May 5 2023, from https://towardsdatascience.com/generate-meaningful-word-clouds-in-python-5b85f5668eeb

[5] Microsoft. *Bing News Search API overview*. Retrieved April 3, 2023, from https://learn.microsoft.com/en-us/bing/search-apis/bing-news-search/overview

[6] OpenAI Chat. Retrieved March 18, 2023, from https://chat.openai.com/

[7] Python Tutorial. *Python write to Text File.* Retrieved April 25, 2023, from https://www.pythontutorial.net/python-basics/python-write-text-file/

[8] Real Python. *Python Counter.* Retrieved May 2, 2023, from https://realpython.com/python-counter/

[9] Stack Overflow. (2015, August 12). *How to embed image or picture in Jupyter Notebook either from a local machine or from a web resource?* Retrieved May 7, 2023, from https://stackoverflow.com/questions/32370281/how-to-embed-image-or-picture-in-jupyter-notebook-either-from-a-local-machine-o

[10] Stack Overflow. (2020, May 27). *Plotly list of valid country names from ISO 3 code*. Retrieved May 7, 2023, from https://stackoverflow.com/questions/62566605/plotly-list-of-valid-country-names-from-iso-3-code

[11] Wikipedia. (2023, April 14). *Wagner Group*. Retrieved April 2, 2023, from https://en.wikipedia.org/wiki/Wagner_Group