PROJECT MILESTONE-2
Group 2: Embedding Eagles

Introduction: We were asked to take a dataset and explore it by checking the contents of the dataset and applying certain exploratory techniques like Clustering, UMAP/t-SNE, Clustering etc. The goal is to understand the dataset and apply any type of operations to see how operations work.

Input: We were given five datasets and asked to select any dataset of our choice. If we want to use any other dataset of interest, then we must submit a request for permission to use the third party dataset.

Output: Required output should include documentation (writeup) about the steps we used to explore the dataset, results of evaluation, including questions and ideas for next steps in the project. We need to attach a Jypter notebook file where we present the code demonstrating all the operations we have done on the dataset. We can also include any tables, figures or screenshots if necessary. All the files should be made into a zip and submitted collectively as a group.

Preprocessing the data:

1. **Appropriate Unit of Analysis:** As far as our knowledge, the most appropriate and best approach to do analysis is dividing a large set of data into smaller parts. It would look very easy and convenient to look at and operate. We can also split them into small parts using logical differences or based on any other similarity like grouping all the similar things at one place.
   Not only us, even models and systems will be able to provide detailed analysis with greater accuracy if the input dataset is organized perfectly based on the differences or any other factors like similarity. If you give your dataset after doing basic pre-processing things like removing any redundant data, adding any missing values.

   If we look more deeply, the unit of analysis, especially sentence analysis, is more efficient for JSON Lines (.jsonl) data is critical to the effectiveness and efficiency of the project, particularly when working with natural language processing (NLP), machine learning, or data exploration tasks. It also include lot of advantages such as:
   - Analyzing data at the sentence level can simplify the processing of complex documents by breaking them down into smaller, more manageable units.
   - Sentence-level processing enables more effective parallelization of computational tasks as sentences can be processed independently.
   - Certain analysis tasks may benefit from increased accuracy when performed at the sentence level, as the context is more localized.
   - Natural Language Processing (NLP's) will benefit more from using sentence analysis as they demonstrate ideas effectively like Sentiment analysis.
2. **Challenges with Data:** There will be different challenges faced while working with the datasets. In some cases, the source of data is also a confusing one. We must verify the

credibility of the author to trust whether to use data or not. If you are not focusing more then there is a scope of virus attacks etc. More challenges are:

- Sometimes, when data is very large, like in bytes, we require large computing resources to process the data.
- In some instances, data will have several inconsistencies, like missing values, etc. That needs to be taken care of before processing.
- It would be very problematic to synchronize various types of data.
- Using data means you are accepting privacy and security issues. When dealing with sensitive and confidential data, we must be very careful.

Steps involved in the process:

- Firstly, we have selected one dataset named "**ecfr-title-12.jsonl**" from the given five datasets. It looked a bit simpler than other ones, and we are comfortable handling JSON files.
- Before going with exploration, we loaded the dataset into Google Colab. We have just uploaded it to session storage conveniently. Then we loaded it using the "loads" method of the JSON module.
- After loading the data, we explored data using the for loop to get a sample of how data is divided and displayed in the JSONL dataset.
- After watching that, we were able to ensure that data was organized adequately and saw how data was stored in a proper format.
- Then we wrote a Python script for getting details of meta data to get the number of records present in the dataset, what keys (fields) are present, and data types of the data.
- Then we wrote a Python script to get other metadata like author, name, and title of dataset. But we did not get any response as the author name and dataset name were not explicitly present. It may be a combination of keys; otherwise, it might not be there.
- To get those missing details, we checked the data manually and succeeded in getting the dataset name and the person who modifies the dataset. We found some links navigating to the website too.
- We then wrote another Python script to get the location of unstructured data and see the format.
- Now we evaluated the quality of the dataset, especially by using two metrics 'completeness' and 'consistency'. Dataset was not of quality as it had some inconsistencies in the meta field for all the entries, and there was one missing value in the text field.
- Finally, we tried to implement one of the exploratory techniques, Sentence Embedding Visualization. There was no problem with the script, but we did not get the output. So, we did not explain it.

Evaluation:

We have evaluated the dataset using two metrics, completeness and consistency, which measure the quality of the dataset. The dataset had inconsistencies with all the entries in the meta column, and there was one value missing in the text column. Hence, the dataset was not a quality one. We tried to further explore using the advanced technique "Sentence Embedding

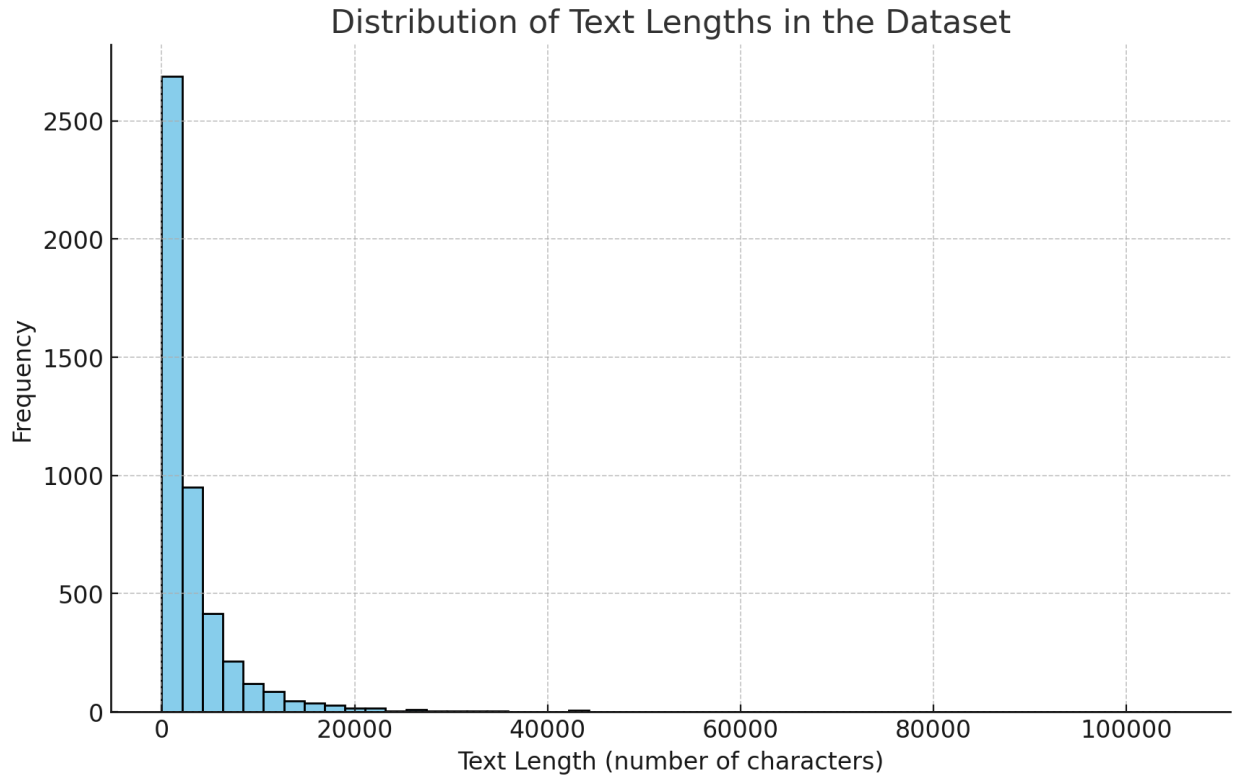
Visualization”. Unfortunately, something went wrong, as we did not get the accurate expected output even after executing the right script.

Questions and ideas:

We got the following ideas and questions while doing this project:

- How can we find the meta data (source, author, title of the article) using the dataset?
- Why did we get error in sentence embedding visualization script, and how do we correct it?
- What file format are the letters in, and how does this format affect data loading and processing?
- What are the potential limitations of analysis? How can we improve them?
- For what applications is this type of analysis useful?
- We could pick a problem in the domain of the dataset and give solution with an idea next time.
- What insights were gained from using techniques like Sentence Embedding Visulization for topic modeling? Include any interesting findings or visualizations.

Appendix:



Source of histogram: OpenAI. (2024). *ChatGPT* (4) [large language model]. https://chat.openai.com