Climate Numbers Extraction Report

1. Introduction

The objective of this project was to utilize Natural Language Processing (NLP) techniques to extract and compile climate-related numerical data from a collection of research papers focused on food and sustainability. The significance lies in understanding and quantifying the environmental impact, emissions, and sustainability metrics associated with food production.

2. Methodology

The methodology involved the following steps:

Document Loading and Text Processing: The code addressed the removal of newline characters ('\n') that appeared without a preceding period ('.'). This action aimed to handle instances where newline characters were placed at the end of lines within sentences, potentially disrupting the natural flow of text. The decision to preserve '\n' characters following periods ('.') was intentional, marking the end of paragraphs within the text.

Text Chunking: Utilizing RecursiveCharacterTextSplitter, the documents were segmented into smaller chunks for efficient processing. The Recursive Character Text Splitter operates by recursively breaking down larger blocks of text into smaller, manageable segments based on predefined separators. It splits the text into chunks while ensuring that the resulting segments remain coherent and retain the structural integrity of the original content.

When employing "\n" and "." as separators:

- '\n' (Newline Character):

By utilizing '\n' as a separator in the text splitter, it ensures that chunks contain complete paragraphs, allowing for coherent analysis within each chunk.

- '.' (Period/Full Stop):

In the context of the text splitter, the '.' serves as a boundary marker to prevent sentence fragmentation while segmenting the text into chunks.

By combining '\n' and '.' as separators, the aim is to create chunks that encapsulate entire paragraphs (indicated by '\n') without disrupting sentence boundaries (indicated by '.'). This approach maintains both paragraph and sentence structures within the segmented text chunks, aiding in subsequent analysis and comprehension.

OpenAI GPT-3.5 Model Usage: The OpenAI API was employed to extract climate-related numerical data from the processed text chunks. The GPT-3.5 model was prompted with guidelines to focus exclusively on climate metrics. By setting a temperature of 0, the model was directed to generate more deterministic and focused responses, minimizing the introduction of randomness or variations in its outputs.

The selection of the GPT-3.5 Turbo model was based on its capabilities in natural language processing, particularly in generating coherent and contextually relevant text responses. GPT-3.5 Turbo represents an upgraded version of the GPT-3 model, incorporating improvements in speed, performance, and accuracy in understanding and generating human-like text. (OpenAI, 2023)

Key components of the prompt include:

- Focus on Climate Metrics: Explicit instructions are given to concentrate solely on climate impact, emissions, and environmental metrics, excluding any nutrition, price, or unrelated data.
- Formatting Guidelines: Clear guidelines are provided on the expected format of the output, emphasizing the structure of "<Metric>: <Value> <Unit>" or "<Metric>: <Value> <Unit> per <Unit> if possible" to ensure consistency and clarity in the extracted data.
- Data Exclusion Criteria: Instructions are provided to exclude metrics where numerical data is unavailable, ensuring the response remains focused on available climate-related numerical values.

3. Challenges Faced

Addressing API Rate Limiting: During the data extraction process, challenges included API rate-limiting issues due to frequent requests. To mitigate this, a delay of 2 seconds was introduced between API calls to avoid exceeding the token rate per minute.

Ensuring GPT-3.5 Adherence to Instructions: A challenge arose in ensuring the GPT-3.5 model's consistent adherence to provided prompts. Despite instructing the model to exclusively extract climate-related numerical data and format responses as directed, occasional deviations occurred in the generated outputs. To address this, a conditional statement was integrated into the code, acting as a filter to exclude responses that didn't comply with the specified format or contained irrelevant and unwanted data. While utilizing GPT-4 might have presented a potential solution due to its advancements in language processing, its implementation was expensive, which influenced the decision to work within the limitations of GPT-3.5.

Additionally, it was observed that the GPT-3.5 model occasionally produced itemized responses, such as bulleted lists, which deviated from the prescribed single-line format. Despite efforts to guide the model through prompts, the occasional generation of itemized outputs persisted, requiring further optimization strategies to improve the consistency of the model's responses in adhering to the specified output format.

Encountering multiple lines of different numerical information for the same climate metric posed challenges in consolidating and organizing the data. A systematic solution can be maintaining a record of past GPT-3.5 model responses. This response history can be utilized in each batch, guiding the model to integrate new data into the existing records. Through iterative accumulation across batches, a comprehensive dataset is built. Validation steps should also be implemented to compare and confirm numerical values associated with identical metrics, enhancing the accuracy of the extracted information.

4. Insights Gained

The outcome showcases an extensive array of extracted climate-related numerical data, encompassing diverse metrics ranging from carbon footprints to emissions factors across various agricultural and production practices. Insights gained from this data include carbon footprint assessments for different food products, greenhouse gas emission factors for specific agricultural processes, nitrogen content and excretion rates in livestock production, and crop yield variations between conventional and organic practices. These insights collectively offer a comprehensive understanding of the environmental implications associated with food production, aiding in the evaluation and potential mitigation of its ecological footprint.

5. Conclusion

In conclusion, the application of NLP techniques combined with the OpenAI GPT-3.5 model facilitated the extraction of crucial climate-related numerical data from research papers on food and sustainability. The extracted information, encompassing emissions, environmental impact, and sustainability metrics, sheds light on the relationship between food production and its environmental consequences.

6. GitHub Repository

The code, extracted data, and documentation for this project are available in the GitHub repository: Link to Repository

7. References

OpenAI. (2023, November 6). *New models and developer products announced at DevDay*. Openai.com. https://openai.com/blog/new-models-and-developer-products-announced-at-devday