**Capstone Project**

**Text Summarization Application using NLP.**

**“Snap Summ”**

A black background with a black square

Description automatically generated with medium confidence

Shyambhargavreddy Allam and Gowtham Reddy Gurrala

Department of Data Science, University of Maryland Baltimore County

DATA 606: Capstone in Data Science

Dr. Antonio Diana

May 2023

**Abstract**

In the era of information abundance, researchers grapple with the overwhelming volume of data, often needing more time constraints. This paper addresses the challenge of using a Natural Language Processing (NLP) algorithm to condense extensive texts into meaningful summaries. Our approach involves training the T5 model around ML, AI, and NLP topics. Prioritizing accuracy and speed, we aim to deliver concise and contextually rich summaries promptly. This research contributes to the field of text summarization, offering valuable tools to alleviate the burden of information overload for researchers dealing with vast datasets. By exploring the T5 NLP algorithm, we seek to empower researchers with adequate means of extracting critical insights from extensive textual content.

**Introduction**

In an ever-changing information ecosystem, researchers face the daunting task of extracting meaningful content within severe time limitations. This study addresses a crucial challenge: How effectively can the T5 model, selected for its efficiency and accuracy, automate text summarization to expedite information retrieval? Summarization, as explored in this research, encapsulates three crucial attributes: the inclusion of vital information, avoidance of redundancy, and the creation of coherent, grammatically sound summaries. Our approach utilizes the simple T5 library to train the T5 model efficiently, streamlining the process of generating coherent and concise summaries with minimal coding effort. (Rush et al., 2015).

Delving into the nuanced task of extracting critical phrases from scientific articles, we recognize this endeavor's essential and rewarding nature. Utilizing the T5 model, trained on datasets annotated with keywords based on concept taxonomies and author judgments, our study tackles the unique challenges presented by scientific papers. Additionally, our research demonstrates the T5 model's adeptness at summarizing extensive content from multi-page PDF documents, showcasing its versatility in applying advanced abstractive summarization techniques (See et al., 2017).

Beyond academia, this research holds pragmatic significance for professionals navigating a sea of information. Optimizing automated summarization tools aspires to empower users with efficient means of distilling pertinent knowledge from voluminous textual sources. The detailed examination of the T5 model provides practical insights into its strengths and limitations in text summarization (Rush et al., 2015). This research advances the T5 summarization technique and enriches the broader discourse on information retrieval, processing, and comprehension, demonstrating the model's effectiveness in practical applications.

**Related Work**

Automated text summarization, a burgeoning research domain, is gaining prominence with a specialized focus on scientific documents. Erera et al. (2019) introduced a dedicated summarization system tailored for computer science publications. The system, informed by user scenarios, efficiently retrieves and summarizes scientific documents, underscoring the significance of tailored approaches for compelling document exploration. This user-centric methodology aligns with the evolving landscape of summarization needs, acknowledging the pivotal role of context in information extraction.

In parallel efforts, Cano and Bojar (2019) contributed to the discourse by proposing a neural extractive summarization model explicitly designed for handling the intricacies of long scientific documents. The model's incorporation of both global and local contexts addresses the nuances inherent in scientific literature. This strategic integration of contextual information underscores the importance of considering document structure and content relationships, particularly in lengthy scientific discourse.

Addressing the intricacies of abstractive summarization, Xiao and Carenini (2019) explored key phrase generation to distill essential information from scientific articles. Despite leveraging advanced deep learning models, their findings underscore the challenges in surpassing more straightforward unsupervised methods. This highlights the inherent complexities of abstractive summarization beyond explicit text content. The study emphasizes the necessity of nuanced approaches to capture and represent the latent information encapsulated within the scientific literature.

Building upon this, Sefid and Giles (2022) introduced SciBERTSUM, a summarization framework explicitly designed to handle the unique challenges of long scientific documents. Diverging from traditional summarization methods, SciBERTSUM extends BERTSUM with a section embedding layer and a sparse attention mechanism, showcasing superior performance in summarizing lengthy scientific papers. This approach acknowledges the distinctive structural and contextual characteristics inherent in scientific discourse.

Concluding the literature review, Widyassari et al. (2022) conducted a comprehensive systematic review of automatic text summarization techniques and methods. Spanning over a decade, the review provided insights into the evolution from extractive to abstractive summarization. It identified key trends, features, problems, and methods, showcasing the field's progression. This retrospective analysis serves as a foundation for understanding the current state of research, highlighting the need for continual advancements in addressing the unique challenges posed by the summarization of scientific literature.

**Methodology**

**Text Extraction:**

In the text extraction phase, the research employs a dedicated Python function, extract\_text\_from\_pdf, to extract pertinent information from PDF documents. The extraction library facilitates the PyMuPDF process, specifically the Fitz module. The function takes an uploaded PDF file as input and opens it using Fitz. open(stream=uploaded\_file.read(), filetype="pdf"). A systematic iteration through each page of the PDF document is performed, and for each page, the text content is extracted using page.get\_text(). The extracted text from each page is then concatenated, resulting in a comprehensive textual representation of the entire PDF document. This extracted text forms the foundation for subsequent processing and analysis within the summarization application, enabling efficient content handling for further summarization techniques.

**Abstractive Summarization:**

Abstractive summarization represents a transformative approach, surpassing the mere extraction of sentences by generating concise and coherent summaries through paraphrasing and rephrasing. Our methodology exclusively employs the T5 model, explicitly leveraging the simple T5 library to implement the T5ForConditionalGeneration model. This approach excels in generating human-like summaries that adeptly capture the essence of the original text. By converting all language problems into a text-to-text format, the T5 model adds a layer of sophistication to our summarization process.

Using the SimpleT5 setup, we dynamically manage the granularity of our summaries based on predefined levels—Detailed, Moderate, or Brief—modulating the chunk percentage of the text to alter the depth of the summary generated. This method ensures tailored summarization that aligns with specific user needs. This focused approach harnesses the power of advanced NLP techniques provided by Hugging Face's T5 model, streamlined through the simple T5 library, ensuring an effective and efficient solution for generating concise and coherent summaries.

***T5 Summarization Methodology:***

Expanding our abstractive summarization capabilities, we incorporate the T5ForConditionalGeneration model, a transformer-based model developed by Hugging Face. Initially designed for abstractive summarization, T5 excels in generating summaries by paraphrasing input text. In our approach, T5 plays a pivotal role in selecting sentences based on their relevance and importance to the document, contributing a sophisticated layer to our summarization strategy.

***Training the T5 model:***

We streamlined the training of the T5 model using the SimpleT5 library, starting with data preparation, where we installed necessary packages, loaded, and reformatted our dataset project\_data1.csv. Inputs were prefixed with "summarize: " to specify the summarization task. Shivanandroy. (2021). The model, configured with the t5-base variant, was trained on an 80/20 train-test split, with source and target token lengths set to 128 and 50, respectively. The training was executed over three epochs with a batch size of 8, using GPU acceleration to improve performance. Finally, the model was saved to Google Drive for deployment and further use, demonstrating its efficacy with an example summarization.

**Additional Features**

**Text-to-Speech Conversion:**

Text-to-speech (TTS) conversion is a pivotal aspect of our research, bridging the gap between written content and auditory experiences. In this module, we explore integrating TTS technology to dynamically convert textual information into spoken words, offering an alternative content consumption mode. Central to our TTS implementation is the gTTS (Google Text-to-Speech) library, a powerful tool that leverages Google Translate's TTS functionality. By utilizing HTTPS, we tap into a robust system for generating natural-sounding speech from textual input, expanding the accessibility and usability of our summarization outputs.

**Language Translation:**

Language translation is a cornerstone in our research, breaking down linguistic barriers and enabling a broader audience to access diverse content. Our study introduces a comprehensive language translation functionality, allowing users to effortlessly translate summarized content from the source language to many target languages. This feature enhances accessibility, making information available to individuals regardless of their native language. At the heart of our language translation capabilities lies the Googletrans library, a robust tool harnessing Google Translate's powerful translation engine. By incorporating Googletrans, we empower our system to seamlessly translate content between various languages, expanding the reach of our summarization outputs to a global audience.

**Chat Support Feature Implementation:**

Our project features a chat support page that utilizes OpenAI's GPT model to offer real-time assistance to users. Integrated through Streamlit, the chatbot initiates conversations with a welcoming message and responds dynamically to user inputs by maintaining a session state that tracks the conversation history. This setup allows the chatbot to provide contextually relevant and intelligent responses, leveraging the powerful natural language processing capabilities of the GPT model. Users interact with the chatbot through a simple input field, making the feature intuitive and effective for enhancing user engagement.

**Outcomes**

**T5 Summarization:**

****

Fig 1. Abstractive T5 Summarization

In this representation, Fig 1 demonstrates the Abstractive Summarization using the T5 (Text-To-Text Transfer Transformer) method from the Hugging Face Transformers library. T5 employs a text-to-text framework, treating summarization as a translation task that converts the input text into a target summary. The model is trained on diverse tasks, allowing it to generate summaries by understanding the context and relationships within the input content.

**Metrics:**

We employ semantic similarity metrics to evaluate the effectiveness of the text summarization performed by the Simple T5 model. These metrics quantitatively assess how well the meanings of the original texts are preserved in the summaries generated by the model. Specifically, we compute the cosine similarity between the embeddings of the original and summarized texts. This approach provides a clear, numerical value that indicates the degree to which the semantic content of the original text is mirrored in the summary. Using these metrics, we can objectively measure the quality of the model's output, ensuring that the summaries are concise and maintain the integrity of the original text's meaning.

**Conclusion**

In conclusion, implementing the Simple T5 model in our text summarization application has demonstrated high efficacy, consistently generating structured and meaningful summaries. The semantic similarity scores, ranging from 0.85 to 0.95, indicate a robust preservation of the original text's meaning within the summaries. As illustrated in Fig 1, a representative sample achieved a similarity score 0.93, underscoring the model’s capability to maintain substantial semantic integrity. These results affirm that the Simple T5 model accurately summarizes text and effectively retains the essential information and context. Overall, the application provides a valuable tool for researchers and professionals, enhancing efficiency by delivering concise yet comprehensive summaries of extensive texts.

**References**

Çano, E., & Bojar, O. (2019). Keyphrase Generation: A Text Summarization Struggle. *Arxiv*, 1904.00110. <https://doi.org/10.18653/v1/n19-1070>

Erera, S., Shmueli-Scheuer, M., Feigenblat, G., Nakash, O. P., Boni, O., Roitman, H., Cohen, D., Weiner, B., Mass, Y., Rivlin, O., Lev, G., Jerbi, A., Herzig, J., Hou, Y., Jochim, C., Gleize, M., Bonin, F., & Konopnicki, D. (2019). A Summarization System for Scientific Documents. *Arxiv*, 1908.11152. <https://doi.org/10.18653/v1/d19-3036>

Shivanandroy. (2021). Quickly train T5/mT5/byT5/CodeT5 models in just 3 lines of code: <https://medium.com/geekculture/simplet5-train-t5-models-in-just-3-lines-of-code-by-shivanand-roy-2021-354df5ae46ba>

Rush, A. M., Chopra, S., & Weston, J. (2015). A Neural Attention Model for Abstractive Sentence Summarization. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 379–389.

See, A., Liu, P. J., & Manning, C. D. (2017). Get To the Point: Summarization with Pointer-Generator Networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1073–1083.

Xiao, W., & Carenini, G. (2019). Extractive Summarization of Long Documents by Combining Global and Local Context. Arxiv, 1909.08089. <https://doi.org/10.18653/v1/d19-1298>

Sefid, A., & Giles, C. L. (2022). SCIBERTSUM: Extractive Summarization for Scientific Documents. In Lecture Notes in Computer Science (pp. 688–701). <https://doi.org/10.1007/978-3-031-06555-2_46>

Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., Affandy, A., & Setiadi, D. R. I. M. (2022). Review of automatic text summarization techniques and methods. Journal of King Saud University - Computer and Information Sciences, 34(4), 1029–1046. <https://doi.org/10.1016/j.jksuci.2020.05.006>