Republic of the Philippines
PALAWAN STATE UNIVERSITY
Puerto Princesa City
College of Sciences

SUMMARIZING PORTABLE DOCUMENT FORMAT FILES USING LANGCHAIN

Adviser

- Research Title should not start with abbreviation.

Title is changed from “PDF SUMMARIZATION TROUGH ANALYSIS AND IMPLEMENTAION WITH LANGCHAIN” to the current title “SUMMARIZING PORTABLE DOCUMENT FORMAT FILES USING LANGCHAIN”.

An Undergraduate Thesis Presented to the Faculty of the

College of Science

Palawan State University

Puerto Princesa City

CABITAC, JOHN MARK L.

DELGADO, JOHN ISAAC A.

PADON, GILLBERT M.

VALLESCAS, JAY ALVIN C.

Table of contents

Title……1

Table of content……2

List of Figures and Tables……4

Chapter I

Introduction……5

Background of the Study……5

Statement of the Problem……7

Objectives……8

Significance……9

Scope……10

Limitations……11

Operational Definition of Terms……12

Conceptual Framework……15

Chapter II

Review of Related Literature ……16

Local Literature……16

Foreign Literature ……28

Chapter III

Theoretical Background.....39

A. Preprocessing and Creation of Embeddings…….....40

Word Embedding……..……..….40

B. Loading the Model…..……..……42

Neural Networks…..……..…….42

Transformer Model……...……..……..45

Activations used in the Transformer Model……..……45

Softmax function…..……..……45

Scaled Dot-Product Attention…….……46

Cross-entropy loss function…….……..……47

Gradient Descent……...……..……48

Backpropagation……..……..……50

C. Loading the Agents and Tools……..……51

D. Create Chains, Perform Chains, and Getting the Output……..……51

E. LangChain…….51

Bibliography……...……..……..53

List of Figures, List of Tables

Figure 1: Conceptual Framework……..……..……15

Figure 2: Prototype (No UI)……....……..……15

Figure 3: Technical Architecture……..……....39

Figure 4: Example embeddings about royalty……..…41

Figure 5: A representation of a neural network……..……..43

Figure 6: Common Activation Functions……..…….44

Figure 7: Incremental steps on a gradient descent……..……...49

Adviser

- Introduction and followed by the background of the study.

Dr. Floredith Jeanne Alcid

- Add “Chapter I” before Introduction.

CHAPTER I

Introduction

Background of the Study

In 2021, it is estimated that there are 15 trillion files on the internet, with PDFs ranking third with 2.5 trillion files, comprising 16.67% of all files (Cloud Files, 2023). To provide context, as of 2021, the world population is 7.9 billion (Word Bank, 2021), which means that if we evenly distribute all PDF files on the internet, each person would have approximately 316 PDF files. This gives us a better understanding of how vast the amount of digital information available on the internet. The use of PDF documents is widespread across various industries, but it can be time-consuming and challenging to extract essential information from lengthy PDFs. This challenge has led to a need for effective and precise methods of summarizing PDF documents.

Summarization is a process of automatically condensing and rewriting a large chunk of text to create a small, crisp summary. A summarization system should give the reader most of the information present in the original document while also ensuring that no important information has been lost during condensation (Vendantu, 2023). The aim is to provide a concise and informative summary that will help readers quickly grasp the main ideas of a document without having to read through the entire text.

There are two types of summarizations; abstractive summarization generates a summary that is not limited to the original text's wording, while extractive summarization selects and rearranges sentences from the original text. Extractive summarization tends to be more straightforward, but abstractive summarization has the potential to produce more human-like and coherent summaries.

In this study, the researchers propose developing a PDF summarizing system utilizing LangChain, which makes use of language modeling to ensure the accuracy, and transparency of the material that is summarized. The LangChain-based PDF summarization system will utilize large language models to analyze and extract relevant information from PDF documents and generate a concise summary that captures the essence of the original content. The developed system will be evaluated on a diverse dataset of PDF documents, including research articles, legal documents, and financial reports. The evaluation will focus on the accuracy and efficiency of the system.

The proposed PDF summarization system using LangChain has the potential to greatly enhance the accessibility and usability of PDF documents, particularly in fields where time is of the essence. By providing an informative summary of lengthy documents, the LangChain-based system can help users quickly identify the key points of a document. Through rigorous evaluation and further development, this system has the potential to revolutionize the way in which PDF documents are analyzed and utilized in various fields.

Adviser

- AI tools can be used to summarize PDF files, and adding new elements can make the research more inventive.

- Concrete proof or citation of stated facts.

- Discussion of why we used the LangChain platform.

-Be consistent on how to write the word “LangChain”.

- Discuss how to measure the accuracy of generated pdf summarize

Dr. Floredith Jeanne Alcid

- Be more specific and add a scenario.

Statement of the Problem

General Statement of the Problem

The general problem is that lengthy PDF files are difficult and time-consuming to read and comprehend, especially when there are time constraints. This poses a challenge for individuals who need to quickly obtain information from these documents, as they must invest significant effort and spend more time than desired in order to fully understand the contents of the file.

Specific Problems

1. Current PDF/text summarization tools that rely on extractive summarization algorithms that results to broad summaries.
2. The need to evaluate LangChain technology against existing extractive summarization techniques to determine its effectiveness in producing high-quality and non-redundant summaries.
3. LangChain's generated summaries lack of proper formatting, making them difficult to understand and use effectively.

Adviser

Dr. Floredith Jeanne Alcid

-Revise the whole statement of the problem statement.

Prof. Rene Buliag

- Add General statement of the problem and at least 3 specific statement of the problem and its key information.

Objectives

General Objective

The general objective of this study is to develop a PDF summarization tool that can efficiently and accurately extract essential information and insights from lengthy Portable Document Format (PDF) documents. The tool should be capable of processing large volumes of information in a timely manner while maintaining the accuracy and relevance of the extracted information. The end goal is to provide users with a tool that can significantly reduce the time and effort required to comprehend and utilize the information contained within lengthy PDF documents.

Specific Objectives

1. To employ LangChain to the development of Portable Document Format (PDF) summarization tools that utilize advanced, non-extractive algorithms to produce more precise and informative summaries of documents.
2. To compare the performance of LangChain technology with existing extractive summarization techniques in terms of summarization quality, coherence, and redundancy.
3. To improve LangChain's summarization tool by enhancing its formatting capabilities, resulting in more coherent and well-structured summaries that are easier to understand and utilize effectively.

Adviser Dr. Floredith Jeanne Alcid

Prof. Rene Buliag

- Revise the whole Objective to match the statement of the problem.

- Add General Objective and at least 3 specific Objective and its key information.

Significance

The aims of the study entitled “Summarizing Portable Document Format Filles Using LangChain” is to extract information from document and produced summary. This study would be beneficial for the following fields:

Students

LangChain's PDF summarization tool offers students a convenient and efficient way to extract important information from lengthy academic documents, allowing them to save valuable time and effort in their studies.

Academia

With LangChain's advanced summarization technology, academic researchers and professionals can now process large volumes of information with greater ease and effectiveness. By providing them with concise and accurate summaries of complex research papers and reports, it can potentially enhance their productivity and streamline their work processes.

Society

The use of LangChain's summarization tool can significantly benefit society by enabling faster and more efficient processing of large amounts of information. This can prove especially useful in fields such as journalism, where the ability to quickly identify and summarize critical information from news articles and reports can help keep the public informed on important issues.

Scope

The scope of this study revolves around exploring the role of PDF summarization as a specialized tool specifically designed for summarizing PDF files. With a primary focus on PDF documents, this approach acknowledges the unique features and layout of such files, allowing for the extraction of essential information and key points.

The LangChain-based system in this study focuses exclusively on the English language, enabling it to generate summaries solely for English-language PDF documents. This targeted approach optimizes the summarization algorithm for English's linguistic nuances, grammar, and syntax, leading to accurate and effective summaries.

Moreover, this study focuses on the field of Computer Science in Palawan State University Main Campus, Puerto Princesa City, Palawan, Philippines, where the utilization of PDF summarization techniques holds great potential across various subfields and applications. One particular area of interest is academic research, where scholars can employ PDF summarization to effectively summarize theses and efficiently review and synthesize pertinent literature for their studies.

Prof. Jent Carlos Gardoce

- Study should focus on the field of Computer Science in Palawan State University Main Campus, Puerto Princesa City, Palawan, Philippines.

Limitation

PDF summarization, although a useful tool, comes with certain limitations. Firstly, the accuracy and efficacy of the summarization process heavily depend on the quality and complexity of the document being summarized. If the original PDF is poorly structured or contains intricate information, the summarization algorithm may struggle to extract the most relevant and coherent summary. Another limitation arises when dealing with languages other than English. Summarization algorithms are often trained on English text, and their effectiveness may decrease when applied to documents written in other languages due to linguistic nuances and differences in grammar and syntax. Additionally, when it comes to PDF file size, PDF summarization techniques may encounter certain limitations. One of the primary challenges is posed by large PDF files, which contain extensive amounts of text and graphics. These files tend to be more complex and may contain numerous pages, resulting in increased processing requirements for the summarization algorithms. As a consequence, summarizing large PDF files can be computationally intensive and time-consuming.

Furthermore, formatting issues can affect the summarization process. Complex formatting, such as intricate layouts, intricate tables, or information presented within images or graphs, may be difficult for the summarization algorithm to interpret accurately. As a result, the summarization output may not fully capture the intended meaning or important details embedded within these visual elements.

Dr. Floredith Jeanne Alcid

Prof. Jent Carlos Gardoce

- English language only, file size should 10Gigabyte below.

Technical Definition of Terms

- **PDF** – A type of file which is compatible with all devices display.
- **Abstractive summarization** - Abstractive summarization is a text summarization technique that uses NLP and machine learning to generate new sentences for a shorter summary that captures the main ideas of the original text.
- **Extractive summarization** - Extractive summarization is a text summarization technique that selects and rearranges existing sentences from the original text to create a summary.
- **LangChain** - A platform that uses natural language processing and machine learning to analyze and summarize PDF documents.
- **Machine Learning** - A type of artificial intelligence that uses data to learn and improve performance on a specific task.
- **Natural Language Processing** - The study of algorithms for analyzing and generating human language, with the aim of creating human-like computer processing.
- **Language Modeling** - A type of artificial intelligence that predicts the probability of a sequence of words in a language.
- **Semantic** - Refers to the meaning of language. It's all about how words and phrases relate to one another to create meaning within a larger context.
- **Query** - Is a way to ask a computer to find specific information for you.
- **Natural Language Inference (NLI)** - Is the task of checking if one sentence can be logically inferred from another sentence.

- **Artificial Intelligence (AI)** - Is the ability of machines or computer systems to perform tasks that would typically require human intelligence, such as learning, problem-solving, and decision-making.
- **Sequence-to-Sequence (Seq2Seq)** - A type of machine learning model that can translate one sequence of data into another.
- **CNN corpus** - The CNN corpus is a collection of over 300,000 news articles published by CNN between 2000 and 2016, used for natural language processing research and development.
- **GPT-3** - Is a powerful language model developed by OpenAI that uses deep learning to generate human-like text for various applications. It's one of the largest and most advanced language models available.
- **Large language models (LLMs)** - Are neural network models used in natural language processing (NLP) that are pre-trained on large amounts of text data to learn and represent the patterns and structures in language.
- **ROUGE metrics** - Is a set of metrics used to evaluate the quality of automatic summarization and machine translation outputs.
- **Natural Language Inference (NLI)** - Is an NLP task that involves determining the logical relationship between two given sentences, typically a premise and a hypothesis.
- **Doc2Vec** - Is a neural network model used for generating fixed-length feature vectors or embeddings for documents in natural language processing.
- **Word Embedding** - Is a technique used in NLP to represent words or phrases as numerical vectors, preserving their meaning and relationships.
- **Neural Networks** – Arre machine learning models inspired by the human brain that can recognize complex patterns in data.

- **ReLU** - Is an activation function used in neural networks to introduce non-linearity. It sets negative inputs to zero and leaves positive inputs unchanged.
- **Transformer Model** - Is a neural network architecture for NLP tasks that processes input in parallel using self-attention mechanisms.
- **Query in transformer model** - In Transformer model, query is the input used to compute attention scores between the current position or "word" being processed and all other positions in the input sequence.
- **Key in transformer model** - In the Transformer model, a key is a vector used to compute the attention score and determine relevant context during the self-attention mechanism.
- **Value in transformer model** - In the Transformer model, "values" are the input embeddings for each position in a sequence that are used to compute attention scores with the queries, and then a weighted sum of the values is calculated to provide context information for each position in the sequence.
- **Gradient descent** - Is an optimization algorithm used in machine learning to minimize model errors by adjusting the model parameters in the direction of the negative gradient of the loss function.
- **API calls** - Is a request made by a software to access data or functionality from a remote server using an API.
- **Iteration** - Is a process of repeating a set of steps or instructions until a desired outcome is achieved.

Adviser

- Add more Definition of Terms.

Dr. Floredith Jeanne Alcid

- Operational definition of terms should be technical definition of terms and add more important technical definition of terms

Conceptual Framework

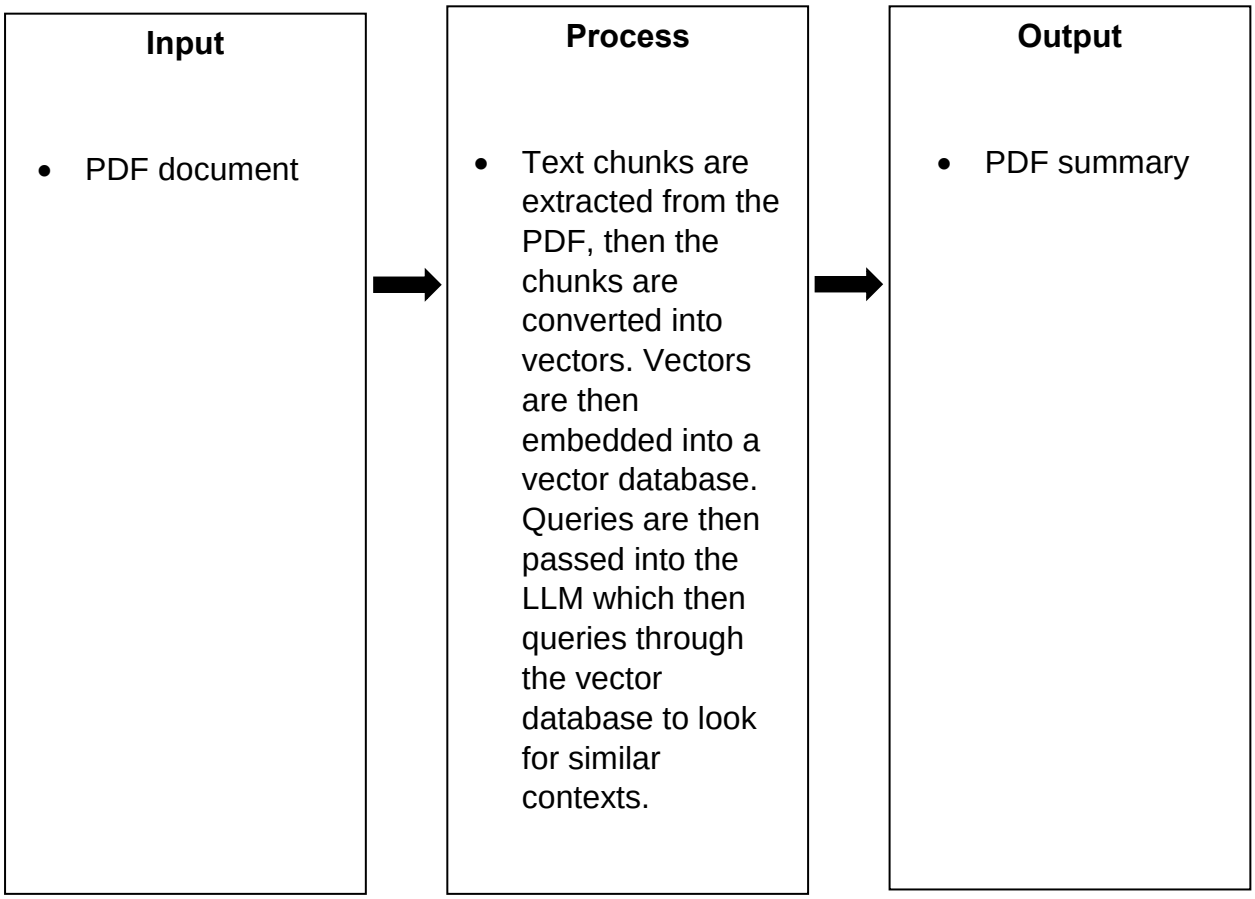
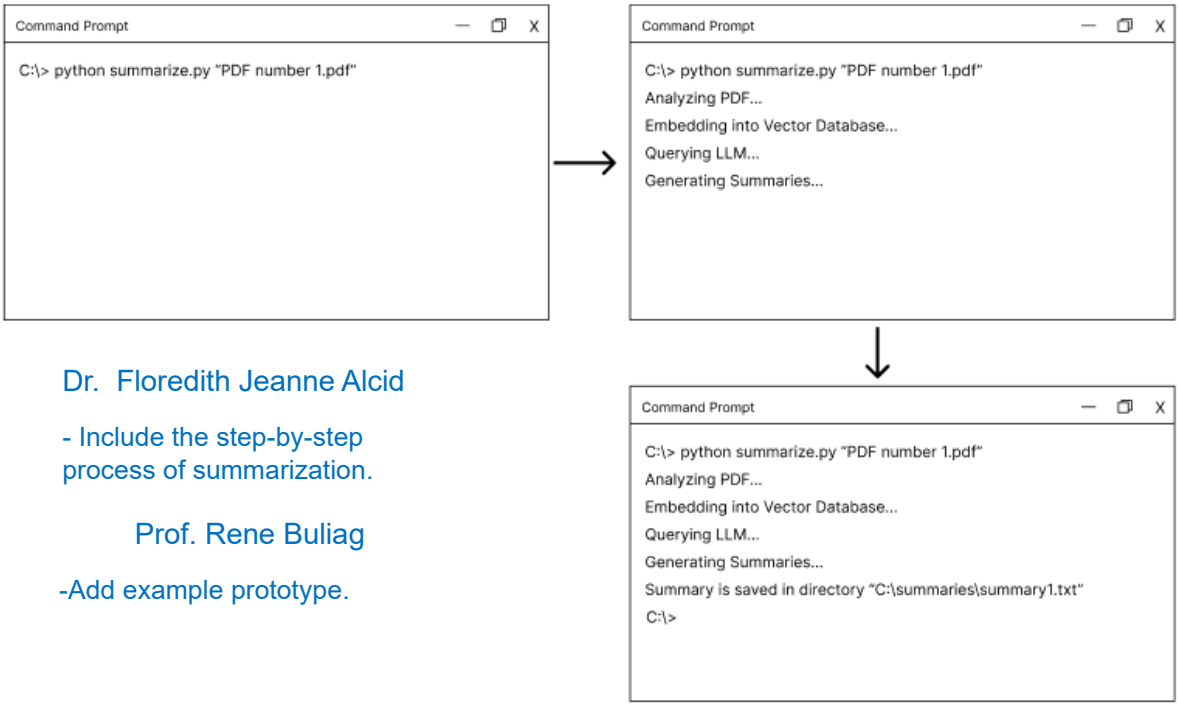


Figure 1: *Conceptual Framework*



Dr. Floredith Jeanne Alcid

- Include the step-by-step process of summarization.

Prof. Rene Buliag

-Add example prototype.

Figure 2: *Prototype (No GUI)*