PDF Question Answering System Report

Table of Contents

1. Introduction

Problem Statement: The primary objective of this project is to develop a system that can answer questions based on the content of PDF documents. Given the vast amount of information contained in PDFs, having a tool that can quickly extract and understand this information to answer specific questions is highly valuable.

Objectives:

- Extract text from PDF documents.
- Tokenize the extracted text into manageable sentences.
- Use a pre-trained Word2Vec model to create numerical vector representations of these sentences.
- Develop a mechanism to compare a user's question with these sentence vectors to find the most relevant answer.

2. Approach

The methodology of the project is divided into several key steps:

2.1 Extract Text from PDF: I used the PyPDF2 library to read and extract text from PDF files. This process involves reading each page of the PDF and compiling the text into a single string. This step is essential for

converting the document's content into a format that can be processed further.

2.2 Tokenize Sentences: Using the Natural Language Toolkit (NLTK), I broke the extracted text into individual sentences. Tokenizing the text helps in handling smaller, more manageable pieces of data, which are crucial for the subsequent embedding process.

2.3 Load and Use Word2Vec Model: I utilized a pre-trained Word2Vec model from the Gensim library. This model, trained on a large corpus of text, can convert words into vectors that represent their meanings. The model was downloaded, saved, and loaded to optimize processing time.

2.4 Create Sentence Embeddings: For each sentence, I created a numerical representation (embedding) by averaging the vectors of the words in the sentence. This approach gives a single vector for each sentence that captures its overall meaning.

2.5 Question Embedding and Similarity Calculation: When a user asks a question, it is converted into an embedding using the same method. This question embedding is then compared with all sentence embeddings from the PDF using cosine similarity. The sentence with the highest similarity score is considered the most relevant answer.

3. Failed Approaches

During the development of this project, several approaches did not yield satisfactory results:

Naive Approach:

- Initially, I tried using a simple Bag-of-Words (BoW) model. This model counts the occurrences of words but fails to capture their meanings and context. This approach resulted in poor performance because it could not understand the semantics of the sentences.

TF-IDF Vectorization:

- The Term Frequency-Inverse Document Frequency (TF-IDF) method was also tested. While TF-IDF considers the importance of words within the document, it struggles with understanding context and semantics, leading to less accurate answers. This method could not effectively match the user's question with the relevant sentences from the PDF.

4. Results

The final system successfully answers questions based on the content of PDF documents by using Word2Vec embeddings. The key results are:

- The system accurately retrieves relevant sentences from the PDF that answer the user's question.

- Cosine similarity is an effective metric for comparing the similarity between question embeddings and sentence embeddings.
- The use of pre-trained Word2Vec embeddings enhances the system's performance by capturing the semantic meaning of words and phrases.

5. Discussion

The results demonstrate that using pre-trained word embeddings significantly improves the performance of a question-answering system. The Word2Vec model captures the semantic meaning of words and phrases, allowing for more precise answer retrieval.

Significance:

- The project showcases the potential of word embeddings in natural language processing tasks, particularly in understanding and retrieving information from large text corpora.
- It highlights the importance of context and semantics in question-answering systems, as simple frequency-based models are insufficient for such tasks.

Insights:

- The quality and size of the pre-trained Word2Vec model play a crucial role in the system's effectiveness.

- More advanced models like BERT (Bidirectional Encoder Representations from Transformers) could further improve the system by providing better context understanding.

6. Conclusion

In conclusion, the PDF Question Answering System efficiently extracts text from PDFs, preprocesses it, and uses Word2Vec embeddings to answer user queries. This project demonstrates the effectiveness of using word embeddings in natural language processing tasks. Future improvements could involve:

- Incorporating more advanced models like BERT for better context understanding.
- Extending the system to handle more complex documents and questions.
- Enhancing the user interface to make the system more accessible and user-friendly.

7. References

- https://arxiv.org/pdf/1707.07328.pdf.
- https://arxiv.org/pdf/1810.04805.pdf.
- https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.
- http://jalammar.github.io/illustrated-transformer/