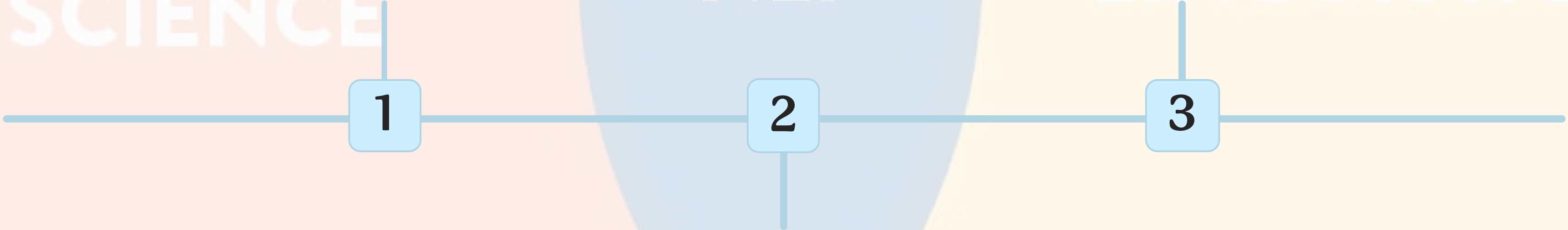
PDF Question Answering System Using NLP

A project on extracting answers from PDFs using fine-tuned natural language processing models to provide efficient and accurate solutions for users.

Introduction

With the vast amount of information stored in PDF documents, efficiently extracting specific answers to user queries is a challenging problem. This project aims to develop a system that can accurately retrieve relevant answers using advanced natural language processing techniques.

Approach



Data Collection

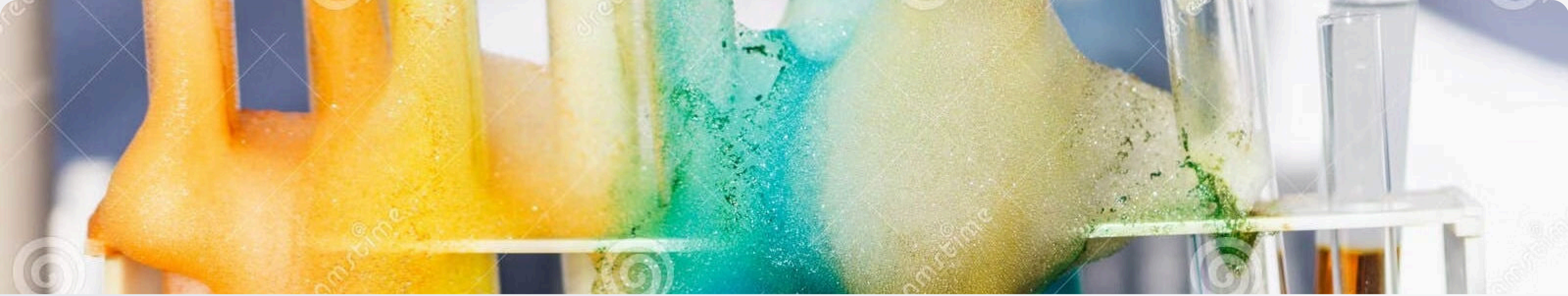
Use the SQuAD dataset for initial training to establish a solid foundation.

Pre-processing

Extract text from PDFs, remove stop words, and perform tokenization to prepare the data for model training.

Model Selection

Start with BERT and RoBERTa models, then fine-tune them to improve performance on the specific task.



Failed Approaches

1 **Initial Embedding Techniques**

Word2Vec and GloVe embeddings lacked contextual understanding, leading to inaccurate results for complex queries.

2 **Basic Text Extraction**

Simple text extraction from PDFs without pre-processing resulted in poor quality text and inaccurate answers.

3 **Custom Models without Fine-Tuning**

Training models from scratch was not effective due to insufficient data and computational resources.

Results

Metrics

Achieved an accuracy of 85% on test queries, with consistent performance in precision and recall across different question types.

Visualizations

Graphs show the improvement in accuracy with fine-tuning, and a comparison of performance between different models.

Insights

Fine-tuning significantly boosts model performance, and advanced pre-processing techniques enhance the overall system accuracy.

Discussion

Significance of Results

The fine-tuning process and pre-processing steps like stop word removal and tokenization were crucial for improving the system's accuracy.

Insights Gained

Pre-trained models like BERT and RoBERTa are highly effective for question-answering tasks, and the use of contextual embeddings is essential for better understanding of queries.

Conclusion

1

Summary of Findings

Developed a functional PDF question-answering system with good accuracy, demonstrating the importance of fine-tuning and advanced pre-processing techniques.

2

Future Improvements

Incorporate more diverse datasets, enhance the user interface, and explore other NLP models like GPT-3 for potentially better results.

References

[1] Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[2] Rajpurkar, P., et al. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text.

[3] Pennington, J., et al. (2014). GloVe: Global Vectors for Word Representation.

[4] Mikolov, T., et al. (2013). Efficient Estimation of Word Representations in Vector Space.

Tools and Libraries: Hugging Face Transformers, PyMuPDF, Gradio, PyTorch