CMPE255 Team Project: Automated Summarization of Research Papers

Team Name: Bay Area Rockers
Team Members: Shawn Chumbar, Dhruval Shah, and Sajal Agarwal
Project Title: Research Paper Summarization

The project plan is transcribed below as markdown for ease of reading.

Project Description

This project aims to develop an automated system for summarizing academic research papers, leveraging the power of Natural Language Processing (NLP) and deep learning. By implementing both extractive and abstractive summarization techniques, the project seeks to create concise, coherent summaries of lengthy research documents, facilitating easier comprehension and quicker review processes. Additionally, we plan to design a user-friendly web application that allows users to upload research papers in PDF format and receive summarized content.

Background

With the ever-increasing volume of academic literature, the ability to quickly comprehend and digest lengthy research papers is becoming crucial. Traditional manual summarization is time-consuming and labor-intensive. Automated text summarization using NLP and AI offers a promising solution, with potential to revolutionize the way researchers and academics interact with literature.

Objectives

- **Develop an Extractive Summarization Model**: To identify and extract key sentences and phrases directly from research papers.
- **Develop an Abstractive Summarization Model**: To generate concise paraphrased summaries that capture the essence of the research papers.
- **Design a Web Application**: To provide a platform for users to easily upload PDF documents and obtain summaries.
- **Compare and Analyze Performances**: To evaluate the effectiveness of both models in producing accurate and coherent summaries.

Methodology

1. **Data Collection**: Accumulating a dataset of over 50 research papers, each with corresponding summaries.
2. Data Preprocessing: Cleaning and preparing the text data for model training.
3. Model Training

a. **Extractive Summarization**: Develop models to identify and extract key sentences or phrases directly from the text. This includes techniques like sentence ranking based on relevance, clustering, or graph-based models.

b. **Abstractive Summarization**: Implement models that generate new text that summarizes the original content. Use advanced language models like ChatGPT-4 to understand and generate human-like text.

4. Model Evaluation
    a. **ROUGE Scores**: Use Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics to evaluate the quality of the summaries against reference summaries.

Technologies to be Used

- **Hugging Face's Transformers**: For pre-trained NLP models.
- **PyTorch or TensorFlow**: As the deep learning framework.
- **Google Colab**: For leveraging GPU resources for training and testing models.
- **Python**: As the primary programming language.
- **Web Development Frameworks**: Such as Flask or Django for building the web application.
- **Front-End Technologies**: Like HTML, CSS, and JavaScript for developing the user interface.

Expected Outcomes

- **Effective Summarization**: The ability to produce summaries that are both concise and representative of the original texts.
- **Model Comparison**: Insights into the comparative effectiveness of extractive vs. abstractive summarization techniques.
- **Scalability**: A framework that can be adapted for summarizing papers across various academic disciplines.

Project Deliverables

Please see below for a list of deliverables that we plan to complete for this project.

1. **Project Plan:** Document detailing the project's scope, objectives, methodology, and expected outcomes.
2. **Project Report:** Document detailing what we learned from performing this project. This document will also include the results of our analysis and any conclusions we draw from them, as well as any related charts and plots.
3. **README.md:** File which contains the project plan and details about deliverables. This file will also document how to use the Google Colab File, and details about the data used in the project.
4. **Google Colab Project File:** A Google Colab Notebook File containing the code for the project.

5. **Dataset:** Dataset containing the data that was used for this project. This dataset will be uploaded to the GitHub repository.
6. **Trained Models**: The final trained extractive and abstractive models.
7. **Codebase**: All code developed for the project, including data preprocessing, model training, and evaluation scripts.
8. **Documentation**: Comprehensive documentation detailing the methodologies, usage, and evaluation of the models.
9. **Web Application**: A functional web platform for users to upload PDFs and receive summaries.

Conclusion

This project aims to contribute significantly to the field of NLP by addressing the challenge of automated research paper summarization. The successful implementation of this project will not only demonstrate the capabilities of current AI and NLP technologies in understanding and processing complex academic texts but also pave the way for future advancements in automated text summarization.