Report: Text Summarization Project

For: Assignment-AI-NLP Intern-Emplay Inc

Prepared By: Aditi Sharma
Aditi0712sharma@gmail.com

Introduction

The text Summarization Project aimed to develop a system capable of summarizing content from Wikipedia articles. The primary objective was to extract relevant information and present it in a concise and coherent manner. The project involved several design decisions, encountered challenges, and required a thorough error analysis to ensure the quality of the generated summaries.

Design Decisions

1. Web Scraping and Content Extraction

- The project utilized web scraping techniques to extract content from Wikipedia articles.
- Design choice: Use BeautifulSoup for parsing HTML and extracting relevant text content.

2. Content Structuring

- The content was organized based on headings and subheadings to maintain a structured summary.
- Design choice: A dictionary structure was employed to store content under each heading and subheading.

3. Summarization Model

- The project incorporated the *Hugging Face* Transformers pipeline with the BART model for text summarization.
- Design choice: Utilize the 'facebook/bart-large-cnn' model for generating summaries.

4. Metric Evaluation:

- Multiple metrics (BLEU Score, ROUGE F1 Score, Flesch Reading Ease, etc.) were chosen to evaluate the quality of the summaries.
- Design choice: Calculate various metrics to assess information retention, readability, and coherence.

5. Random Paragraph Generation for Evaluation

- To perform quick evaluations, random paragraphs were generated for applying the summarization model.
- Design choice: Implement a function to generate random paragraphs from the content dictionary.

6. Metrics Evaluation Report:

- A detailed report was generated to interpret the metrics and provide insights into the performance of the summarization model.
- Design choice: Create a DataFrame to store and analyze metric scores for each generated summary.
- A random piece of content was selected and the model was evaluated using the evaluation metrics discussed above and the following report was generated:

Metric	Score	Interpretation
Average Sentence Length	13.30	Moderate sentence length, suggesting a balanced level of detail. Longer sentences might enhance retention.
Average Word Length	5.47	Reasonable word length, contributing to readability. A mix of word lengths improves overall engagement.
BLEU Score	1.0	Perfect match with reference summaries, indicating high precision and accuracy.
Cosine Similarity	1.0	Identical generated summaries, demonstrating high internal consistency and coherence.
Flesch Reading Ease	49.52	Moderately easy to read, falling within the standard readability range, contributing to information retention.
ROUGE F1 Score	1.0	Perfect overlap with reference summaries, reinforcing the high quality and coherence of the content.

Access this GitHub Repository to access the code using this link:

https://github.com/Aditi0712/Text-summarization-wiki

Challenges Encountered

1. Web Scraping Challenges:
- The Wikipedia article had complex HTML structures, leading to challenges in accurately extracting content.
- Solution: Implemented robust parsing strategies to handle variations in HTML.

2. Model Limitations:
- The summarization model has a maximum sequence length, leading to potential truncation issues for longer content.
- Solution: Implemented chunking of content to fit within the model's limitations.

3. Warning Messages:
- Warning messages were encountered during metric evaluation, such as the BLEU Score warning about 0 counts of 4-gram overlaps.
- Solution: Addressed warnings by considering smoothing functions for BLEU Score calculations.

4. Time and Resource Constraints
The project encountered time and resource constraints, especially when dealing with a large volume of content. Optimizing the code and managing computational resources effectively became crucial to meet project deadlines.

Error Analysis

1. Warnings:
- Warning messages, particularly the BLEU Score warning, indicated potential issues with the reference summaries.
- Warnings related to token length accepted by models.
- Analysis: Investigated the cause of warnings and adjusted the evaluation approach to improve accuracy.

2. Metric Interpretation:
- The metric scores, such as Flesch Reading Ease and Cosine Similarity, required interpretation to understand their implications.
- Analysis: Conducted a thorough examination of metric definitions and their relevance to summarization quality.

Resources Utilized

1. Kaggle : As an IDE for implementing the code.
2. Hugging Face : To access the pretrained model.
3. BeautifulSoup documentation: For Web Scraping.
4. Text summarization Methods: To learn about text summarization techniques.