Text Summarization with SpaCy and BART

This script explores two primary methods for summarizing text: extractive and abstractive. It utilizes the CNN/Daily Mail dataset, a collection of news articles and their corresponding summaries, and also includes a real-world example for demonstration.

1. Data Preparation:

- The script begins by loading the CNN/Daily Mail dataset using the datasets library.
- It then applies a preprocessing step to clean the text data. This involves removing unnecessary newline characters and extra spaces to ensure the text is well-formatted for analysis.

2. Extractive Summarization using SpaCy:

- This approach leverages the SpaCy library and its English language model (en_core_web_sm) for natural language processing.
- It focuses on identifying and extracting the most important sentences from the original text to form the summary.
- A simple heuristic is used to select the top 3 longest sentences as the summary, assuming they contain the most crucial information.

3. Abstractive Summarization using BART:

- This method utilizes the pre-trained BART model, specifically the facebook/bart-large-cnn version, which is designed for abstractive summarization tasks.
- BART generates a concise summary by understanding the meaning of the input text and paraphrasing it, rather than simply extracting sentences.
- A summarization pipeline is created using the BART model and its tokenizer to streamline the summary generation process.
- To ensure the summaries are both informative and concise, length constraints are applied, limiting the output to a specific range of words.

4. Evaluation and Demonstration:

- To illustrate the effectiveness of both summarization methods, the script uses two examples:
    - A sample article is selected from the test set of the CNN/Daily Mail dataset.

    - A real-world news article about SpaceX is used to showcase the practical application of the techniques.
- Both extractive and abstractive summaries are generated for each example. To see the output, run the code. This allows for a comparison of the two approaches and their ability to capture the essence of the original text.

Conclusion:

This script provides a basic implementation of extractive and abstractive text summarization using the SpaCy and BART libraries, respectively. By applying these techniques to both a standard dataset and a real-world scenario, it demonstrates their effectiveness in generating informative and concise summaries from textual data.