Project #2 for the Biomedical Information Retrieval Course

Name: Phan Ben (潘班)

ID Student: P76127051

Project Overview:

This project focuses on text analysis and processing using advanced techniques. It involves calculating the Zipf Distribution for a set of PubMed text documents, implementing Porter's algorithm for text normalization, and comparing the outcomes. The system also allows dynamic programming-based Edit distance computations for text matching and presents retrieval results, making it a versatile tool for research and analysis. The choice of programming language is flexible.

Key Features:

1. **Download PubMed Articles :** The application simplifies the acquisition of PubMed articles by enabling users to directly download content from the PubMed Central (PMC) database. This feature supports local storage and future analysis of articles.

2. **Frequency Analyst:** The Frequency Analyst tool offers a suite of advanced algorithms and features for deepening the understanding of PubMed articles:

- **Zipf Distribution Computation:** The application includes a visual representation of the Zipf Distribution, showcasing the distribution of term frequencies within the dataset. This visualization is instrumental in identifying term significance and fine-tuning research efforts.
- **Porter Stemmer:** Porter Stemmer, a renowned algorithm for text normalization, is integrated into the application to process keywords and terms. This improves the consistency and relevance of keyword analysis and searches.
- **Edit Distance Computation:** Users can calculate the edit distance between keywords and terms within the articles. This feature is invaluable for locating approximate matches and retrieving relevant results, enhancing the efficiency of literature research.

Conclusion:

The PubMed Article Analyzer is a powerful platform for biomedical literature research. It offers advanced algorithms for PubMed article downloads, keyword analysis, Zipf Distribution visualizations, edit distance calculations, and text normalization. This tool streamlines research for professionals, researchers, and students, enhancing efficiency and results.