Project #4 for the Biomedical Information Retrieval Course

Name: Phan Ben

ID Student: P76127051

Project Overview:

This project involves implementing and analyzing term weighting technology for text documents in the vector space. It focuses on integrating Porter's algorithm and explores 2-3 TF-IDF methods, including Standard TF-IDF, Smoothed TF-IDF, or Probabilistic TF-IDF. The goal is to rank documents using a chosen similarity measure, either cosine or euclidean.

Key Features:

1. ***Text Preprocessing Options***
    - **Lowercasing:** Convert all text to lowercase.
    - **Remove Stopwords:** Eliminate common words that do not contribute significantly to document meaning.
    - **Porter Stemming:** Reduce words to their root form for improved analysis.
2. ***TF-IDF Algorithms***
    - **Standard TF-IDF:** Calculates the standard TF-IDF weights for document analysis.
    - **Smoothed TF-IDF:** Incorporates sublinear term frequency and smooth inverse document frequency for enhanced performance.
    - **Probabilistic TF-IDF:** Implements a custom probabilistic TF-IDF calculation.
3. ***Similarity Measures***
    - **Cosine Similarity:** Measures the cosine of the angle between two vectors.
    - **Euclidean Distance:** Represents the "inverse" of the Euclidean distance between TF-IDF vectors.
4. ***Visualization***
    - **Ranked Documents:** Displays documents sorted by similarity to the input text.
    - **Total Similarity per Category:** Offers a bar plot ranking document categories based on total similarity.

Conclusion:

In summary, this project presents a flexible framework for term weighting in text document analysis. The integration of Porter's algorithm, coupled with the option to choose from various TF-IDF methods and similarity measures, offers a customizable approach to document ranking. This project contributes to a deeper understanding of text mining techniques and their applications in information retrieval.