**NTLK:**

In this study, the effectiveness of the Universal Sentence Encoder (USE) in measuring text similarity is investigated. The methodology involves preprocessing the textual data to ensure standardization and then leveraging the pre-trained USE model, which is renowned for its ability to generate dense embeddings that encapsulate semantic meaning. By utilizing USE embeddings to compare pairs of sentences from two text files, the study demonstrates the model's capacity to accurately capture semantic similarity, yielding an impressive average cosine similarity score of approximately 86.4119%. This performance underscores the robustness of USE in capturing nuanced semantic relationships between text fragments, showcasing its suitability for various natural language processing tasks. The results affirm the utility of USE in facilitating tasks such as information retrieval and document summarization. Nonetheless, further exploration is warranted to assess the model's generalizability across diverse datasets and contexts. Overall, this study highlights the efficacy of the Universal Sentence Encoder in advancing the field of text processing and analysis, underscoring its importance in the broader landscape of natural language processing research and application.

**Gensim:**

The research methodology outlined in this study focuses on the assessment of semantic similarity between sentences through the utilization of pre-trained word embeddings. Employing the renowned 'word2vec-google-news-300' model, the investigation embarks on a journey to elucidate the intricate semantic relationships inherent within textual data. The methodology begins by adopting a robust preprocessing approach, incorporating the application of regular expressions to standardize the text. This step involves the removal of punctuation and the conversion of all characters to lowercase, ensuring consistency across the dataset. Subsequently, the study introduces a cosine similarity calculation method, facilitating the evaluation of semantic similarity between pairs of sentences. By leveraging the computed word embeddings, the cosine similarity metric serves as a reliable measure to quantify the semantic relatedness of textual content.

Central to the research endeavor is the meticulous evaluation of the proposed methodology's efficacy. Through a meticulous examination of pairs of sentences drawn from distinct text files, the study demonstrates a remarkable accuracy of 91.21716308713199% in capturing semantic similarity. This significant achievement underscores the proficiency of the employed word embedding model in encapsulating semantic nuances within textual data. Such a high accuracy rate substantiates the viability of the proposed methodology for diverse natural language processing tasks, including but not limited to information retrieval, document summarization, and sentiment analysis. Moreover, the findings of this study underscore the indispensable role of word embeddings in advancing the field of natural language understanding and fostering innovative solutions to complex linguistic challenges.

**BERT:**

The research methodology elucidated in this study is dedicated to exploring semantic similarity estimation between pairs of sentences, employing state-of-the-art BERT-based embeddings. The methodology commences with a rigorous text preprocessing phase, encompassing tokenization, stop-word elimination, punctuation removal, and lemmatization, ensuring the normalization and standardization of the textual data. Subsequently, leveraging the 'bert-base-nli-stsb-mean-tokens' pre-trained model, the study harnesses BERT-based embeddings for sentence representation, encapsulating rich semantic information. Pairs of sentences sourced from two distinct text files are subjected to similarity assessment using cosine similarity metrics based on BERT embeddings. Notably, the results showcase a substantial average cosine similarity score of 93.1153025344064%, affirming the robustness and efficacy of BERT embeddings in capturing nuanced semantic relationships within textual data. This exemplary performance underscores the potential of BERT-based approaches in advancing natural language understanding tasks, paving the way for innovative solutions in various domains such as information retrieval, sentiment analysis, and text summarization.

**Ensemble:**

The research methodology presented in this study is dedicated to the comprehensive assessment of semantic similarity between pairs of text segments. The methodology integrates various techniques, including pre-processing steps and the utilization of multiple semantic similarity models. Initially, the text data undergoes rigorous pre-processing, encompassing tokenization, lemmatization, and stop-word removal to standardize the textual content. Subsequently, the study employs three distinct semantic similarity models: a custom-built model leveraging Natural Language Toolkit (NLTK), ConceptNet, and a pre-trained BERT-based Sentence Transformer model. Each model provides a unique perspective on semantic similarity, offering diverse insights into the textual content. Furthermore, the study employs an ensemble learning approach to aggregate predictions from the individual models, aiming to enhance overall predictive performance. The ensemble model combines predictions from the base models, leveraging their complementary strengths to achieve a more robust assessment of semantic similarity. The methodology concludes with the calculation of an average ensemble prediction score, providing a comprehensive evaluation of semantic similarity across the text segments. Notably, the accuracy of the ensemble prediction, as measured against ground truth data, is reported to be 85.27%. Through this multi-faceted approach, the study aims to offer a nuanced understanding of semantic relationships within textual data, fostering advancements in natural language understanding and related fields.