#Summarize the entire interface in one paragraph.

#The first interface options allow you to choose the document type and language.

#There are then options for what defines a word or character as well as options to choose how many words you want to use for the analysis.

#The statistics setting allows you to choose statistical methods and define the delta distance for analysis

#The sampling tab lets you choose your sampling type and size.

#The output tab gives you settings for your analysis output including file type and plot size and features.

#Make two tables, one for the statistical approaches and one for Delta Distances

#Cluster Analysis: This approach performs hierarchical clustering of analyzed texts, producing a dendrogram that illustrates the similarity between texts.

#Multidimensional Scaling (MDS): MDS visualizes the dissimilarities between texts in a low-dimensional space, enabling the detection of underlying patterns or structures.

#Principal Component Analysis (PCA) using a covariance matrix: PCA identifies the principal components in the data, revealing the main sources of variation among texts.

#Principal Component Analysis (PCA) using a correlation matrix: Similar to covariance-based PCA, this method uses correlations to uncover underlying patterns, particularly effective for English texts.

#Consensus Tree: This approach combines multiple cluster analyses results for varying parameters to produce a statistically justified representation of text similarities.

#Classic Delta: Utilizes normalized word frequencies to measure the dissimilarity between texts, offering a reliable method for comparing texts with different word counts.

#Argamon’s Linear Delta: Applies Euclidean distance to normalized word frequencies, making it sensitive to the number of texts in a corpus and suitable for analyzing large datasets.

#Eder’s Delta: A modified version of standard Burrows’s distance, enhancing the weighting of frequent words and rescaling less frequent ones to minimize the influence of random uncommon words, particularly useful for highly inflected languages.

#Eder’s Simple: Provides a straightforward normalization technique independent of corpus size, effectively accounting for Zipf’s law implications in word frequency distributions.

#Canberra Distance: Effective for languages with sparse frequencies and sensitive to rare vocabulary usage, though prone to noise due to its sensitivity to minor differences in word occurrences.

#Cosine Distance: Measures the cosine of the angle between word frequency vectors, offering a classical approach to quantifying textual similarity, especially suitable for high-dimensional spaces.

Statistical Operations:

Cluster analysis – Uses algorithms to group sets of object or data points into different groups based on similarities. The purpose of cluster analysis is to reveal structures and patterns in the data set to reveal underlying patterns, relationships, and associations.

MDS(Multidimensional scaling) vs PCA(Principle component analysis) –

Similarly to cluster analysis , MDS and PCA both focus on grouping data. The main difference is in how they approach it. MDS focuses on pairwise data while PCA focuses on minimizing variance or covariance. MDS works on a distance matric while PCA works on the original data matrix.

t-SNE – is a non-linear statistical method much like the three previous. It groups objects and data based on similarities but does it in higher dimensions as well.

Consensus tree – A way to group data by similarities but it also defines similarities between groups by classifying them in branches