Experiment No.6

Title: Applying similarity measures on the Textual datasets and Handling Skew in Dataset.

Batch: **Roll No.:** **Experiment No.: 5**

Aim: Applying similarity measures on the textual datasets and Handle skewness in dataset.

Resources needed: Any programming language, any data source (RDBMS/Excel/CSV)

Theory:

Textual similarity refers to the degree of likeness or resemblance between two pieces of text. It is a measure of how closely related or similar two texts are in terms of their content, structure, or meaning. Textual similarity can be assessed using various techniques and algorithms, depending on the specific aspects of similarity that you want to measure.

In this lab session will explore these three Textual similarity in data:

Cosine Similarity: Cosine similarity is a commonly used technique to measure the similarity between two text documents. It calculates the cosine of the angle between two vectors representing the texts in a high-dimensional space. The closer the cosine similarity score is to 1, the more similar the texts are.

Cosine Distance = 1 - Cosine Similarity

Cosine Similarity $(A, B) = (A \cdot B) / (\|A\| * \|B\|)$

Where, $(A \cdot B)$ represents the dot product of vectors A and B.

$\|A\|$ and $\|B\|$ represent the Euclidean norms (magnitudes) of vectors A and B, respectively.

To calculate cosine distance between two vectors, you subtract the cosine similarity value from 1. A value of 0 indicates no similarity (orthogonal vectors), and a value of 1 indicates perfect similarity (vectors point in the same direction).

Jaccard Similarity: Jaccard similarity measures the similarity between two sets by comparing their intersection and union. In text analysis, it's often used for comparing the similarity of two sets of words, such as the words in two documents.

Levenshtein Distance (Edit Distance): This metric quantifies the minimum number of single-character edits (insertions, deletions, or substitutions) needed to transform one text into another. It is often used for measuring the similarity between two strings.

Computing the edit distance, also known as the Levenshtein distance or edit distance, between two strings is a common problem in computer science and text processing. The edit distance measures the minimum number of edit operations (insertions, deletions, and substitutions) required to transform one string into another. You can compute the edit distance using dynamic programming.

Longest Common Subsequence(LCS) : LCS which is a technique used in string comparison and text similarity measurement. The Longest Common Subsequence (LCS) is a dynamic programming algorithm that finds the longest subsequence that two sequences of characters share, without requiring consecutive elements. It is a measure of similarity between two strings or textLCS is useful for measuring the similarity between two texts or strings when you want to find common subsequences that may not be consecutive. It's often used in applications like plagiarism detection, DNA sequence alignment, and text comparison.

Skewness Computation:

Skewness is a statistical measure that describes the asymmetry or lack of symmetry in the distribution of data points in a dataset. It indicates whether the data is skewed to the left (negatively skewed), skewed to the right (positively skewed), or if it has a relatively symmetrical distribution.

A symmetrical distribution has a balanced, bell-shaped curve, and it is neither positively nor negatively skewed. In a symmetrical distribution, the mean, median, and mode are approximately equal.

How to Calculate Skewness:

1. There are several methods to calculate skewness, but one of the most commonly used methods is Pearson's First Coefficient of Skewness formula:
2. Skewness=3(Mean−Median)Standard DeviationSkewness=Standard Deviation3(Mean−Median)
3. Here's a step-by-step guide to calculating skewness:
4. Calculate the mean (average) of the dataset.
5. Calculate the median (middle value) of the dataset.
6. Calculate the standard deviation of the dataset.
7. Plug these values into the formula to calculate skewness.

Central Moments:

Central moments are statistical measures that provide information about the shape and distribution of data around its mean. The central moments of data are typically calculated using the following formulas:

1. The first central moment (mean) is simply the mean of the data.
2. The second central moment (variance) measures the spread or dispersion of the data.
3. The third central moment measures the asymmetry or skewness of the data distribution.
4. The fourth central moment measures the kurtosis, which indicates the tails and the presence of outliers in the data.

Procedure / Approach /Algorithm / Activity Diagram:

1. Identify the suitable attributes to apply the textual similarity measures among one of them and write python code to calculate Longest common subsequence, edit distance, cosine or Jaccard similarity measures on it.
2. Find skewness in data with any of the correlation coefiicent viz. Pearson coefficient ,Bowleys coefficient .Finally find the different moments among data. Write a python code for computation of skwenss.

Results: (Program printout with output / Document printout as per the format)

Questions:

1. What are the different applications of Textual similarity measure?
2. What are the different applications of finding similarity between textual attributes?

Outcomes:

Conclusion: (Conclusion to be based on the objectives and outcomes achieved)

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

References:

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3^{nd} Edition

2. Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to data mining. Pearson Education India, 2016.