TEXT SIMILARITY

PART A: BUILD A MODEL

Initially the required libraries were imported. The dataset was loaded into a variable namely, ‘data’. The shape of the data described that it contains 3000 instances present in 2 columns (text1 and text2). There were no missing values. The data had to be pre-processed in order to perform the further functionalities.

Pre-process:

1. **Punctuations**
   The punctuations consist of special characters that were present in the text.

2. **Lower**
   The entire texts were converted into lower case.

3. **Tokenization**:
   The texts were broken into words.

4. **Lemmatization**
   Replaced the words that can be synonymously used.

5. **Digits**
   The digits present in the text were removed.

6. **Space**
   Unwanted spaces were removed.

Using the pipe function all the above mentioned pre-processing stages was removed.

MODEL:

CountVectorizer and TfidfVectorizer were used to compute the similarity scores. The developed scores were added as additional columns in the dataset. TfidfVectorizer is better than Count Vectorizers because it not only focuses on the frequency of words present in the corpus but also provides the importance of the words.