NLP task 2-

Text Similarity:

We will begin with creating a function to read the text files we want to check for similarity. Then we will remove punctuations and convert all words to lower case by using the translation table.

Next we will make a function that counts the frequency of the words in the files. In this function we will have a dictionary that takes the word as key and we can iterate through the word list to get count of each word.

For calculating the cosine similarity we need to calculate the dot product of the two word lists of the given files. To do this we will iterate throungh the word lists and if a word is present in both the lists then the product of the count of word in list 1 and list 2 will be added to the dot product.

To calculate modulus of the word list vectors we will take thier dot product with the same vectors. Cosine of the angle between the vectors will be given by dot product of both vectors divided by the modulus of both vectors.

Hence similarity of the documents will be given by the cosine of the angle between the vectors of the word lists of the files