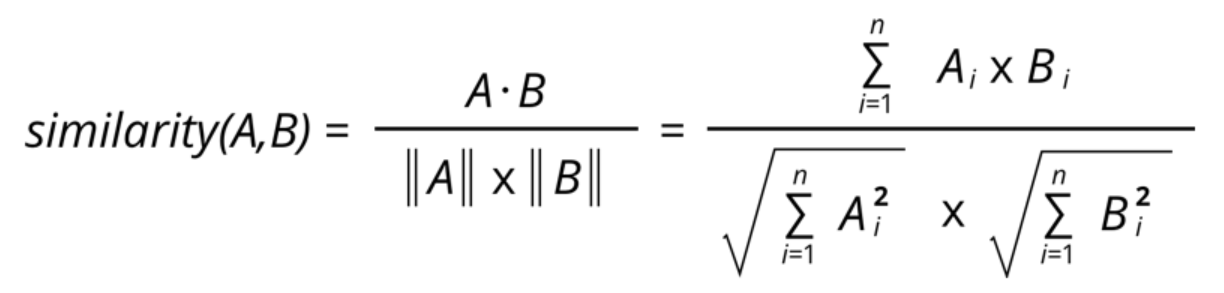
There are various ways to calculate the similarity between 2 texts/documents. The metric I have used is cosine similarity. In this method, we will create 2 vectors based on unique words from both the texts. The formula for calculating cosine similarity is:



**Following assumptions are taken into consideration while implementing this code:**

1. Only English text is accepted as input in order to compute an optimal score.

2. All the punctuations and special characters are ignored in the comparison.

3. All the English stop words (the words that are used frequently) are not counted.

4. Case is ignored while comparing.

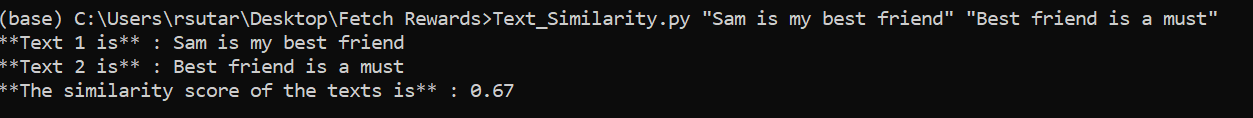
**In order to run the code, please follow the below steps:**

1. Save the file 'Text\_Similarity.py' to a folder

2. Open the terminal and change the directory to the one where the above file is saved.

3. Run the command - Text\_Similarity.py text1 text2 (text1 and text2 are texts to be compared and passed as arguments here)

**Below is the screenshot with an example of command line and the output:**



**An example is given below for illustration:**

Text1 list 🡪 ['sam', 'is', 'my', 'best', 'friend']

Text2 list 🡪 ['best', 'friend', 'is', 'a', 'must']

Text1 frequency of words after ignoring the case, punctuation and stop

words 🡪 {'sam': 1, 'best': 1, 'friend': 1}

Text2 frequency of words after ignoring the case, punctuation and stop

words 🡪 {'best': 1, 'friend': 1, 'must': 1}

Unique words from Text1 and Text2 🡪 {'best', 'must', 'friend', 'sam'}

Text1 vector 🡪 [1, 0, 1, 1]

Text2 vector 🡪 [1, 1, 1, 0]

Score (calculated using above formula) 🡪 0.67