Q1.) How to Compute the similarity between two text documents.

ANS. The Document similarity is calculated by calculating the Document Distance.

- Document Distance is a concept where words (documents) are treated as Vectors and is calculated as the Angle between two given Document Vectors.
- Document vectors are the Frequency of Occurrences of words in a given document.
- For Example:-

Given two documents D1 and D2 as:-

D1 : "This is a Geek"

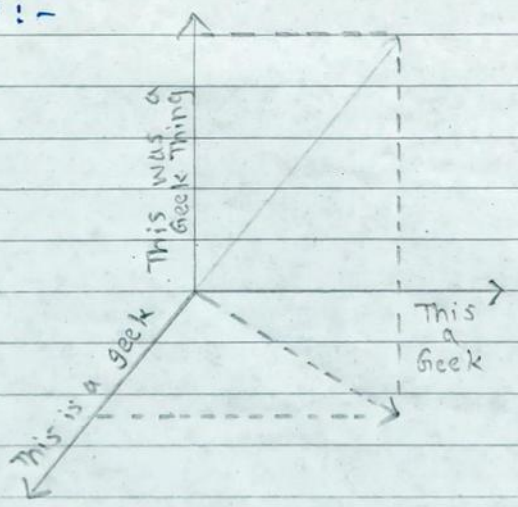
D2 : "This was a Geek Thing"

The similar words in both these documents then become.

"This is a Geek"

⇒ The 3D Representation of this as Vectors by taking D1, D2 is:-

Diagram:



If we take a dot product of two document vectors D1 and D2 is:-

1. D2 =

"This", "This" + "~~was~~", "~~was~~ is" + "a", "a" + "geek", "geek" + "thing", "0"

D1·D2 = 1 + 0 + 1 + 1 + 0 = 3

D1·D2 = 3

Now, We can calculate the Angle between the Document

$$\cos d = \frac{D1 \cdot D2}{|D1||D2|}$$

Here, d is the document distance.
Its value ranges from 0 to 90 degrees.
where, i.) 0° means cos(0) = 1.
The two documents are exactly identical.
ii.) 90° means cos(90) = 0
The two documents are very different.

CONCLUSION:-

→ Hence, We have studied how to compute the similarity between two text documents using a Python program.