**PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE**

**ACADEMIC YEAR: 2023-24**

## **DEPARTMENT OF COMPUTER ENGINEERING DEPARTMENT**

**CLASS: B.E. SEMESTER: I**

**SUBJECT: LP-IV**

|  |  |
| --- | --- |
| **ASSIGNMENT NO.** | A1 |
| **TITLE** | **Compute similarity between two text documents.** |
| **PROBLEM STATEMENT /DEFINITION** | Write a program to compute similarity between two text documents. |
| **OBJECTIVE** | To compute similarity between two text documents by making use of the **Vector Space** model.  To understand the working of the **Vector Space** model. |
| **OUTCOME** | Students will be able to -   * Compute the similarity between two text documents. * Understand the working of the **Vector Space** model. |
| **S/W PACKAGES AND**  **HARDWARE APPARATUS USED** | Windows 10 (64-bit),  Intel I5 4GB RAM 256 GB SSD,  Python 3.9.0,  VS Code |
| **REFERENCES** | 1. C.J. Rijsbergen, &quot;Information Retrieval&quot;, (http://www.dcs.gla.ac.uk/Keith/Preface.html)  2. W.R. Hersh, ―Information Retrieval: A Health and Biomedical Perspective‖, Springer, 2002.  3. G. Kowalski, M.T. Maybury. &quot;Information storage and Retrieval System&quot; , Springer, 2005 |
| **STEPS** | Refer to theory, algorithm, test input, test output |
| **INSTRUCTIONS FOR**  **WRITING JOURNAL** | 1. Date  2. Assignment no.  3. Problem definition  4. Learning objective  5. Learning Outcome  6. Concepts related Theory  7. Algorithm  8. Test cases  10. Conclusion/Analysis |

**Prerequisites:**

**Concepts related Theory:**

**Vector Space model**

Vector space model or term vector model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers (such as index terms). It is used in information filtering, information retrieval, indexing and relevancy rankings. Its first use was in the SMART Information Retrieval System.

Documents and queries are represented by vectors as follows:

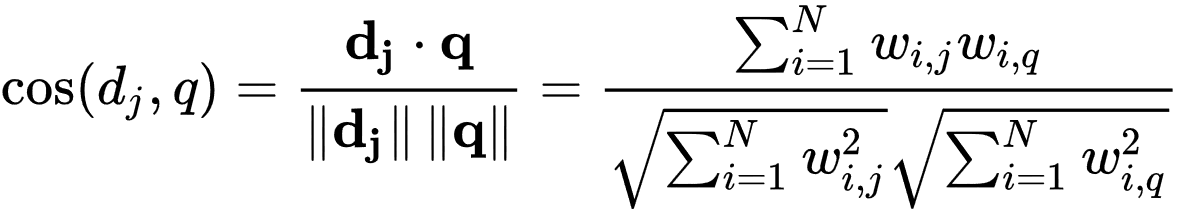
**dj  = ( w1,j , w2,j, …….. wn,j )**

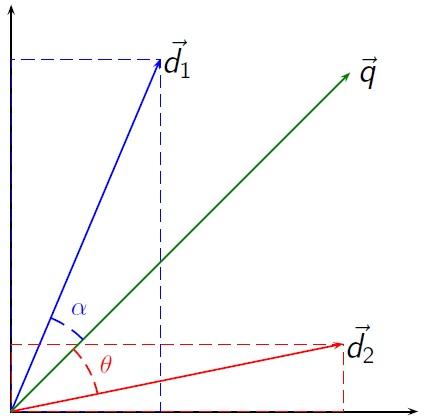
**q = ( w1,j , w2,j, …….. wn,j)**

Our aim is to find the weight vector based on the frequency or occurrences of terms.

The definition of term depends on the application. Typically terms are single words, keywords, or longer phrases. If words are chosen to be the terms, the dimensionality of the vector is the number of words in the vocabulary (the number of distinct words occurring in the corpus).

The correlation between the document **dj**andthe query **q** is given as the cosine of the angle between two vectors.





If the value of cos(dj, q) is less than the document is more similar and vice versa then we can get a degree of similarity between 0 and 1. A cosine value of zero means that the query and document vector are orthogonal and have no match (i.e. the query term does not exist in the document being considered) while a cosine value of 1 means that the query and document vector are exactly identical.

**Algorithm:**

1. Read the text file.
2. Tokenizing the lines into words.
3. Removing the punctuation marks and stop words.
4. Counting the frequency of each word in the query and document.
5. Taking the dot product of the query vector and the document vector.
6. Finding the angle value by taking the cosine inverse of the dot product.

**Conclusion:** We have successfully learned and implemented a vector space model and found similarity between the two text documents.

**Review Questions**:

Q1. What is text Summarization?

Q2. What is tokenization?

Q3. What do you mean by Text Extraction and Cleanup?

Q4. What are the steps to follow when building a text classification system?