**PSG COLLEGE OF TECHNOLOGY**

**Department of Applied Mathematics and Computational Sciences**

**VIII Semester M.Sc. SS – 2023-2024**

**Information Retrieval Lab**

**Exercise 2**

1. To maintain the integrity of your data, it’s a good idea to reduce duplicate documents in the system. Desktop Plagiarism Checker helps us to detect documents. The database has N documents. In each document, the title is separated from the rest of the text by: (colon). A Plagiarism Checker reads a new document and adds it to the database if it is not a duplicate of any document in the database. Design a Plagiarism Checker with the following features.

A) Verify if the titles are exactly same (Apply BinaryDistance(u,v), which gives the binary distance between vectors u and v, equal to 0 if they are identical and 1 otherwise.). If same, label the document as duplicate and discard it else proceed to second part of the Checker.

B) Represent documents (excluding the title) as term document vectors with weight of a term in a document computed as



C) Identify a document as duplicate if the similarity of the document is more than the threshold α. (α=0.85). Apply cosine similarity for similarity computations.

D1: information requirement: query considers the user feedback as information requirement to search.

D2: information retrieval: query depends on the model of information retrieval used.

D3: prediction problem: Many problems in information retrieval can be viewed as prediction problems

D4: search: A search engine is one of applications of information retrieval models.

New documents

D5: Feedback: feedback is typically used by the system to modify the query and improve prediction

D6: information retrieval: ranking in information retrieval algorithms depends on user query