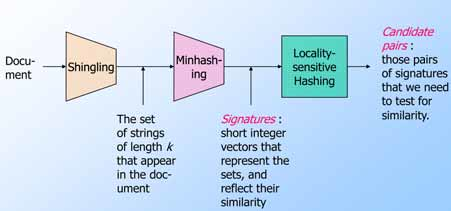
**INFORMATION RETRIEVAL - ASSIGNMENT 2**

|  |  |
| --- | --- |
| **Name** | **ID** |
| **Poosarla Divakar** | **2017A7PS0225H** |
| **Yadala Akhil** | **2017AAPS0456H** |
| **Nihal Jain** | **2017A7PS0325H** |
| **Praneetha Vaddamanu** | **2017AAPS0319H** |

**Project Architecture**

Objective of this assignment is to develop a plagiarism detector using Locality Sensitive Hashing (LSH) algorithm. This Document describes the architecture implemented for plagiarism detection. Corpus consists of hundred documents of average length 200 words. Locality Sensitive Hashing is an efficient technique that groups similar documents by hashing them to same bucket. This process involves three main steps namely Shingling, Min hashing, LSH.



Shingling is performed to obtain set of characters of length 6 from each document of corpus. Document matrix is constructed by representing each document as a set of shingles. Corpus is preprocessed prior to shingling by removing special characters and case folding to lower case. Document Matrix is too sparse to fit in primary memory for large corpus. Min hashing is performed to resolve space-complexity by converting each document to a small signature using hundred randomly generated hash functions with out loosing similarity between documents. LSH is then used on signature matrix to identify candidate pairs (similarity greater than desired threshold t) which involves partitioning signature matrix into b bands each having r rows. Values of b and r are tuned such that when each column of each band of signature matrix is hashed, similar documents (candidate pairs) hash to same bucket.