***Plagiarism Checker***

**Objective**:- The project is to retrieve plagiarised documents based on the document query given by the user from the database containing 100 documents of which some are plagiarised.

**Architecture**:- The database consists of 100 documents. All the docs are first changed into lowercase. Shingles of length 8 are made for each document and stored in a list for all the documents in the corpus. A dictionary of lists is made where each key is a shingle and the corresponding value is a list containing all the docid’s having that shingle. A Boolean matrix(npmatrix) of (shingle-docid) is made which indicates whether the shingle is present in the corresponding doc or not.

**Flow of code:-**

* For the creation of the hash matrix, a total of 96 hash functions are created.
* Hash Matrix is (num\_of\_shingles-num\_of\_hash\_functions) where each hash function is of the form (a+bx) where a,b both are independent hash functions.
* Sign matrix is developed from the hash matrix created above which is of the form (num\_of\_hash\_functions-docid’s)

**Query Processing:**

- Whenever the query is given, the signature matrix is divided into bands and rows.

- For each band all the partial columns are hashed into buckets.

The bucket containing the query docid is taken and all other docs in that bucket are stored in a set.

-The above procedure is performed for each band ans the set is returned as answer.

**Retrieval Time:**

Runtime=(num\_of\_hash\_functions\*num\_of\_documents)

In this case Runtime= O(96\*100).

Precision(Varies with query):

Avg=68%

Max=87%