**README**

Input file used : A csv file called news\_summary.csv containing 4500 articles

Python Libraries used:

* numpy
* pandas
* math
* os

To run the project save news\_summary.csv in the same folder where the executable Shingles.py file is stored. Give threshold value and no of documents as input to the file to obtain the following results:

* All candidate pairs
* Jaccard score of all candidate pairs
* Cosine score of all candidate pairs
* List of dictionary containing all hashtables: This is named dictionary in the project and contains a hashmap from name of column vector (toCompress) to document number.

Output files

The output is finally printed to a file which is stored in the same folder. This output file contains various scores of document pairs which have been described in the design document.