**Documentation**

These are the functions used for the project:

1. **cal\_jaccard\_score(sigmat,d1,d2)** : This function takes signature matrix as input along with document number d1 and d2 and calculates jaccard score between the column vectors of documents d1 and d2 in signature matrix.
2. **cal\_cosine\_score(sig\_mat,d1,d2):** This function takes signature matrix as input along with document number d1 and d2 and calculates cosine score between the column vectors of documents d1 and d2 in signature matrix.
3. **hash\_signature2(sig\_mat, b, r):** This function takes signature matrix as input along with number of bands(b) and number of rows in each band(r) and calculates the hashtable for each band in the signature matrix and appends it in a list of dictionaries named “dictionary”.
4. **cal\_jacc\_score\_candidate\_pairs(threshold):** This function takes a threshold for the jaccard score as input and returns all the candidate pairs in the set of documents above the given threshold.
5. **cal\_cosine\_score\_candidate\_pairs(threshold):** This function takes a threshold for the cosine score as input and returns all the candidate pairs in the set of documents above the given threshold.
6. **build\_shingling\_matrix()**: From the dictionary from shingles to list of documents this function creates the shingling matrix used for the creation of signature matrix.
7. **build\_signature\_matrix()**: This function uses the shingling matrix created in the build\_shingling\_matrix() function to create the signature matrix using minhashing with 100 randomly generated hash functions.