MinHash class:

1. Given the folder path to MinHash class, each file within the folder is saved in an ArrayList by calling the method allDocs(). allDocs() calls the class FileNames in which the listFiles() method of class File is called to get all the filenames under the given directory.

For each file name in the ArrayList the file is read with a file reader and the content is filtered to remove the STOP words as given in the assignment pdf. Then content is split on “ “ character and each term is added to an ArrayList terms in such a way that no duplicate terms are added to the terms ArrayList.

1. After getting the final terms ArrayList each term in the ArrayList is given an integer starting from 0 to the size of the ArrayList and saved in another ArrayList called hterms.
2. Initially for each document a mapping is done to the integer present in the hterms ArrayList and the index positions are saved in a separate ArrayList called indexs for each document.

An ArrayList with integers from 0 to the number of Terms, m is shuffled using the java’s in built method Collections.shuffle(list), the list is shuffled from 1 to number of permutation, k times and each time it is shuffled for each document the corresponding intex in the corresponding indexs ArrayList is taken and the integer in that index in the shuffled list is extracted and the minimum of all the integers extracted is found. This minimum number is saved in the minHash matrix.

MinHashAccuracy class:

MinHashTime class:

LSH class:

1. The implementation challenge we faced was how to create and initialize b tables, and if b is large number then the memory being used was a concern. Even if there were b tables doing a parallel hashing and saving them in b tables to reduce time of computation was not possible. Somehow after experimentations the best we wrote was hashing each bands one by one. Hence we used only one hash table, in our implementation it is a hash map called bandTable. And for each band we hashed each documents min hash signature within that band and hash mapped with the hashvalue,HashSet docName pairs.

We then maintained another HashMap called lshsimilardocs, which will be the hashmap were each document has its similar docs mapped in docName,HashSet docName pairs.

Then traversed the bandTable and for each key in the hash map keyset and for each document in the document hashset linked to this key we added the entries to lshsimilardocs like (document, document hashset) pairs.

Then we moved to the next band segment in the minHashMatrix and repeat the computations.

1. Each r-tuple minhash signature of a document in a band is converted into a string like below:

For(int i= 0; i<rTupleLength;i++){

String s = minHashSig(i)+”.”;

This string is then hashed using FNV hash function and the hashvalue is got as below:

hashvalue = fnvHash(s);

hashvalue = hashvalue % bigPrime;

hashvalue = Math.abs(hashvalue);

here bigPrime is a prime number larger than the size of the documents, N.

1. Note that the near duplicates of method returns all the documents mapped into the same bucket of the b hash tables and do not consider the threshold by which they are similar, since the similarity threshold was not mentioned in the LSH class section in the assignment and was only addressed in the NearDuplicates class in the assignment. Hence there is an implementation in the NearDuplicates that returns s similar documents.

The pseudocode of nearDuplicateOf is as below:

ArrayList<String> nearDuplicatesOf(String docName)

1. similarDocs = new ArrayList<String>
2. index = get position of the docName from the documents list
3. from the lshsimilardocs hashmap which is an <Integer, HashSet<Integer>> mapping, we get the HashSet of integers saved in the index and save it in a temp HashSet.

HashSet<Integer> temp = lshsimilardocs.get(index);

1. get a list of docnames from the integers in the temp HashSet by getting the docnames saved in the corresponding position in the docnames list
2. return the list of docnames.

nearDuplicateDetector

We ran the nearDuplicateDetector on the files from F16PA2.zip on 10 different inputs with the threshold 0.9 and got the following results:

With filename “space-0.txt”

[space-0.txt.copy1, space-0.txt.copy2, space-0.txt.copy3, space-0.txt.copy4, space-0.txt.copy5, space-0.txt.copy6, space-0.txt.copy7]

with filename “baseball51.txt”

[baseball51.txt.copy1, baseball51.txt.copy2, baseball51.txt.copy3, baseball51.txt.copy4, baseball51.txt.copy5, baseball51.txt.copy6, baseball51.txt.copy7]

with filename “hockey857.txt”

[hockey857.txt.copy1, hockey857.txt.copy2, hockey857.txt.copy3, hockey857.txt.copy4, hockey857.txt.copy5, hockey857.txt.copy6, hockey857.txt.copy7]

with filename “space-205.txt”

[space-205.txt.copy1, space-205.txt.copy2, space-205.txt.copy3, space-205.txt.copy4, space-205.txt.copy5, space-205.txt.copy6, space-205.txt.copy7]

with filename “baseball777.txt”

[baseball777.txt.copy1, baseball777.txt.copy2, baseball777.txt.copy3, baseball777.txt.copy4, baseball777.txt.copy5, baseball777.txt.copy6, baseball777.txt.copy7]

with filename “hockey47.txt”

[hockey47.txt.copy1, hockey47.txt.copy2, hockey47.txt.copy3, hockey47.txt.copy4, hockey47.txt.copy5, hockey47.txt.copy6, hockey47.txt.copy7]

with filename “space-378.txt”

[space-378.txt.copy1, space-378.txt.copy2, space-378.txt.copy3, space-378.txt.copy4, space-378.txt.copy5, space-378.txt.copy6, space-378.txt.copy7]

with filename “hockey582.txt”

[hockey582.txt.copy1, hockey582.txt.copy2, hockey582.txt.copy3, hockey582.txt.copy4, hockey582.txt.copy5, hockey582.txt.copy6, hockey582.txt.copy7]

with filename “baseball292.txt”

[baseball292.txt.copy1, baseball292.txt.copy2, baseball292.txt.copy3, baseball292.txt.copy4, baseball292.txt.copy5, baseball292.txt.copy6, baseball292.txt.copy7]

with filename “hockey5.txt”

[hockey5.txt.copy1, hockey5.txt.copy2, hockey5.txt.copy3, hockey5.txt.copy4, hockey5.txt.copy5, hockey5.txt.copy6, hockey5.txt.copy7]

with filename “baseball910.txt”

[baseball910.txt.copy1, baseball910.txt.copy2, baseball910.txt.copy3, baseball910.txt.copy4, baseball910.txt.copy5, baseball910.txt.copy6, baseball910.txt.copy7]