**COMS 535 – Programming Assignment 2**

Hung Phan(799388622)

Shruti Sahu (106202120)

30 October 2017

1. ***MinHash:***

* *Term Collection:*

1. Collect all the file path in a file array by using listFiles() method from the directory where all the files are kept.
2. Iterate over all the files and remove all the stop words by only collecting the words that are not stop words in a list.

* *Data Structure*

1. To store the list of terms collected from each file: ArrayList<String>
2. To store all the words (vocabulary) in all the documents: HashSet<String>
3. To store terms and ids : Hashtable<String, Integer>
4. To store the occurrence of each term in the vocabulary each document in binary format: Hashtable<String, BitSet>

* *Assigning integers to terms*

1. Retrieve the HashSet<String> of vocabulary
2. Assign integer from 1 to size of the vocabulary and storing in a Hashtable<String, Integer>

* *Permutation used*

A hash function is used to generate the permutation. The terms a and b are generated randomly and c is the prime number that is fixed for all the hash functions.

1. ***MinHashAccuracy:***

|  |  |  |
| --- | --- | --- |
| **# of Permutation** | **Error Factor** | **Mismatch Count** |
| 400 | 0.04 | 2245 |
| 400 | 0.07 | 3 |
| 400 | 0.09 | 0 |
| 600 | 0.04 | 343 |
| 600 | 0.07 | 0 |
| 600 | 0.09 | 0 |
| 800 | 0.04 | 68 |
| 800 | 0.07 | 0 |
| 800 | 0.09 | 0 |

1. ***MinHashTime:***

Exact jaccard time: 70792

Construct minhash time: 1161

Approx jaccard time: 4695

Total Approx time: 5856

1. ***LSH:***

* *Implementation Challenges:*

1. Impossible to store the binary matrix for the collection of documents.
2. Selecting the number of bands is difficult

* *Hashing algorithm*

1. For each band *i* an array of hashtable values is created of size r.
2. For *j* from *i\*r* to *i+1* do
3. Index = *j%r*

*//* where r is the number of rows in each band

* *Psuedocode for nearDuplicateOf:*

1. Initialize an arraylist
2. Initialize a hashset
3. Using hashing algorithm compute hash values
4. Check if lsh\_hashtable has the hash value
5. Add them to the set
6. add the values of the set to arraylist and return

* *NearDuplicates:*

1) Result for baseball0.txt (7/7=1.0)

FileName approxJaccard exactJaccard

baseball0.txt.copy3 0.9583333333333334 0.9554140127388535

baseball0.txt.copy6 0.955 0.9554140127388535

baseball0.txt.copy2 0.955 0.9554140127388535

baseball0.txt.copy7 0.9533333333333334 0.9554140127388535

baseball0.txt.copy4 0.9516666666666667 0.9554140127388535

baseball0.txt.copy1 0.9416666666666667 0.9554140127388535

baseball0.txt.copy5 0.94 0.9554140127388535

baseball10.txt 24008

2) Result for baseball10.txt (7/7=1.0)

FileName approxJaccard exactJaccard

baseball10.txt.copy2 0.9333333333333333 0.9147442326980942

baseball10.txt.copy7 0.925 0.9147442326980942

baseball10.txt.copy3 0.9216666666666666 0.9147442326980942

baseball10.txt.copy5 0.9116666666666666 0.9147442326980942

baseball10.txt.copy6 0.9083333333333333 0.9147442326980942

baseball10.txt.copy4 0.9016666666666666 0.9147442326980942

baseball10.txt.copy1 0.9 0.9147442326980942

baseball20.txt 24008

3) Result for baseball20.txt (7/7=1.0)

FileName approxJaccard exactJaccard

baseball20.txt.copy4 0.985 0.9722222222222222

baseball20.txt.copy5 0.9633333333333334 0.9722222222222222

baseball20.txt.copy6 0.9566666666666667 0.9722222222222222

baseball20.txt.copy2 0.9483333333333334 0.9722222222222222

baseball20.txt.copy7 0.9466666666666667 0.9722222222222222

baseball20.txt.copy3 0.9466666666666667 0.9722222222222222

baseball20.txt.copy1 0.945 0.9722222222222222

baseball100.txt 24008

4) Result for baseball100.txt (7/7=1.0)

FileName approxJaccard exactJaccard

baseball100.txt.copy2 0.9883333333333333 0.9666666666666667

baseball100.txt.copy1 0.9833333333333333 0.9666666666666667

baseball100.txt.copy6 0.9833333333333333 0.9666666666666667

baseball100.txt.copy3 0.975 0.9666666666666667

baseball100.txt.copy4 0.9683333333333334 0.9666666666666667

baseball100.txt.copy5 0.9683333333333334 0.9666666666666667

baseball100.txt.copy7 0.965 0.9666666666666667

baseball200.txt 24008

5) Result for baseball200.txt (7/7=1.0)

FileName approxJaccard exactJaccard

baseball200.txt.copy4 0.9633333333333334 0.9428571428571428

baseball200.txt.copy6 0.9566666666666667 0.9428571428571428

baseball200.txt.copy5 0.9566666666666667 0.9428571428571428

baseball200.txt.copy7 0.9483333333333334 0.9428571428571428

baseball200.txt.copy3 0.9466666666666667 0.9428571428571428

baseball200.txt.copy1 0.9183333333333333 0.9428571428571428

baseball200.txt.copy2 0.9116666666666666 0.9428571428571428

baseball300.txt 24008

6) Result for baseball300.txt (7/7=1.0)

FileName approxJaccard exactJaccard

baseball300.txt.copy2 1.0 1.0

baseball300.txt.copy3 1.0 1.0

baseball300.txt.copy4 1.0 1.0

baseball300.txt.copy5 1.0 1.0

baseball300.txt.copy1 1.0 1.0

baseball300.txt.copy6 1.0 1.0

baseball300.txt.copy7 1.0 1.0

space-0.txt 24008

7) Result for space-0.txt (7/7=1.0)

FileName approxJaccard exactJaccard

space-0.txt.copy4 0.97 0.9576271186440678

space-0.txt.copy1 0.9633333333333334 0.9576271186440678

space-0.txt.copy6 0.9533333333333334 0.9576271186440678

space-0.txt.copy2 0.9516666666666667 0.9576271186440678

space-0.txt.copy3 0.95 0.9576271186440678

space-0.txt.copy7 0.9483333333333334 0.9576271186440678

space-0.txt.copy5 0.9416666666666667 0.9576271186440678

space-10.txt 24008

8) Result for space-10.txt (7/7=1.0)

FileName approxJaccard exactJaccard

space-10.txt.copy3 0.9516666666666667 0.9357664233576642

space-10.txt.copy6 0.9483333333333334 0.9357664233576642

space-10.txt.copy5 0.9433333333333334 0.9357664233576642

space-10.txt.copy2 0.9366666666666666 0.9357664233576642

space-10.txt.copy1 0.9283333333333333 0.9357664233576642

space-10.txt.copy4 0.925 0.9357664233576642

space-10.txt.copy7 0.9133333333333333 0.9357664233576642

space-20.txt 24008

9) Result for space-20.txt (7/7=1.0)

FileName approxJaccard exactJaccard

space-20.txt.copy4 0.9766666666666667 0.9578947368421052

space-20.txt.copy1 0.97 0.9578947368421052

space-20.txt.copy7 0.9633333333333334 0.9578947368421052

space-20.txt.copy2 0.9566666666666667 0.9578947368421052

space-20.txt.copy5 0.9516666666666667 0.9578947368421052

space-20.txt.copy6 0.95 0.9578947368421052

space-20.txt.copy3 0.9416666666666667 0.9578947368421052

space-100.txt 24008

10) Result for space-100.txt (7/7=1.0)

FileName approxJaccard exactJaccard

space-100.txt.copy5 0.9616666666666667 0.9465648854961832

space-100.txt.copy6 0.9616666666666667 0.9465648854961832

space-100.txt.copy3 0.9583333333333334 0.9465648854961832

space-100.txt.copy7 0.9566666666666667 0.9465648854961832

space-100.txt.copy2 0.9566666666666667 0.9465648854961832

space-100.txt.copy4 0.9466666666666667 0.9465648854961832

space-100.txt.copy1 0.9366666666666666 0.9465648854961832

space-200.txt 24008

11) Result for space-200.txt (7/7=1.0)

FileName approxJaccard exactJaccard

space-200.txt.copy5 0.9816666666666667 0.9743589743589743

space-200.txt.copy7 0.9816666666666667 0.9743589743589743

space-200.txt.copy6 0.9783333333333334 0.9743589743589743

space-200.txt.copy2 0.975 0.9743589743589743

space-200.txt.copy4 0.97 0.9743589743589743

space-200.txt.copy3 0.9683333333333334 0.9743589743589743

space-200.txt.copy1 0.9533333333333334 0.9743589743589743

space-300.txt 24008

12) Result for space-300.txt (7/7=1.0)

FileName approxJaccard exactJaccard

space-300.txt.copy7 0.965 0.9518072289156626

space-300.txt.copy6 0.9633333333333334 0.9518072289156626

space-300.txt.copy5 0.96 0.9518072289156626

space-300.txt.copy1 0.9566666666666667 0.9518072289156626

space-300.txt.copy4 0.9533333333333334 0.9518072289156626

space-300.txt.copy3 0.9516666666666667 0.9518072289156626

space-300.txt.copy2 0.94 0.9518072289156626

hockey0.txt 24008

13) Result for hockey0.txt (7/7=1.0)

FileName approxJaccard exactJaccard

hockey0.txt.copy4 0.9283333333333333 0.9228971962616822

hockey0.txt.copy2 0.9266666666666666 0.9228971962616822

hockey0.txt.copy5 0.9233333333333333 0.9228971962616822

hockey0.txt.copy1 0.9216666666666666 0.9228971962616822

hockey0.txt.copy6 0.9166666666666666 0.9228971962616822

hockey0.txt.copy3 0.915 0.9228971962616822

hockey0.txt.copy7 0.905 0.9228971962616822

hockey10.txt 24008

14) Result for hockey10.txt (7/7=1.0)

FileName approxJaccard exactJaccard

hockey10.txt.copy3 0.9316666666666666 0.9205607476635514

hockey10.txt.copy4 0.9133333333333333 0.9205607476635514

hockey10.txt.copy7 0.9116666666666666 0.9205607476635514

hockey10.txt.copy2 0.9116666666666666 0.9205607476635514

hockey10.txt.copy5 0.9116666666666666 0.9205607476635514

hockey10.txt.copy1 0.9083333333333333 0.9205607476635514

hockey10.txt.copy6 0.8916666666666667 0.9205607476635514

hockey20.txt 24008

15) Result for hockey20.txt (7/7=1.0)

FileName approxJaccard exactJaccard

hockey20.txt.copy3 0.97 0.953125

hockey20.txt.copy1 0.96 0.953125

hockey20.txt.copy2 0.96 0.953125

hockey20.txt.copy5 0.9583333333333334 0.953125

hockey20.txt.copy6 0.9566666666666667 0.953125

hockey20.txt.copy4 0.9533333333333334 0.953125

hockey20.txt.copy7 0.9466666666666667 0.953125

hockey100.txt 24008

16) Result for hockey100.txt (7/7=1.0)

FileName approxJaccard exactJaccard

hockey100.txt.copy2 0.9866666666666667 0.9629629629629629

hockey100.txt.copy1 0.9783333333333334 0.9629629629629629

hockey100.txt.copy3 0.9666666666666667 0.9629629629629629

hockey100.txt.copy4 0.955 0.9629629629629629

hockey100.txt.copy5 0.9533333333333334 0.9629629629629629

hockey100.txt.copy6 0.9516666666666667 0.9629629629629629

hockey100.txt.copy7 0.9516666666666667 0.9629629629629629

hockey200.txt 24008

17) Result for hockey200.txt (7/7=1.0)

FileName approxJaccard exactJaccard

hockey200.txt.copy7 0.945 0.9399141630901288

hockey200.txt.copy3 0.945 0.9399141630901288

hockey200.txt.copy6 0.94 0.9399141630901288

hockey200.txt.copy2 0.94 0.9399141630901288

hockey200.txt.copy1 0.9333333333333333 0.9399141630901288

hockey200.txt.copy4 0.925 0.9399141630901288

hockey200.txt.copy5 0.92 0.9399141630901288

hockey300.txt 24008

18) Result for hockey300.txt (7/7=1.0)

FileName approxJaccard exactJaccard

hockey300.txt.copy5 0.98 0.9607843137254902

hockey300.txt.copy1 0.965 0.9607843137254902

hockey300.txt.copy3 0.9583333333333334 0.9607843137254902

hockey300.txt.copy6 0.9583333333333334 0.9607843137254902

hockey300.txt.copy4 0.9566666666666667 0.9607843137254902

hockey300.txt.copy7 0.9566666666666667 0.9607843137254902

hockey300.txt.copy2 0.9466666666666667 0.9607843137254902