**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.

1. ***MinHashAccuracy:***

|  |  |  |
| --- | --- | --- |
| **# of Permutation** | **Error Factor** | **Mismatch Count** |
| 400 | 0.04 | 296654 |
| 400 | 0.07 | 157891 |
| 400 | 0.09 | 92792 |
| 600 | 0.04 | 139382 |
| 600 | 0.07 | 41225 |
| 600 | 0.09 | 18581 |
| 800 | 0.04 | 90889 |
| 800 | 0.07 | 17986 |
| 800 | 0.09 | 24492 |

1. ***MinHashTime:***
2. ***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:*
* *NearDuplicates:*

1) Result for baseball0.txt (7/9798=7.144315166360481E-4)

FileName approxJaccard exactJaccard

baseball0.txt.copy7 0.97375 0.9777396600155359

baseball0.txt.copy6 0.9725 0.9777396600155359

baseball0.txt.copy3 0.97 0.9777396600155359

baseball0.txt.copy2 0.9625 0.9777396600155359

baseball0.txt.copy1 0.945 0.9777396600155359

baseball0.txt.copy5 0.93375 0.9777396600155359

baseball0.txt.copy4 0.93 0.9777396600155359

2) Result for baseball10.txt (5/9054=5.52242102937928E-4)

FileName approxJaccard exactJaccard

baseball10.txt.copy3 0.92375 0.9625334218796219

baseball10.txt.copy2 0.92125 0.9625334218796219

baseball10.txt.copy5 0.91625 0.9625334218796219

baseball10.txt.copy6 0.9 0.9625334218796219

baseball10.txt.copy1 0.8825 0.9625334218796219

3) Result for baseball20.txt (2/9054=2.208968411751712E-4)

FileName approxJaccard exactJaccard

baseball20.txt.copy6 0.94 0.9870962335856491

baseball20.txt.copy2 0.9325 0.9870962335856491

4) Result for baseball100.txt (4/8768=4.562043795620438E-4)

FileName approxJaccard exactJaccard

baseball100.txt.copy1 0.97375 0.982607368881035

baseball93.txt.copy5 0.67375 0.7559289460184545

baseball93.txt.copy2 0.655 0.7559289460184545

baseball93.txt.copy3 0.65125 0.7559289460184545

5) Result for baseball200.txt (4/8768=4.562043795620438E-4)

FileName approxJaccard exactJaccard

baseball200.txt.copy4 0.9675 0.9724084693648638

baseball200.txt.copy6 0.96375 0.9724084693648638

baseball200.txt.copy1 0.935 0.9724084693648638

baseball200.txt.copy2 0.93375 0.9724084693648638

6) Result for baseball300.txt (0/8768=0.0)

FileName approxJaccard exactJaccard

7) Result for space-0.txt (5/10969=4.558300665511897E-4)

FileName approxJaccard exactJaccard

space-0.txt.copy2 0.955 0.9756654533819864

space-0.txt.copy7 0.95125 0.9756654533819864

space-0.txt.copy6 0.94875 0.9756654533819864

space-0.txt.copy5 0.9475 0.9756654533819864

space-531.txt.copy7 0.6475 0.8233056133593915

8) Result for space-10.txt (3/10377=2.891008962127783E-4)

FileName approxJaccard exactJaccard

space-10.txt.copy1 0.96 0.9686084075574665

space-10.txt.copy4 0.9275 0.9686084075574665

space-10.txt.copy7 0.9275 0.9686084075574665

9) Result for space-20.txt (1/10377=9.636696540425942E-5)

FileName approxJaccard exactJaccard

space-20.txt.copy4 0.96125 0.9766049484994053

10) Result for space-100.txt (5/9798=5.103082261686059E-4)

FileName approxJaccard exactJaccard

space-100.txt.copy2 0.9525 0.97182531580755

space-100.txt.copy5 0.94125 0.97182531580755

space-100.txt.copy4 0.93125 0.97182531580755

space-100.txt.copy6 0.91875 0.97182531580755

space-100.txt.copy1 0.91 0.97182531580755

11) Result for space-200.txt (1/9798=1.0206164523372117E-4)

FileName approxJaccard exactJaccard

space-200.txt.copy6 0.92625 0.9839989676081822

12) Result for space-300.txt (3/9798=3.061849357011635E-4)

FileName approxJaccard exactJaccard

space-300.txt.copy2 0.95875 0.9734465847056577

space-300.txt.copy4 0.955 0.9734465847056577

space-300.txt.copy1 0.9375 0.9734465847056577

13) Result for hockey0.txt (3/10969=2.7349803993071385E-4)

FileName approxJaccard exactJaccard

hockey0.txt.copy1 0.93875 0.9653072991634227

hockey0.txt.copy6 0.9175 0.9653072991634227

hockey0.txt.copy5 0.9025 0.9653072991634227

14) Result for hockey10.txt (4/10377=3.8546786161703766E-4)

FileName approxJaccard exactJaccard

hockey10.txt.copy7 0.94375 0.9619120153500137

hockey10.txt.copy4 0.935 0.9619120153500137

hockey10.txt.copy2 0.915 0.9619120153500137

hockey10.txt.copy6 0.8925 0.9619120153500137

15) Result for hockey20.txt (6/10377=5.782017924255566E-4)

FileName approxJaccard exactJaccard

hockey20.txt.copy2 0.975 0.9762812094883317

hockey20.txt.copy1 0.9675 0.9762812094883317

hockey20.txt.copy6 0.96625 0.9762812094883317

hockey20.txt.copy5 0.94625 0.9762812094883317

hockey20.txt.copy7 0.935 0.9762812094883317

hockey20.txt.copy3 0.92625 0.9762812094883317

16) Result for hockey100.txt (2/9798=2.0412329046744235E-4)

FileName approxJaccard exactJaccard

hockey100.txt.copy5 0.9675 0.9817613873476319

hockey100.txt.copy2 0.95875 0.9817613873476319

17) Result for hockey200.txt (3/9798=3.061849357011635E-4)

FileName approxJaccard exactJaccard

hockey200.txt.copy3 0.95375 0.9713641887040998

hockey200.txt.copy4 0.94 0.9713641887040998

hockey200.txt.copy2 0.90875 0.9713641887040998

18) Result for hockey300.txt (7/9798=7.144315166360481E-4)

FileName approxJaccard exactJaccard

hockey300.txt.copy2 0.9875 0.9801960588196068

hockey300.txt.copy1 0.98375 0.9801960588196068

hockey300.txt.copy3 0.96625 0.9801960588196068

hockey300.txt.copy5 0.96375 0.9801960588196068

hockey300.txt.copy4 0.95 0.9801960588196068

hockey300.txt.copy7 0.9375 0.9801960588196068

hockey300.txt.copy6 0.9175 0.9801960588196068