Com S 435 PA2-Report

Jialin Xing

MinHash-This class mainly constructs MinHashMatrix and termMatrix for N documents. To collect all terms of the documents, a hashtable is created and stores each term as key and value is like {1,2,3,4…M} represented as each term after preprocessing the words. This hashtable only collects each term once. For permutation, a 2D integer list is used to store the random permutations. To generate random permutation, an array list is used to store each row of random number by calling collections.shuffle and use one for loop to store in each row of 2D integer.

MinHashAccuracy- The number of pairs for which approximate and exact similarities differ by more than ϵ $(0.04, 0.07, 0.09)$: 400 permutations – 46187, 10743, 2338 ; 600 permutations – 71901, 7154, 3016 ; 800 permutations – 33136, 4783, 3708. Generally speaking, it is clear to see that number of pairs decrease when the ϵ is larger. Although it is not always the case, it shows that two similarities value are more closer when the more permutations are applied.

MinHashTime- The total run time to compute exact Jaccard similarities and approximate Jaccard similarities when running 600 permutations on space.zip is 438064ms = 7.3 minutes.(Constructor time: 295682ms;Exact: 436615ms;Approximate: 1449ms)

NearDuplicateDetector- nearDuplicateDetector("baseball0.txt")
1st input: numPerm:400 s:0.9 return: baseball0.txt.copy5 baseball0.txt.copy6 baseball0.txt.copy2 baseball0.txt.copy3 baseball0.txt.copy7 baseball0.txt.copy4 baseball0.txt.copy1

2nd input: numPerm:600 s:0.9 return: baseball0.txt.copy7 baseball0.txt.copy1 baseball0.txt.copy5 baseball0.txt.copy4 baseball0.txt.copy3 baseball0.txt.copy2 baseball0.txt.copy6

3rd input: numPerm:800 s:0.9 return: baseball0.txt.copy2 baseball0.txt.copy7 baseball0.txt.copy1 baseball0.txt.copy5 baseball0.txt.copy6 baseball0.txt.copy4

4th input: numPerm:600 s:0.87 return: baseball0.txt.copy5 baseball0.txt.copy7baseball0.txt.copy6 baseball0.txt.copy4 baseball0.txt.copy2 baseball0.txt.copy1 baseball0.txt.copy3

5th input: numPerm:600 s:0.91 return: baseball0.txt.copy1 baseball0.txt.copy3baseball0.txt.copy4 baseball0.txt.copy6 baseball0.txt.copy7 baseball0.txt.copy2 baseball0.txt.copy5

6th input: numPerm:600 s:0.92 return: baseball0.txt.copy6 baseball0.txt.copy4baseball0.txt.copy3 baseball0.txt.copy7 baseball0.txt.copy2 baseball0.txt.copy5 baseball0.txt.copy1

7th input: numPerm:400 s:0.93 return: baseball0.txt.copy3 baseball0.txt.copy6baseball0.txt.copy2 baseball0.txt.copy4 baseball0.txt.copy7 baseball0.txt.copy1 baseball0.txt.copy5

8th input: numPerm:400 s:0.95 return: baseball0.txt.copy1 baseball0.txt.copy4baseball0.txt.copy5 baseball0.txt.copy6 baseball0.txt.copy2 baseball0.txt.copy3

9th input: numPerm:400 s:0.97 return: baseball0.txt.copy6

10th input: numPerm:600 s:0.98 return: null