Michael Koopmann

COM S 435

PA2 Report

* MinHash:
  + I used an ArrayList of Strings in the Document class that would take the strings that matched the criteria for being a term and then would check if that term has already been stored if so then continue if not then it adds it to the ArrayList.
  + I go through all the terms stored in the document. First I use hashCode() to get a base value then I use the permutation formula (ax+b)%p to get a permutation value then I compare that to the current minimum permutation value and I store the smallest value.
  + In the HelperFunctios class I have a getParameters function that takes the number of permutations and terms as inputs then I generate pairs of random values a and b using random generator with numTerms as the limit. I do this for each number of Permutaions then I return an ArrayList of these Parameters.
* MinHashAccuracy:
  + Results

|  |  |  |  |
| --- | --- | --- | --- |
|  | 400 | 600 | 800 |
| .04 | 20768 | 13474 | 10440 |
| .07 | 456 | 195 | 65 |
| .09 | 84 | 75 | 64 |

* + Clearly you can tell that as you increase the allowed error parameter the number of pairs that are reported as being larger than said parameter decreases because we’ve moved our range. What I find most interesting is the change from the 400 permutations to the 800 but specifically the difference from .04 to .09 because from 400 to 800 for .04 about cut the amount reported in half while for the .09 it only decreased by 20 which is still a fourth but it’s nowhere near as dramatic.
* MinHashTime:
  + Using 600 permutations on the files from space.zip report the run time
    - Construct MinHashSimilarities is 1837 milliseconds
    - Compute every exact Jaccard is 8140 milliseconds
    - Estimate every Jaccard is 19 milliseconds.
* SimilaritySearch:

|  |  |  |  |
| --- | --- | --- | --- |
| simThreshold | numPermutations | File name | Results |
| .9 | 600 | baseball0.txt | baseball0.txt  baseball0.txt.copy1  baseball0.txt.copy2  baseball0.txt.copy4  baseball0.txt.copy5  baseball0.txt.copy6  baseball0.txt.copy7 |
| .2 | 600 | baseball0.txt | baseball0.txt  baseball0.txt.copy1  baseball0.txt.copy2  baseball0.txt.copy4  baseball0.txt.copy5  baseball0.txt.copy6  baseball0.txt.copy7  baseball31.txt  baseball31.txt.copy1  baseball31.txt.copy2  baseball31.txt.copy3  baseball31.txt.copy4  baseball31.txt.copy5  baseball31.txt.copy6 |
| .2 | 200 | baseball0.txt | baseball0.txt  baseball0.txt.copy1  baseball0.txt.copy2  baseball0.txt.copy4  baseball0.txt.copy5  baseball0.txt.copy6  baseball0.txt.copy7 |
| .2 | 600 | baseball1.txt | baseball1.txt  baseball1.txt.copy1  baseball1.txt.copy2  baseball1.txt.copy3  baseball1.txt.copy4  baseball1.txt.copy5  baseball1.txt.copy6 baseball1.txt.copy7 |
| .5 | 600 | baseball1.txt | baseball1.txt  baseball1.txt.copy1  baseball1.txt.copy2  baseball1.txt.copy3  baseball1.txt.copy4  baseball1.txt.copy5  baseball1.txt.copy6  baseball1.txt.copy7 |
| .7 | 200 | baseball2.txt | baseball2.txt  baseball2.txt.copy1  baseball2.txt.copy2  baseball2.txt.copy3  baseball2.txt.copy4  baseball2.txt.copy5  baseball2.txt.copy6  baseball2.txt.copy7 |
| .15 | 400 | baseball2.txt | baseball2.txt  baseball2.txt.copy1  baseball2.txt.copy2  baseball2.txt.copy3  baseball2.txt.copy4  baseball2.txt.copy5  baseball2.txt.copy6  baseball2.txt.copy7  baseball0.txt  baseball0.txt.copy1  baseball0.txt.copy2  baseball0.txt.copy3  baseball0.txt.copy4  baseball0.txt.copy5  baseball0.txt.copy6  baseball0.txt.copy7  baseball31.txt  baseball31.txt.copy1  baseball31.txt.copy2  baseball31.txt.copy3  baseball31.txt.copy4  baseball31.txt.copy5  baseball31.txt.copy6  baseball31.txt.copy7  baseball71.txt  baseball71.txt.copy1  baseball71.txt.copy2  baseball71.txt.copy3  baseball71.txt.copy4  baseball71.txt.copy5  baseball71.txt.copy6  baseball71.txt.copy7 |
| .1 | 400 | baseball2.txt | baseball2.txt.copy2  baseball2.txt.copy3  baseball2.txt.copy5  baseball2.txt.copy6  baseball2.txt  baseball2.txt.copy4  baseball2.txt.copy7  baseball47.txt  baseball47.txt.copy1  baseball47.txt.copy2  baseball47.txt.copy3  baseball47.txt.copy4  baseball47.txt.copy5  baseball47.txt.copy6  baseball47.txt.copy7  baseball70.txt  baseball70.txt.copy1  baseball70.txt.copy2  baseball70.txt.copy3  baseball70.txt.copy4  baseball70.txt.copy6  baseball31.txt  baseball31.txt.copy1  baseball31.txt.copy2  baseball31.txt.copy3  baseball31.txt.copy4  baseball31.txt.copy5  baseball31.txt.copy6  baseball31.txt.copy7  baseball38.txt  baseball38.txt.copy3  baseball38.txt.copy4  baseball38.txt.copy5  baseball38.txt.copy6  baseball38.txt.copy7  baseball4.txt  baseball4.txt.copy1  baseball4.txt.copy2  baseball4.txt.copy3  baseball4.txt.copy5  baseball4.txt.copy6  baseball4.txt.copy7  baseball43.txt  baseball43.txt.copy1  baseball43.txt.copy2  baseball43.txt.copy3  baseball43.txt.copy4  baseball43.txt.copy5  baseball43.txt.copy6  baseball43.txt.copy7  baseball73.txt  baseball73.txt.copy1  baseball73.txt.copy2  baseball73.txt.copy3  baseball73.txt.copy4  baseball73.txt.copy5  baseball73.txt.copy6  baseball73.txt.copy7  baseball78.txt  baseball78.txt.copy1  baseball78.txt.copy2  baseball78.txt.copy3  baseball78.txt.copy4  baseball78.txt.copy5  baseball78.txt.copy6  baseball78.txt.copy7  baseball80.txt  baseball80.txt.copy1  baseball80.txt.copy2  baseball80.txt.copy5  baseball80.txt.copy6  baseball80.txt.copy7  baseball98.txt  baseball98.txt.copy1  baseball98.txt.copy2  baseball98.txt.copy3  baseball98.txt.copy4  baseball98.txt.copy5  baseball98.txt.copy6  baseball98.txt.copy7  baseball99.txt  baseball99.txt.copy1  baseball99.txt.copy2  baseball99.txt.copy5  baseball99.txt.copy6  baseball99.txt.copy7 |
| .05 | 400 | baseball2.txt | baseball2.txt.copy1  baseball2.txt.copy2  baseball2.txt.copy3  baseball2.txt.copy4  baseball2.txt.copy6  baseball2.txt  baseball2.txt.copy5  baseball2.txt.copy7  baseball33.txt  baseball33.txt.copy1  baseball33.txt.copy2  baseball33.txt.copy3  baseball33.txt.copy4  baseball33.txt.copy5  baseball33.txt.copy6  baseball33.txt.copy7  baseball1.txt  baseball1.txt.copy1  baseball1.txt.copy2  baseball1.txt.copy3  baseball1.txt.copy5  baseball1.txt.copy7  baseball25.txt  baseball25.txt.copy1  baseball25.txt.copy3  baseball25.txt.copy5  baseball25.txt.copy6  baseball25.txt.copy7  baseball85.txt  baseball85.txt.copy1  baseball85.txt.copy3  baseball85.txt.copy4  baseball85.txt.copy5  baseball85.txt.copy6  baseball85.txt.copy7  baseball46.txt  baseball46.txt.copy1  baseball46.txt.copy2  baseball46.txt.copy3  baseball46.txt.copy4  baseball46.txt.copy5  baseball46.txt.copy6  baseball46.txt.copy7  baseball85.txt.copy2  baseball87.txt  baseball87.txt.copy1  baseball87.txt.copy2  baseball87.txt.copy3  baseball87.txt.copy4  baseball87.txt.copy5  baseball87.txt.copy6  baseball87.txt.copy7  baseball90.txt  baseball90.txt.copy1  baseball90.txt.copy2  baseball90.txt.copy3  baseball90.txt.copy4  baseball90.txt.copy5  baseball90.txt.copy6  baseball90.txt.copy7  baseball0.txt  baseball0.txt.copy1  baseball0.txt.copy2  baseball0.txt.copy4  baseball0.txt.copy5  baseball0.txt.copy6  baseball0.txt.copy7  baseball31.txt  baseball31.txt.copy1  baseball31.txt.copy2  baseball31.txt.copy3  baseball31.txt.copy4  baseball31.txt.copy5  baseball31.txt.copy6  baseball31.txt.copy7  baseball56.txt  baseball56.txt.copy1  baseball56.txt.copy2  baseball56.txt.copy3  baseball56.txt.copy4  baseball56.txt.copy5  baseball56.txt.copy6  baseball56.txt.copy7  baseball0.txt.copy3  baseball28.txt  baseball28.txt.copy1  baseball28.txt.copy2  baseball28.txt.copy3  baseball28.txt.copy4  baseball28.txt.copy5  baseball28.txt.copy6  baseball28.txt.copy7  baseball32.txt  baseball32.txt.copy1  baseball32.txt.copy2  baseball32.txt.copy3  baseball32.txt.copy4  baseball32.txt.copy5  baseball32.txt.copy6  baseball32.txt.copy7  baseball7.txt  baseball7.txt.copy1  baseball7.txt.copy2  baseball7.txt.copy3  baseball7.txt.copy4  baseball7.txt.copy5  baseball7.txt.copy6  baseball7.txt.copy7  baseball4.txt  baseball4.txt.copy1  baseball4.txt.copy2  baseball4.txt.copy3  baseball4.txt.copy4  baseball4.txt.copy5  baseball4.txt.copy6  baseball4.txt.copy7  baseball42.txt  baseball42.txt.copy1  baseball42.txt.copy2  baseball42.txt.copy3  baseball42.txt.copy4  baseball42.txt.copy5  baseball42.txt.copy6  baseball42.txt.copy7  baseball66.txt  baseball66.txt.copy1  baseball66.txt.copy2  baseball66.txt.copy3  baseball66.txt.copy4  baseball66.txt.copy5  baseball66.txt.copy6  baseball66.txt.copy7  baseball100.txt  baseball100.txt.copy1  baseball100.txt.copy2  baseball100.txt.copy3  baseball100.txt.copy4  baseball100.txt.copy5  baseball100.txt.copy6  baseball100.txt.copy7  baseball53.txt  baseball53.txt.copy1  baseball53.txt.copy2  baseball53.txt.copy3  baseball53.txt.copy4  baseball53.txt.copy5  baseball53.txt.copy6  baseball53.txt.copy7  baseball74.txt  baseball74.txt.copy1  baseball74.txt.copy2  baseball74.txt.copy3  baseball74.txt.copy5  baseball74.txt.copy6  baseball74.txt.copy7  baseball1.txt.copy4  baseball1.txt.copy6  baseball74.txt.copy4 |
| .02 | 400 | baseball2.txt | baseball2.txt.copy1  baseball2.txt.copy2  baseball2.txt.copy3  baseball2.txt.copy4  baseball2.txt.copy5  baseball2.txt.copy6  baseball2.txt.copy7  baseball2.txt  baseball0.txt  baseball0.txt.copy1  baseball0.txt.copy2  baseball0.txt.copy3  baseball0.txt.copy4  baseball0.txt.copy5  baseball0.txt.copy6  baseball0.txt.copy7  baseball1.txt  baseball1.txt.copy1  baseball1.txt.copy2  baseball1.txt.copy3  baseball1.txt.copy4  baseball1.txt.copy5  baseball1.txt.copy6  baseball1.txt.copy7  baseball10.txt  baseball10.txt.copy1  baseball10.txt.copy2  baseball10.txt.copy3  baseball10.txt.copy4  baseball10.txt.copy5  baseball10.txt.copy6  baseball18.txt  baseball18.txt.copy2  baseball18.txt.copy3  baseball18.txt.copy4  baseball18.txt.copy5  baseball18.txt.copy6  baseball18.txt.copy7  baseball22.txt  baseball22.txt.copy1  baseball22.txt.copy2  baseball22.txt.copy3  baseball22.txt.copy4  baseball22.txt.copy5  baseball22.txt.copy6  baseball22.txt.copy7  baseball23.txt  baseball23.txt.copy1  baseball23.txt.copy2  baseball23.txt.copy3  baseball23.txt.copy4  baseball23.txt.copy5  baseball23.txt.copy6  baseball23.txt.copy7  baseball40.txt  baseball40.txt.copy1  baseball40.txt.copy2  baseball40.txt.copy3  baseball40.txt.copy4  baseball40.txt.copy5  baseball40.txt.copy6  baseball40.txt.copy7  baseball42.txt  baseball42.txt.copy1  baseball42.txt.copy2  baseball42.txt.copy3  baseball42.txt.copy4  baseball42.txt.copy5  baseball42.txt.copy6  baseball42.txt.copy7  baseball85.txt  baseball85.txt.copy1  baseball85.txt.copy2  baseball85.txt.copy3  baseball85.txt.copy4  baseball85.txt.copy5  baseball85.txt.copy6  baseball85.txt.copy7  baseball87.txt  baseball87.txt.copy1  baseball87.txt.copy2  baseball87.txt.copy3  baseball87.txt.copy4  baseball87.txt.copy5  baseball87.txt.copy6  baseball87.txt.copy7  baseball64.txt  baseball64.txt.copy1  baseball64.txt.copy2  baseball64.txt.copy3  baseball64.txt.copy5  baseball64.txt.copy6  baseball64.txt.copy7  baseball31.txt  baseball31.txt.copy1  baseball31.txt.copy2  baseball31.txt.copy3  baseball31.txt.copy4  baseball31.txt.copy5  baseball31.txt.copy6  baseball31.txt.copy7  baseball74.txt  baseball74.txt.copy1  baseball74.txt.copy2  baseball74.txt.copy3  baseball74.txt.copy4  baseball74.txt.copy5  baseball74.txt.copy6  baseball74.txt.copy7  baseball38.txt  baseball38.txt.copy1  baseball38.txt.copy2  baseball38.txt.copy3  baseball38.txt.copy4  baseball38.txt.copy5  baseball38.txt.copy6  baseball38.txt.copy7  baseball34.txt  baseball34.txt.copy1  baseball34.txt.copy2  baseball34.txt.copy3  baseball34.txt.copy4  baseball34.txt.copy5  baseball34.txt.copy6  baseball34.txt.copy7  baseball26.txt  baseball26.txt.copy1  baseball26.txt.copy2  baseball26.txt.copy3  baseball26.txt.copy4  baseball26.txt.copy5  baseball26.txt.copy6  baseball26.txt.copy7  baseball54.txt  baseball54.txt.copy1  baseball54.txt.copy2  baseball54.txt.copy3  baseball54.txt.copy4  baseball54.txt.copy5  baseball54.txt.copy6  baseball54.txt.copy7  baseball90.txt  baseball90.txt.copy1  baseball90.txt.copy2  baseball90.txt.copy3  baseball90.txt.copy4  baseball90.txt.copy5  baseball90.txt.copy6  baseball90.txt.copy7  baseball51.txt  baseball51.txt.copy1  baseball51.txt.copy2  baseball51.txt.copy3  baseball51.txt.copy4  baseball51.txt.copy5  baseball51.txt.copy7  baseball71.txt  baseball71.txt.copy1  baseball71.txt.copy2  baseball71.txt.copy3  baseball71.txt.copy4  baseball71.txt.copy5  baseball71.txt.copy6 |