IOWA STATE UNIVERSITY, COMPUTER SCIENCE DEPARTMENT

Computer Science 535
Project 2

Jason Hu, Joo Seung Song

November 3, 2019

1 MIN HASH

1.1 TERM COLLECTION AND DATA STRUCTURE

We iterated over all data files, preprocessed them, and collected all the unique terms and stored them in a HashMap `uniqueWordIndex` from every term to its index. The index of the term is the same index for the `termDocumentMatrix`, which is a 2-d integer array. The computation of `termDocumentMatrix` requires another iteration over all files, because we would not know the number of unique terms before computing it. We compute termDocumentMatrix for each document by iterating over all words and increment its corresponding `uniqueWordIndex`. We then compute a 1-d integer array `multiSetUnion`, which is the maximum `termDocumentMatrix` over all documents for each unique word. We also compute a 1-d integer array `multiWordStartIndex`, which maps each unique word to its corresponding multi-set word partition start index. This is needed to compute our goal Multiset Jaccard.

When the data set is too big and `termDocumentMatrix` cannot be stored in memory, term frequency will be computed when on the fly.

1.2 INTEGER ASSIGNMENT TO EACH TERM

`uniqueWordIndex` is a hash map that assigns an integer to every unique word, from zero to the size of vocabulary minus one. To assign an integer to a multi word, which is a tuple (w, n) where w is the unique word for the n-th appearance in the document, we lookup $u = \texttt{uniqueWordIndex}[w]$ and $v = \texttt{multiWordStartIndex}[u] + n - 1$, where v is a unique index for the multi-word tuple (w, n) distinguishing repetition.

1.3 Permutation used

Every permutation f maps sets $\{1, 2, ...$`numTerms()`$\} \rightarrow \{1, 2, ...$`numTerms()`$\}$. The `numTerms()` is computed and stored when computing `multiWordStartIndex`, so we know it when constructing the permutation. The permutation is generated by Fisher Yates Shuffle.

```
Algorithm 1: Fisher Yates(a)
  Input: An array $a$.
  Output: The randomly permuted array.
  for $i = 0 ... len(a) - 2$ do
      $j$= RandomInt($i, n$)
      exchange a[i] and a[j]
  return $a$
```

2 Min Hash Accuracy

The number of pairs exceeding the error margin is reported as follows. Specification: $(a, b) = (b, a)$, and (a, a) is allowed for all documents a, b.

permutations \ ϵ	0.04	0.07	0.09
400	1638	5	0
600	109	0	0
800	16	0	0

The number of pairs exceeding error margin decreases as the number of permutations used increases and as the error margin increases, all other variables the same.

3 Min Hash Time

The program reports as below, elapsed time used, all pairs as specified in section MinHashAccuracy. This is for the space data set. For this data set, the term document matrix can be stored in memory.

Time taken to construct an instance of MinHashSimilarities: 3.487 seconds

Time taken to compute exact Jaccard similarity: 4.595 seconds

Time taken to approximate Jaccard similarity: 0.064 seconds

4 Near Duplicate Detector

The S shape location is set to be lower than the threshold according to the lecture. This allows higher recall, which is more valuable than precision, as false positives can be eliminated after retrieval, but false negatives cannot be retrieved at all.

For $s = 0.99$, we have almost perfect results.

Average precision:0.9893254437073162

Average recall:0.9974175274908363

Average true Jaccard:0.9669295282969471

The example retrieval is in Table 4.1.

For all results, go to `https://gist.github.com/Fuchai/42002c0f939f824f747634c88f1bb0a8`

For $s = 0.9$, the computation is slower so not every original document is queried, so I cannot compute precision or recall.

The example retrieval is in Table 4.2.

For some results, go to `https://gist.github.com/Fuchai/7f85962360c1774b4b07623af164f35d`

Original	Retrieved
space-988.txt	space-988.txt, space-988.txt.copy7, space-988.txt.copy1, space-988.txt.copy2, space-988.txt.copy5, space-988.txt.copy6, space-988.txt.copy3, space-988.txt.copy4
space-969.txt	space-969.txt.copy7, space-969.txt.copy6, space-969.txt.copy5, space-969.txt.copy4, space-969.txt.copy3, space-969.txt, space-969.txt.copy2, space-969.txt.copy1
space-931.txt	space-931.txt.copy2, space-931.txt.copy1, space-931.txt.copy4, space-931.txt.copy3, space-931.txt.copy6, space-931.txt.copy5, space-931.txt.copy7, space-931.txt
space-901.txt	space-901.txt.copy6, space-901.txt.copy7, space-901.txt.copy2, space-901.txt, space-901.txt.copy3, space-901.txt.copy4, space-901.txt.copy5, space-901.txt.copy1
hockey58.txt	hockey58.txt, hockey58.txt.copy1, hockey58.txt.copy2, hockey58.txt.copy3, hockey58.txt.copy4, hockey58.txt.copy5, hockey58.txt.copy6, hockey58.txt.copy7
hockey218.txt	hockey218.txt.copy1, hockey218.txt.copy2, hockey218.txt.copy3, hockey218.txt.copy4, hockey218.txt.copy5, hockey218.txt, hockey218.txt.copy6, hockey218.txt.copy7
baseball67.txt	baseball67.txt, baseball67.txt.copy1, baseball67.txt.copy2, baseball67.txt.copy3, baseball67.txt.copy4, baseball67.txt.copy5, baseball67.txt.copy6, baseball67.txt.copy7
baseball643.txt	baseball643.txt.copy5, baseball643.txt.copy4, baseball643.txt.copy7, baseball643.txt.copy6, baseball643.txt.copy1, baseball643.txt, baseball643.txt.copy3, baseball643.txt.copy2
hockey429.txt	hockey429.txt.copy5, hockey429.txt.copy6, hockey429.txt.copy7, hockey429.txt.copy1, hockey429.txt.copy2, hockey429.txt.copy3, hockey429.txt.copy4, hockey429.txt
hockey414.txt	hockey414.txt, hockey414.txt.copy4, hockey414.txt.copy3, hockey414.txt.copy2, hockey414.txt.copy1, hockey414.txt.copy7, hockey414.txt.copy6, hockey414.txt.copy5

Table 4.1: 10 example retrievals with $s = 0.99$

Original	Retrieved
baseball114.txt	baseball114.txt, baseball114.txt.copy7, baseball114.txt.copy5, baseball114.txt.copy6, baseball114.txt.copy3, baseball114.txt.copy4, baseball114.txt.copy1, baseball114.txt.copy2
baseball113.txt	baseball113.txt.copy3, baseball113.txt.copy2, baseball113.txt.copy5, baseball113.txt.copy4, baseball113.txt.copy1, baseball113.txt.copy7, baseball113.txt.copy6, baseball8.txt.copy6, baseball8.txt.copy5, baseball8.txt.copy7, baseball8.txt.copy2, baseball8.txt.copy1, baseball8.txt.copy4, baseball8.txt.copy3, baseball8.txt, baseball113.txt
baseball127.txt	baseball147.txt, baseball147.txt.copy6, baseball147.txt.copy5, baseball147.txt.copy7, baseball147.txt.copy2, baseball147.txt.copy1, baseball147.txt.copy4, baseball147.txt.copy3, baseball127.txt.copy4, baseball127.txt.copy3, baseball127.txt.copy6, baseball127.txt.copy5, baseball127.txt.copy7, baseball127.txt.copy2, baseball127.txt.copy1, baseball127.txt
baseball152.txt	baseball78.txt.copy6, baseball78.txt.copy7, baseball78.txt.copy4, baseball78.txt.copy5, baseball78.txt.copy2, baseball78.txt.copy3, baseball78.txt.copy1, baseball78.txt, baseball152.txt.copy2, baseball152.txt.copy1, baseball152.txt.copy7, baseball152.txt.copy6, baseball152.txt.copy5, baseball152.txt.copy4, baseball152.txt.copy3, baseball152.txt
baseball162.txt	baseball162.txt, baseball214.txt, baseball165.txt.copy7, baseball165.txt.copy6, baseball165.txt.copy5, baseball165.txt.copy4, baseball165.txt.copy3, baseball165.txt.copy2, baseball165.txt.copy1, baseball165.txt, baseball214.txt.copy5, baseball214.txt.copy4, baseball214.txt.copy7, baseball214.txt.copy6, baseball214.txt.copy1, baseball214.txt.copy3, baseball214.txt.copy2, hockey932.txt, baseball162.txt.copy4, baseball162.txt.copy5, baseball162.txt.copy6, baseball162.txt.copy7, baseball162.txt.copy1, baseball162.txt.copy2, baseball162.txt.copy3
baseball164.txt	baseball17.txt.copy1, baseball17.txt.copy2, baseball17.txt.copy3, baseball17.txt.copy4, baseball17.txt.copy5, baseball17.txt.copy6, baseball17.txt.copy7, baseball164.txt, baseball344.txt, baseball17.txt, baseball164.txt.copy1, baseball164.txt.copy4, baseball164.txt.copy5, baseball164.txt.copy2, baseball164.txt.copy3, baseball164.txt.copy6, baseball164.txt.copy7, baseball344.txt.copy3, baseball344.txt.copy2, baseball344.txt.copy5, baseball344.txt.copy4, baseball344.txt.copy1, baseball344.txt.copy7, baseball344.txt.copy6

baseball165.txt	baseball162.txt, baseball214.txt, baseball165.txt.copy7, baseball165.txt.copy6, baseball165.txt.copy5, baseball165.txt.copy4, baseball165.txt.copy3, baseball165.txt.copy2, baseball165.txt.copy1, baseball172.txt.copy3, baseball172.txt.copy4, baseball172.txt.copy1, baseball172.txt.copy2, baseball172.txt.copy7, baseball172.txt.copy5, baseball172.txt.copy6, baseball295.txt.copy7, baseball295.txt.copy6, baseball295.txt.copy5, baseball295.txt.copy4, baseball295.txt.copy3, baseball295.txt.copy2, baseball295.txt.copy1, baseball165.txt, baseball214.txt.copy5, baseball214.txt.copy4, baseball214.txt.copy7, baseball214.txt.copy6, baseball214.txt.copy1, baseball214.txt.copy3, baseball214.txt.copy2, baseball295.txt, baseball172.txt, baseball162.txt.copy4, baseball162.txt.copy5, baseball162.txt.copy6, baseball162.txt.copy7, baseball162.txt.copy1, baseball162.txt.copy2, baseball162.txt.copy3
baseball168.txt	baseball980.txt, baseball168.txt.copy3, baseball168.txt.copy2, baseball168.txt.copy5, baseball168.txt.copy4, baseball168.txt.copy1, baseball168.txt.copy7, baseball168.txt.copy6, baseball876.txt, baseball565.txt, baseball49.txt.copy6, baseball49.txt.copy7, baseball49.txt.copy4, baseball49.txt.copy5, baseball49.txt.copy2, baseball49.txt.copy3, baseball49.txt.copy1, baseball168.txt, baseball572.txt, baseball92.txt.copy5, baseball92.txt.copy4, baseball92.txt.copy7, baseball92.txt.copy6, baseball92.txt.copy1, baseball92.txt.copy3, baseball92.txt.copy2, baseball565.txt.copy4, baseball565.txt.copy3, baseball565.txt.copy2, baseball565.txt.copy1, baseball565.txt.copy7, baseball565.txt.copy6, baseball565.txt.copy5, baseball876.txt.copy1, baseball876.txt.copy4, baseball876.txt.copy5, baseball876.txt.copy2, baseball876.txt.copy3, baseball876.txt.copy6, baseball876.txt.copy7, baseball49.txt, baseball92.txt, baseball980.txt.copy3, baseball980.txt.copy2, baseball980.txt.copy1, baseball980.txt.copy7, baseball980.txt.copy6, baseball980.txt.copy5, baseball980.txt.copy4, baseball572.txt.copy7, baseball572.txt.copy5, baseball572.txt.copy6, baseball572.txt.copy3, baseball572.txt.copy4, baseball572.txt.copy1, baseball572.txt.copy2
baseball170.txt	baseball170.txt.copy2, baseball170.txt.copy1, baseball170.txt.copy4, baseball170.txt.copy3, baseball170.txt.copy6, baseball170.txt.copy5, baseball170.txt.copy7, baseball170.txt
baseball179.txt	baseball179.txt, baseball179.txt.copy7, baseball179.txt.copy6, baseball179.txt.copy1, baseball179.txt.copy5, baseball179.txt.copy4, baseball179.txt.copy3, baseball179.txt.copy2

Table 4.2: 10 example retrievals with $s = 0.9$