See the Excel file for data to be discussed, as well as how the data was acquired.

From this data, it appears the differences in search times between data sets are negligible; the greatest difference among the document sets’ average times was slightly less than 500ms. We note that the set with the average fastest search times was the Shakespeare set—a much larger set than the two-document Mini example.

This discrepancy does not seem to be explainable by overhead—if the Mini set’s increased search times were due to too much map/reduce bookkeeping, one would think that the bookkeeping itself would establish the minimum amount of time for an operation. However, since the Shakespeare set takes less time than the Mini set, it is possible these fluctuations in performance are due to overall server load, and not necessarily something exclusively reliant on the data set itself.

Conversely, the indexing operations saw the Mini set understandably completing quickly, while the larger sets took longer to process—the Shakespeare sets took over three minutes on average to index. Nevertheless, we noticed some outlier timings when recomputing indices for testing the search functions. At one point, the Given input set was on par with the Mini set for a fast completion time. In this case, correct search results were still printed. So, again, the performance of map/reduce jobs seems to be heavily influenced by server load, and perhaps only slightly by input size. However, this only considers input sizes on the order of megabytes, instead of what may be seen in a typical data center. Therefore, due to the state of the server, the sample size of the data is likely too small to make any meaningful conclusions.