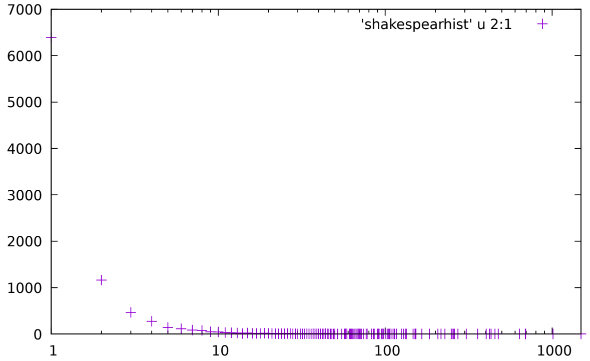
Map Reduce Design : Word Count Frequency

1 Word Count Frequency

This is a design activity. There is no programming involved.

We are given a set of text files. We want to generate for each textfile a histogram that present how many words appear a particular number of times. Look for instance at the results for Shakespeare's Hamlet. There are 6390 words that appeared exactly once, 1162 words who appeared exactly twice and 466 words who appeared exactly three times.



The problem is to design a map reduce application that can perform that analysis on many files at once. That is to say for each file, the application should generate the histogram of how many words appear how many times.

The application should be designed to be a single pipeline. And the application should be designed so that all map and reduce tasks require a constant amount of memory (read: $O(1)$). However, there is no restriction on how many key-value pairs they can generate or ingested by a task. The program should generate the histogram for each file before completing.

Question: Specify the application by describing each phase, whether it is a map phase, a reduce phase, or a cross product. Describe how the tasks are generated/parametrized. Describe what the computation is performed by the tasks in each phase. Describe in between phases the tuples that are generated and what the keys and values are.