

CS425 MP1 Report

Algorithm

We chose to use Python to write this project, mainly because we both felt comfortable with the language, although we were aware there are better options (in terms of performance and existing libraries). We used a total of 10 threads, each spinning up a server, to simulate a distributed system of 10 virtual machines (VMs) that were connected with sockets. VMs could also act as clients and execute a “distributed grep” on all active VMs, each one having a unique file that it would query. Each server would send its results to the client, and the client would aggregate and print the findings.

Tests

We have a total of 11 test cases covering: (1) basic, common patterns, (2) regex patterns, (3) a variety of flags, like case sensitivity, (4) no matches, (5) and server shutdown.

Latency and results

Tests were run on 4 machines, each with a 60MB file, and 5 test runs to compute an average for a run. We used different words with varying, increasing average counts in each file and noticed that the more matches grep finds, the longer the latency for that query. The words we used, alongside their average counts across all 10 VMs, the average latency, and standard deviation, can be found below.

