CS 2150 – mini Google
Project Requirements

To secure full credit, the following requirements must be met.

1. The design and implementation of an indexing component to index data items. The indexing component of your design must strike an appropriate balance to achieve high levels of parallelism and efficient use of the computing cluster. A cluster will be available for this project. The unit of work per indexer can vary from a chapter to a page, or even a finer level of granularity. You may want to think how a distributed structure of directories/folders can be designed to support efficient work assignment and execution, while avoiding excessive contention among workers. Your system should detect and avoid duplicate indexing of the same item.

2. The design and implementation of the distributed inverted index. The design must enable a high level of parallelism and support both indexing and search queries efficiently.

3. The design and implementation of search engine to handle search queries, using different patterns. The system should support simultaneous indexing and searching queries, while preserving consistency of the distributed inverted index and guaranteeing correct results of the search query. You also need to develop a ranking system, including a set of criteria to rank—order hits and a display strategy to list hits.

4. In addition to all the mini Google components discussed above, the design and implementation of the same mechanisms using a Hadoop platform, using MapReduce.

5. Include log files of your experiments on the Hadoop cluster. It is highly recommended to carry out extensive testing **on the Hadoop cluster**, before submission. Results obtained by running the program on your own computer will not be sufficient. It has been the case that programs that work on a local computing environment fail on an actual Hadoop cluster. Therefore, debug the program on the Hadoop cluster before final submission.

6. Grading of this type of projects is time consuming. To ensure fairness, while minimizing grading time, the project submission should include the project software, a compiled version of the project software that captures all functionalities implemented in this project, a **makefile** and a brief **"how-to?" instruction manual** to compile and execute the project on a Hadoop cluster.

7. Extra credit: support for *speculative execution*, whereby when tasks in a job are ending, redundant copies of the remaining tasks are scheduled across several nodes that do not have other work to perform. This ensures that the same input can be processed multiple times in parallel, to exploit differences in machine capabilities.