Edmond Wu
Vincent Xie
CS417

Hadoop Project Phase 1 Readme

Assumptions:

- Word counts are case-insensitive (outputs are all in lower-case). So if the input contains “The” and “the,” it will assume they’re the same word.
- Hyphenated words count as 1 word (ex: “bee-hive” counts as 1 word). This can sometimes lead to unexpected results if you have hyphens separated by delimiters (ex: a hyphen alone and a delimiter could be seen as a word).
- We also made some assumptions in terms of delimiters for words. The delimiter string we used was: `" \t\n\r\f.,;:\"!?@#$%^&*()/+_<>~[]{}|\\".`

Challenges:

This assignment took us about 3-4 hours to finish. The Java source code was not too difficult, as it was just a matter of making an implementation of a mapper and a reducer class. The majority of the time spent on this assignment was attempting to figure out Amazon’s services, using Amazon EMR and AWS to utilize the jar file we created to run the program. We set up an S3 bucket with an input folder (containing various test text files) and the jar file (map-reduce-hadoop.jar). The program generates an output folder containing a file that displays the word counts. If a folder with the same name exists, it won’t run, so delete any old output folders first or specify a new output directory.