HLD Report: Word Frequency Project

Team Members: Sohail Ahmad, Shirwan Kanj, Arthur Fong

1. Background:

When given the various choices of projects to showcase our learning progress in the Python programming language, we found the word frequency analysis project to be the most interesting. This project essentially requires us to write a program that performs analysis on an input text file; it would go through and read the text file, analyze it, and output the necessary information. Once we have a functional program that can perform this analysis, we must test its performance by measuring how quickly it can run on varying amounts of computing power. For this example, we will be testing the performance by changing the number of threads 1-8 allocated to the program.

2. Each Members Work:

Each member of the team worked on conducting some background research that would be useful to our work on this project – with virtual meetings conducted to discuss our ideas and questions.

Sohail Ahmad: Researched information relevant to the project, worked on the HLD report.

Shirwan Kanj: Researched information relevant to the project, worked on the HLD report.

Arthur Fong: Researched information relevant to the project, worked on the HLD report.

3. Dataset:

In this project we will be utilizing a text file as an input for our analysis program. We were introduced to Project Gutenberg, which is a free resource that provides easy access to a library of ebooks in the text file format. We made use of the website to choose “The island

pirate, a tale of the Mississippi", written by Mayne Reid, as our input dataset for the project. This will be the text file that we will be doing our word frequency analysis on.

4. Milestone:

There are several smaller milestones that we plan to work towards in order to complete this project. This will help to break down the problem into smaller problems so we can work on them one at a time. The first milestone would be to be able to understand the process of reading an input text file in Python, which will help get us started with the project. The second milestone would be to figure out a way to store all the string text, and remove all punctuations present in the English language from the text; this would help us to have a simple file with only words separated by spaces. The third milestone we would need to work towards is analyzing that simple text file with just the words, where we would have a function to go through the file and essentially count how many times each word appears in the file. Then, the last milestone would be to sort those words based on their frequency count and output the results in a text file. Through completing these smaller milestones, we should be able to have a final functioning program that should sufficiently fulfill the requirements of this project.