CS 722 Project 2

Kyle Goodwin and Sam Harris

1. Dataset

Ten books from Project Gutenburg, the list and corresponding URLs used in test_books.json is:

2. Analysis of the BookReduce index

There were some anomalies as well as some things we expected to see in the final output of our BookReduce implementation. Our algorithm aggregated all words in the ten books, separated them into 5000 word buckets, and ordered by descending frequency per term occurrence for the final index.

Some results were very predictable, for example the word “i” being seen hundreds of times per book or the word “because” being seen in most if not all buckets with varying frequency. One thing we did not account for nor expect were special characters like the underscore character “_” being seen before words or used consistently in general. This is an interesting results likely caused by the use of metadata tags or characters that were not parsed correctly in the Project Gutenburg .txt files, and were not removed by the tokenizer used to parse the text.

There were also many interesting rare terms, many being numbers or dates (for example “144” in Moby Dick bucket 30, or “15th” in Pride and Prejudice bucker 5.