MP4 Development process/reflections

Development process:

* Utilized file i/o to read in all the strings from a Gutenberg e-book.
* Mapped each string (key) to its number of occurrences, continually updating number of occurrences of each string while reading in from file.
  + kept track of individual unique strings each time a new string was found in a vector.
* Iterated through vector holding unique strings, used strings as key for number of times occurred, used that number to map each string to a vector storing each string that occurred the same number of times.
* User inputs an integer n, program iterates through vector corresponding to n, displays all strings that occurred n times.

Observations:

For the purposes of this project, a “word” was defined as any string of characters with a whitespace on either side. Because of this, there are many “words” that occur only once, as a word with a capitalized first letter or any punctuation attached to it (ex. “Since”, “end.”) is counted as distinct from the same word without a capital or attached punctuation mark (ex. “since”, “end”). Still it is easily observable that words that occur only once (of which there are also many) tend to be unique adjectives or specific things.

There are fewer words that occur at high frequencies, and these are almost all prepositions and conjunctions, (such as “and”, “the”, and “a”). As for words which occur at relatively high frequencies and are more specific descriptors, or that relate to more concrete concepts, I think it would be reasonable to think that these could be used to gauge an author’s main themes/sentiments. The thought here is that if the complexity and frequency of a given word are both high, then it will be a word that is important to the author’s thesis or main ideas.