Project

Sunday, May 5, 2019 6:59 PM

This project consists of three parts. This may sound like a lot, but to complete the project, it took me about 20 lines of new code. You will likely want to do your development work using the smaller 50 megabyte file, but you will need to run the 5 gigabyte file before turning the project in. Also, while you can and probably will want to use the show() action while developing and debugging, you cannot have show() anywhere in the final project.

To start, download the following file:

project3

Part I

Count all the terms in the document containing dashes and display the number. Somewhere in this process you will need to call split(). Be sure to call it with no parameters; calling it this way will split on all whitespace. You do not need to get any fancier than that. Further, please do not try to get any fancier because then we will have different answers.

Part II

Count the number of occurrences of every word. For example, maybe the word "you" occurs 7000 times. List the top 25 words, based on the number of occurrences, in the file. List them in descending order. As mentioned in the first paragraph, please refrain from using the show() action for this work item.

Part III

Find the lengths of all the words, and then determine how many times words of that length occur. For example, "at" and "by" are both words of length two. Perhaps there are a million words of length two. List the top 25 lengths and how many times they occur.

Take screenshots or video of the data from parts I, II and III and submit to Blackboard.

Expected Output

Here's (some of) the output from my project:

4344056

Word | Count
the: 17392842
of: 10666816
and: 8767994
to: 7533475
a: 6020151
in: 5146971
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN: 5024122
…

Length | Count
3: 69164029
50: 64855055
2: 56440252
4: 50871351
5: 37358985
…