a. Harsha Thulasi
b. I spent around 2 hours right after class. I used the 40 min at end of class to setup the docker env but after that I spent around an 1 hr compiling and running the java program in hadoop.
c. I am submitting for all there parts. However, the numbers match between part 1 and 2 but the order/terms match. For part1 I wrote a simple python script and ran it.
d. Output for part 1:

    the 2219
    and 2034
    to 1512
    i 1408
    of 1302
    you 1115
    a 994
    my 914
    that 873
    in 808

e. Output for part 2:

    the 2219
    and 2034
    to 1512
    i 1408
    of 1302
    you 1115
    a 994
    my 914
    that 873
    in 808

f. It's a little tricky to compare books based on terms/tokens. Seeing that they are different stories and time periods I personally feel its not very comparable. However, assuming that two different authors write the same story and are from same era it would make sense to be able to use the terms/tokens ratio. Still very subjective and would be missing the context. If anything just the term count could better illustrate the "richer vocabulary". Perhaps do some sort of histogram or variance spread can illustrate the "richer vocabulary".