

1. Implement the following applications in MapReduce by mrjob with the given document “Trump_Tweet”.
 - a. Count the number of occurrences of each alphabetic character in a file. The count for each letter should be case-insensitive (i.e., include both upper-case and lower-case versions of the letter). Ignore non-alphabetic characters. (20 marks)
 - b. Count the number of occurrences of each word in the file. The count should be case-insensitive and ignore punctuations. (20 marks)
 - c. Based on b, count the number of occurrences of 2-gram in the file. A 2-gram is a sequence that contains the adjacent two words. For example, the 2-gram in the sentence “I like to eat pizza” is “I like”, “like to”, “to eat” and “eat pizza”. (20 marks)
 - d. Based on b, obtain the top 20 most frequently occurred words except the stop words such as “a”, “the”, “this” and so on. The stop words can be imported by stop-words package (see <https://pypi.org/project/stop-words/>) (Hint: use 2 steps and the second step only has the reducer) (30 marks)
 - e. Give a brief summary of the style of Trump’s tweet you have found based on the frequently used words. (10 marks)