BIG DATA HW1

1. The Shakespeare.txt file contains all of Shakespeare's writings. Write a program using SPARK to:

a. Specify the total number of words in this text.

b: What is the number of words without repetition?

c: What are the ten words that have the highest number of repetitions and how many times each?

d: Draw a graph of the runtime by changing the number of cores that program (b) runs on from one kernel to at least four cores. Outputs include code, results, graphs, and analysis.

2. The three files C1, C2, C3 each contain two columns of text information, each row showing the coordinates of a two-dimensional point. Write two programs using SPARK using two methods of clustering K-means ++ and Bisecting K-means.

a: For each data file, plot the cost cluster for k = 2 to 25

b: What is the optimal number of clusters?

c: In the optimal number of clusters and in the least expensive clustering method, specify the center points of each cluster for each data file?

d: In the optimal number of clusters, draw and analyze the runtime graph of two clustering methods on one to four cores for each file.