INF 553 Homework 4

My three scala codes, were saved in a package called Georgios.HW4
As a result in front of each task, I have to specify the package.
Even if by the changes made to Task 2 and Task3, I could have made them in 1 scala code, I decided to have 2 different codes, since this is the way I started approaching the problem. Thus the codes to run the three tasks successfully are the following:

Problem1:
spark-submit --class Georgios.HW4.Kmeans Georgios_Iliadis_Clustering.jar W
spark-submit --class Georgios.HW4.Kmeans Georgios_Iliadis_Clustering.jar T

Problem2:
spark-submit --class Georgios.HW4.KTask2 Georgios_Iliadis_Clustering.jar K
spark-submit --class Georgios.HW4.KTask2 Georgios_Iliadis_Clustering.jar B

Problem3:
spark-submit --class Georgios.HW4.KTask3 Georgios_Iliadis_Clustering.jar K
spark-submit --class Georgios.HW4.KTask3 Georgios_Iliadis_Clustering.jar B

Since Task3 always uses only the small text file, k=16, and maxIterations = 20, I have removed them from the input arguments. I have them commented in the first lines of my code if you want to use them.
Same applies for Task2 and Task1.
The only input arguments I left was the feature for Task1, and the algorithm used for Tasks 2 and 3.

First you have to cd where the jar file is and then use the three commands provided to run Task1,2 and 3. I have the data needed in the directory where the .jar file is.

Output Files:
Once I run the command lines provided, the output files will be created where the .jar files are.
I printed the results in json files as required, trying to have the spacing as similar to the provided output as possible.

Note:
I have included some print statements for debugging purposes. I thought not to remove them in order for you to be able to understand how my implementation works and what is really happening. The output files though are created as requested.
It takes a lot of time to run this program, especially the Task3 on my machine since it uses k=16.
Also while running Bisecting Kmeans with K=16 I get an error:
java.lang.OutOfMemoryError: Java heap space.
//conf.set("spark.yarn.executor.memoryOverhead", "809")
java -Xmx2g Regexer
So I used the above to make it to work on my machine.

Name: Georgios Iliadis
3668057286