**How to run Assignment 2 Task 2 Code?**

1. **Set up IntelliJ**

Note: The tool versions are: **JDK 8, sbt 1.3.4, Scala 2.11.12, Spark 2.3.0**

* Create a new project in Intellij  
  Choose Scala and sbt  
  Select proper versions. I use JDK 11, sbt 1.3.4, Scala 2.11.12

Right click on your project directory, add framework support, choose Scala

Go to src/main/scala, right click and create a new scala class

* Add Spark to IntelliJ

Add libraryDependencies += "org.apache.spark" %% "spark-core" % "2.3.0" to build.sbt

1. **File Structure**

|  |  |
| --- | --- |
| Input | Task2\_data/QA\_data/QA\_data.csv |
| Main source code | kMeans/src/main/scala/K\_means\_clustering.scala |
| Jar file built from sbt | kMeans/target/scala-2.11/kmeans\_2.11-0.1.jar |
| Print output | In the console  I set Intellij configuration to save console output to a log file in Task2\_data/output/output.txt  Search for “centroid: ”, this part is the final result output |
| Visualization of results | I wrote visualization\_kmeans.ipynb to extract useful information from output.txt, visualize and analyze the results. |

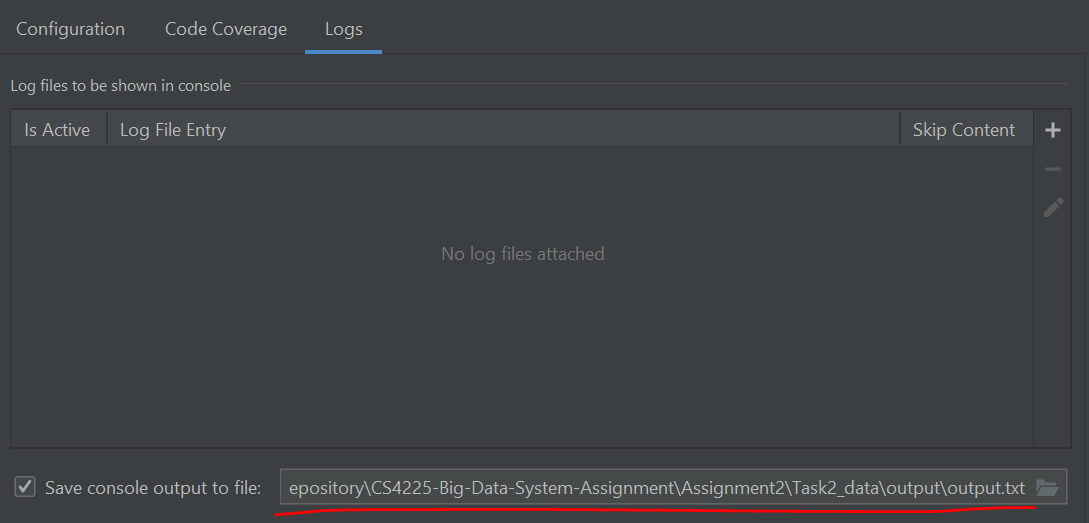
1. **How to run**

* Change the input file paths in the code



* Set the path of log file for the output

Edit running configuration -> go to Logs -> tick “save console output to file” and edit path



The saved log file will be parsed by visualization\_kmeans.ipynb to analyze and visualize the clustering results.

* Click run in Intellij