Unzip the folder AmazonFineFoodReviews. There is a file called AmazonFoodReviewPreprocess.py .This python file is used for preprocessing. You need to run this python script using

The input file finefoods.txt is downloaded from <http://snap.stanford.edu/data/web-FineFoods.html>. Here we get the compressed folder. Unzip the folder and we get a file called finefoods.txt which is the input text file used to run the python script

**python3 AmazonFoodReviewPreprocess.py finefoods.txt output1.csv output2.csv**

or

**python AmazonFoodReviewPreprocess.py finefoods.txt output1.csv output2.csv**

In my case I have python version is 3. So I run the command using python3

To run this python script we need to have pandas installed in your computer. You can use pip to install pandas with command

**pip install pandas** if python version is 2 .You can check python version by running command python –version in command line

**pip3 install pandas** if python version is 3

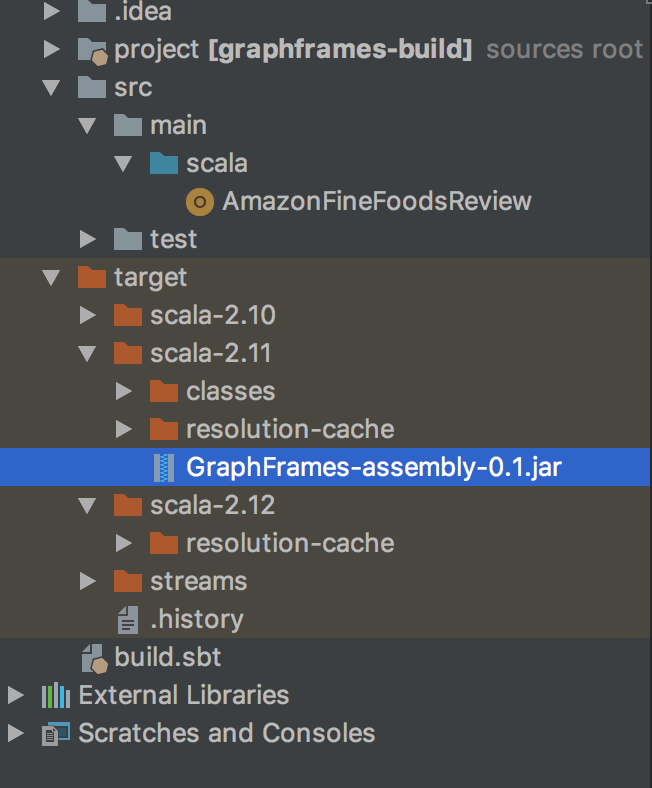
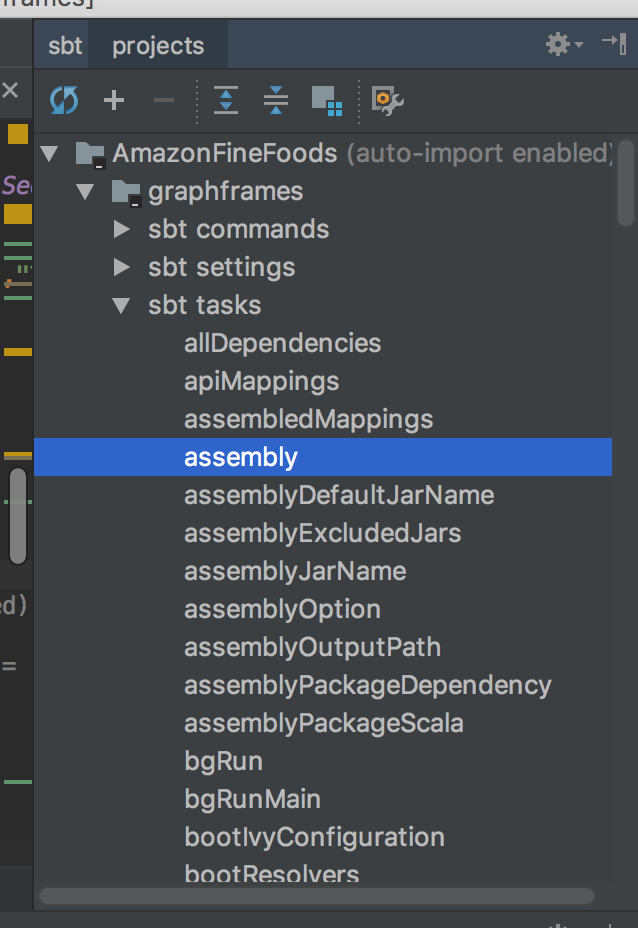
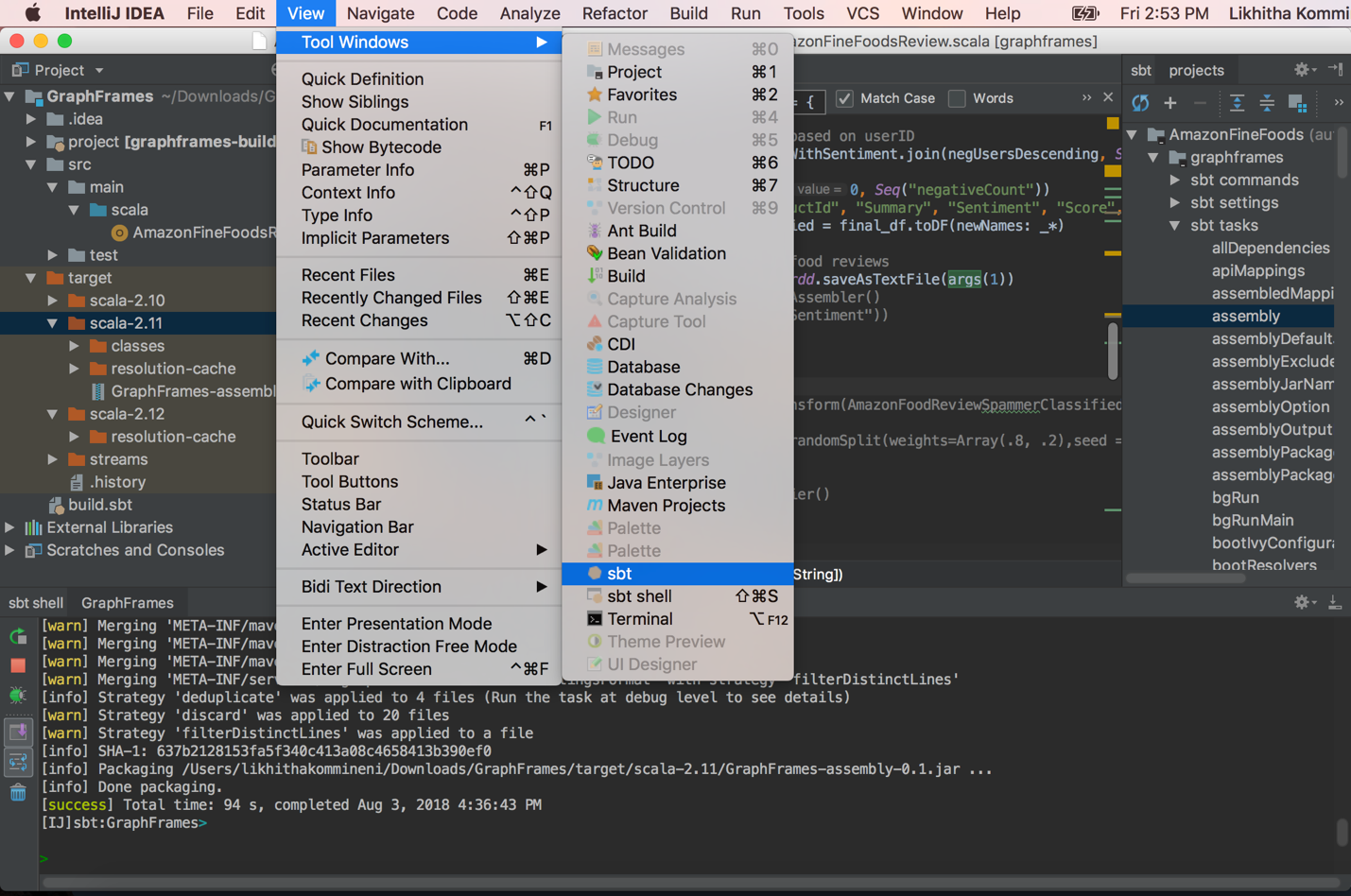
if python is not installed.You need to install python in your computer. You can check whether python is installed or not using command **which python** in Mac Terminal

python3 AmazonFoodReviewPreprocess.py finefoods.txt(input dataset that we got from Stanford) output1.csv(the output csv file1 you can give the csv file path where u want to save.This csv has null rows) output2.csv(to remove the null rows from output1.csv u should give the output csv file 2 path this is final output csv after removing null rows from output1.csv which we use to run the scala project)

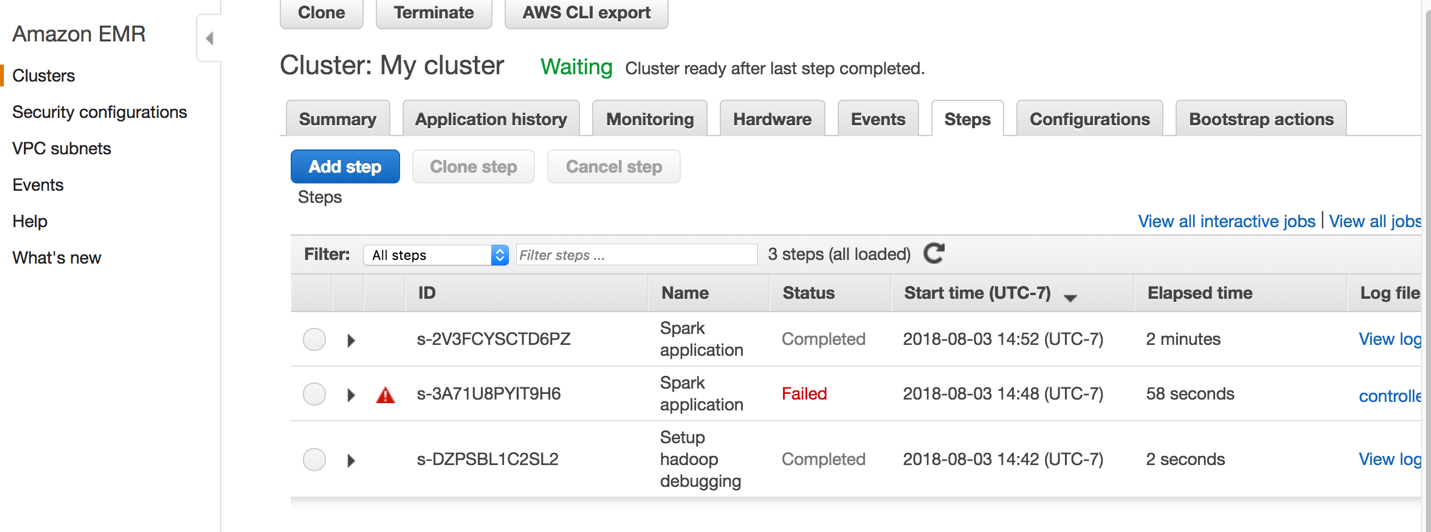
There is a folder called GraphFrames

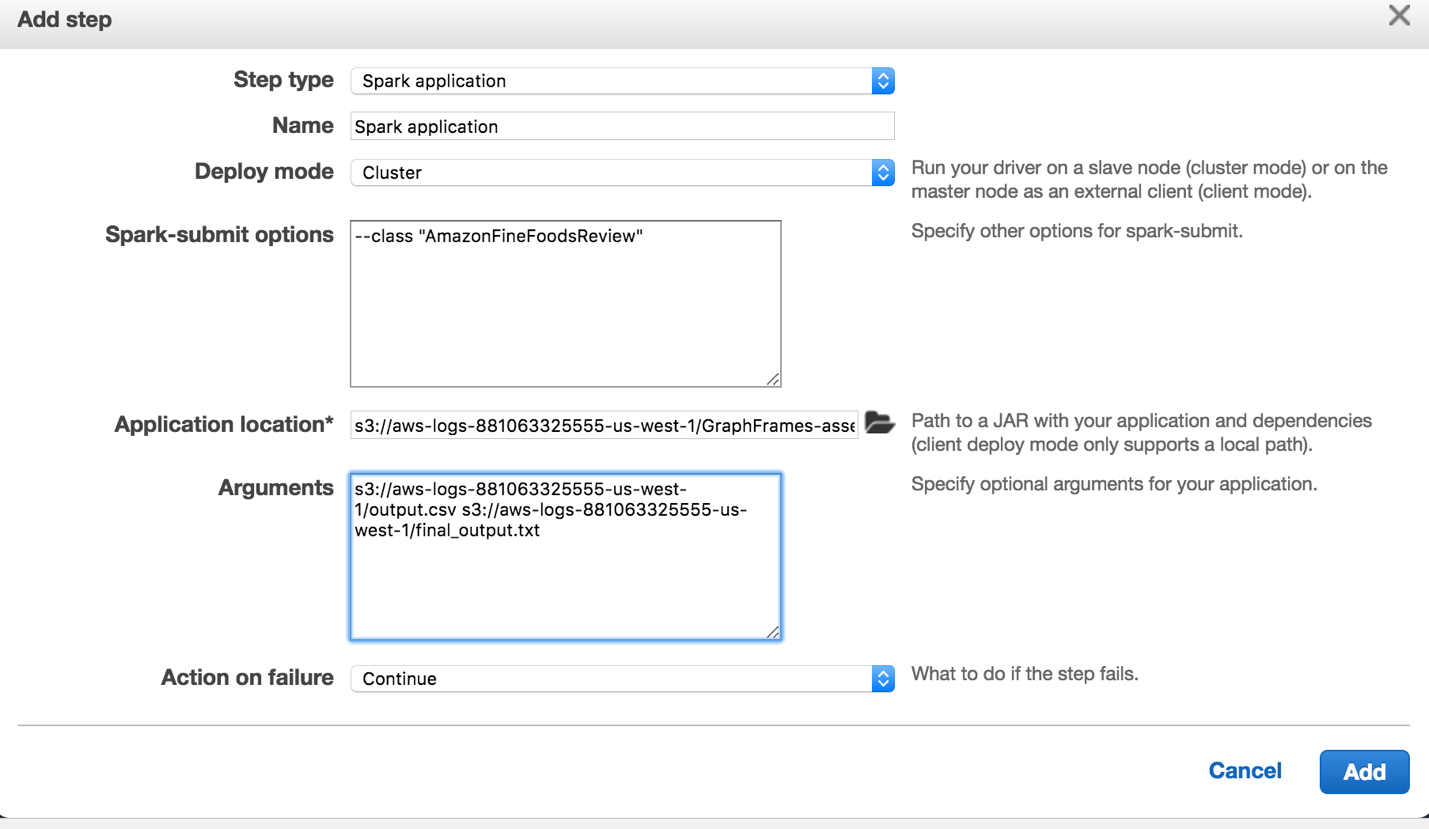
GraphFrames is a sbt project.

Import this project GraphFrames as sbt project in IntelliJ and go to View tab->click on tool windows->Click on sbt.->Go to sbt tasks -> Click on Assembly-> This will generate a jar and you can see it as GraphFrames-assembly-0.1.jar generated in scala-2.11 folder in target folder.This is a fat jar which includes all the dependencies.



This generated jar is used to run the GraphFrames project. Go to aws and s3 and upload the jar file generated i.e GraphFrames-assembly-0.1.jar and output2.csv that’s generated from the python script which doesnot have any null rows

Go to aws and emr and create a cluster with emr-5.15.0 and with applications Hive 2.3.3, Pig 0.17.0, Hue 4.2.0, HBase 1.4.4, Spark 2.3.0, Tez 0.8.4 included .

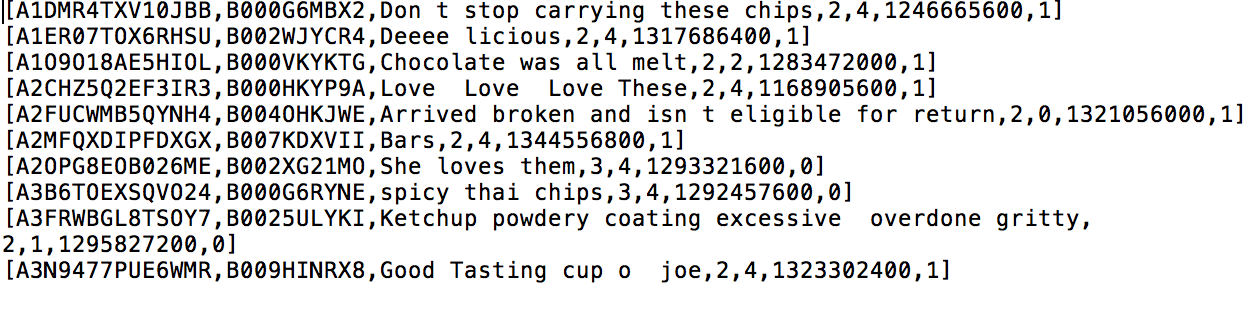
Now after cluster starts .Click on steps tab and click on add step and add step giving the parameters as 

Application location is the path of generated jar file you uploaded in s3

The arguments should be two and 1st argument is the path of second csv file you get after running python script and uploaded in s3 and second argument is the output file path u want to save the output in (note that there should be no directory with same name in s3)

The sample output will be in this manner with the columns in order "UserId", "ProductId", "Summary", "Sentiment", "Score","Timestamp", "spam"

For every row in the output



The last column in every row is the output column we generated i.e spam(spammer) is 1 for spammer users and 0 for non-spammer users.