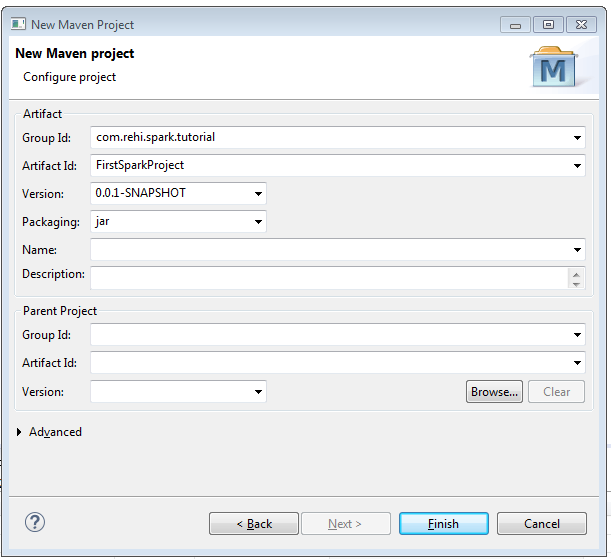
**SPARK PROJECT SETUP:**

1. Install Scala 2.11.8, Java 8, and sbt
2. Download Hadoop\_common\_2.2.0\_bin/zip/master(Hadoop common package as spark requires Hadoop type setup)
3. Download spark\_1.6 bundle with Hadoop for windows.
4. Set Environment variables for Scala\_home, java\_home, path, sbt\_home and Hadop\_home
5. Create a simple maven project and add this dependency:
6. 6.Start eclipse and Create a simple Maven Project:



**THIS IS THE SCREENSHOT OF THE PROJECT**

1. Open pom.xml and add these dependencies:

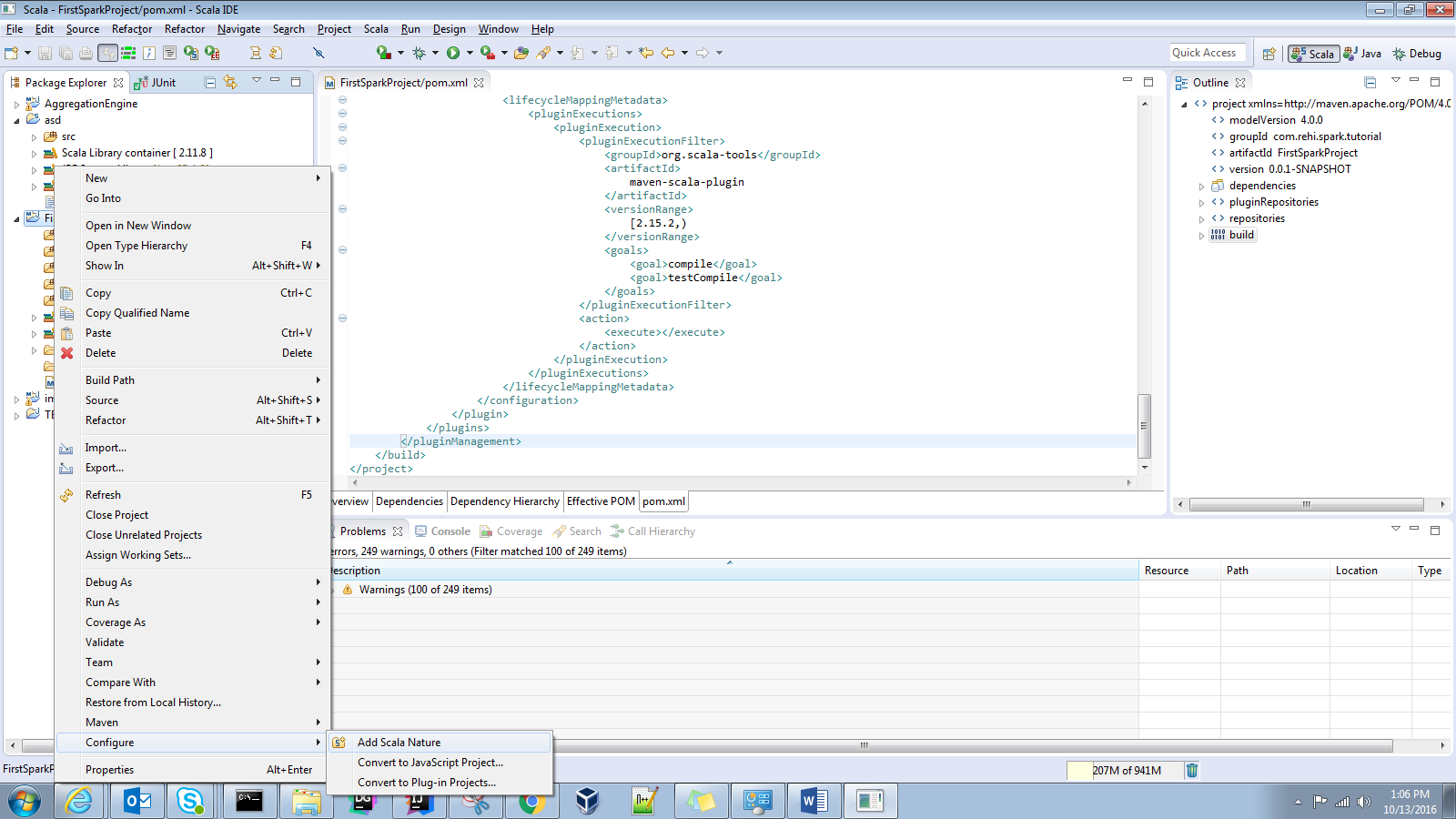


Also a config pom.xml can be found in:

<https://github.com/mcapuccini/spark-tutorial/blob/master/spark-tutorial/pom.xml>

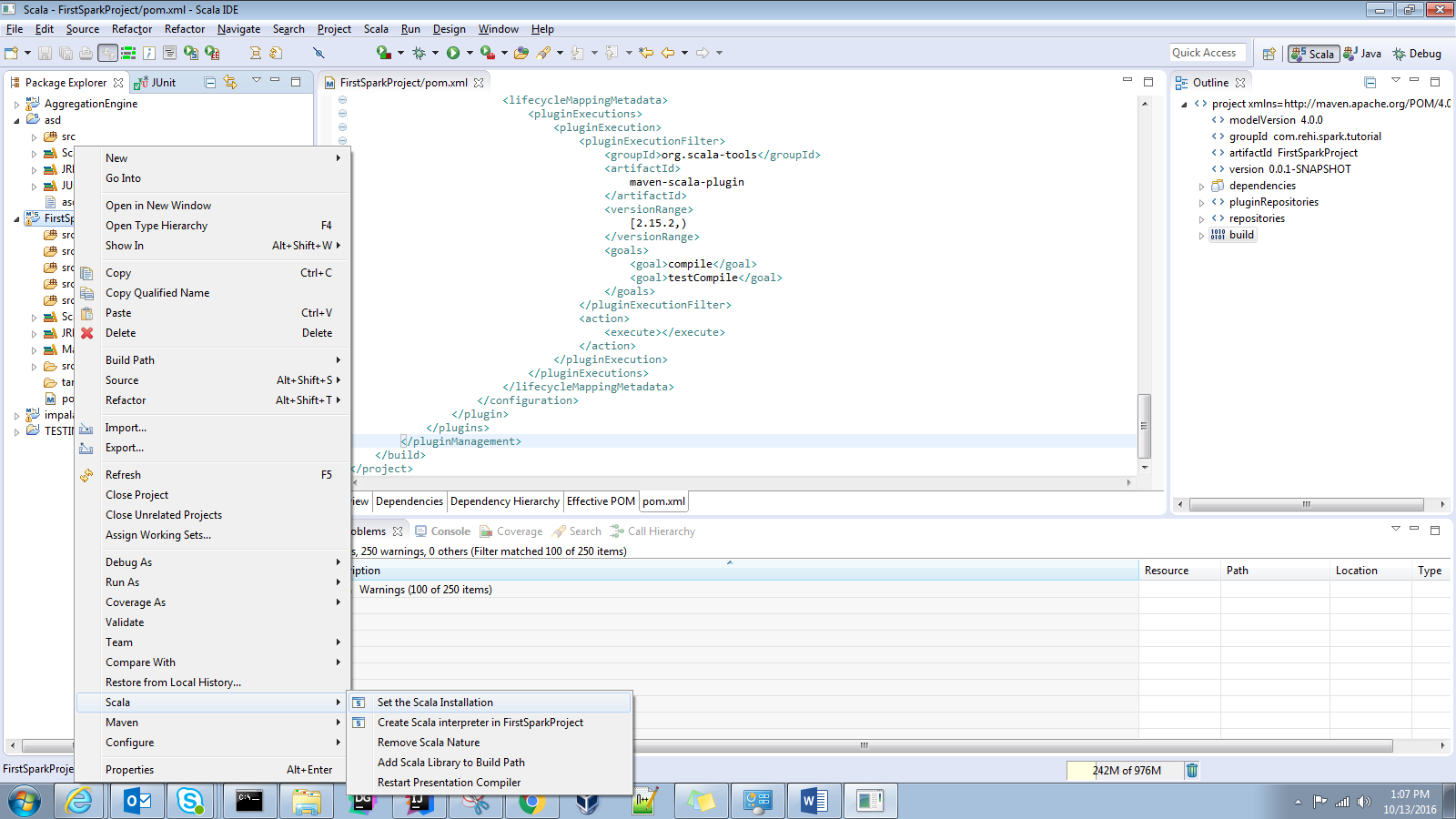
At this point you will get an error the pom.xml is not up to date so just right click on the error and click QuickFIx(The Error will be gone)

1. As of now you cannot create a Scala class as the nature is not present so Just add the Scala nature to the project : (Attached)

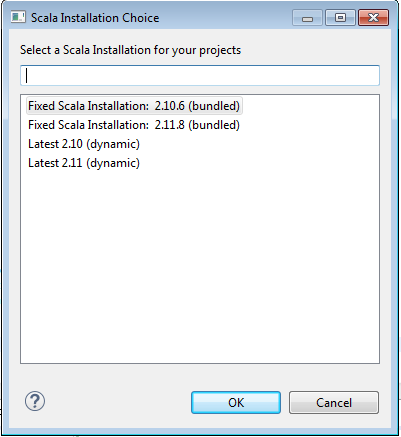


IF you are using latest Scala IDE and the dependencies are of spark-core-2.10

So see this screenshot:

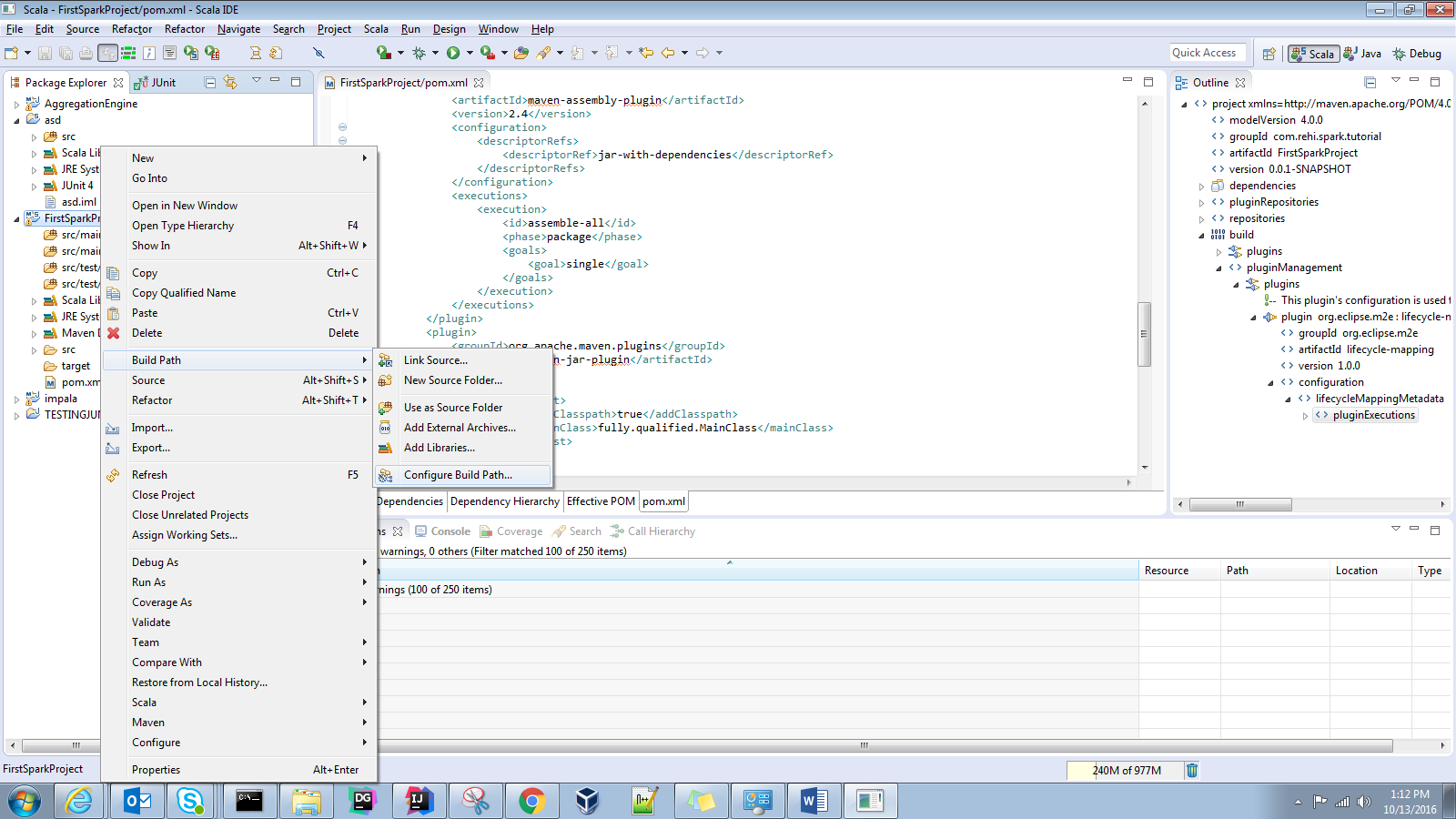


You can set the specific Scala version:

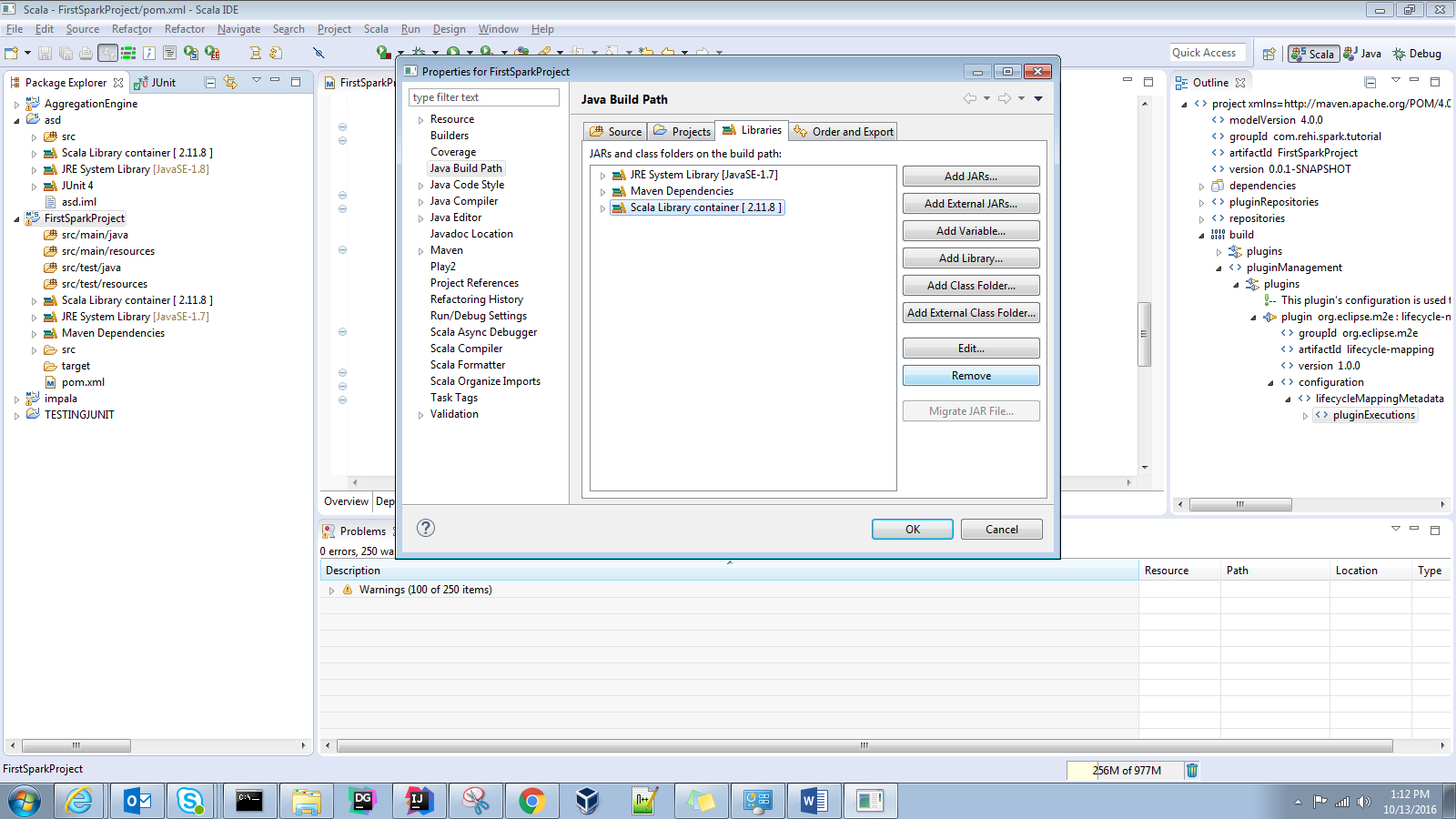


1. Remove the scala\_library Container as this will be pulled by spark core:

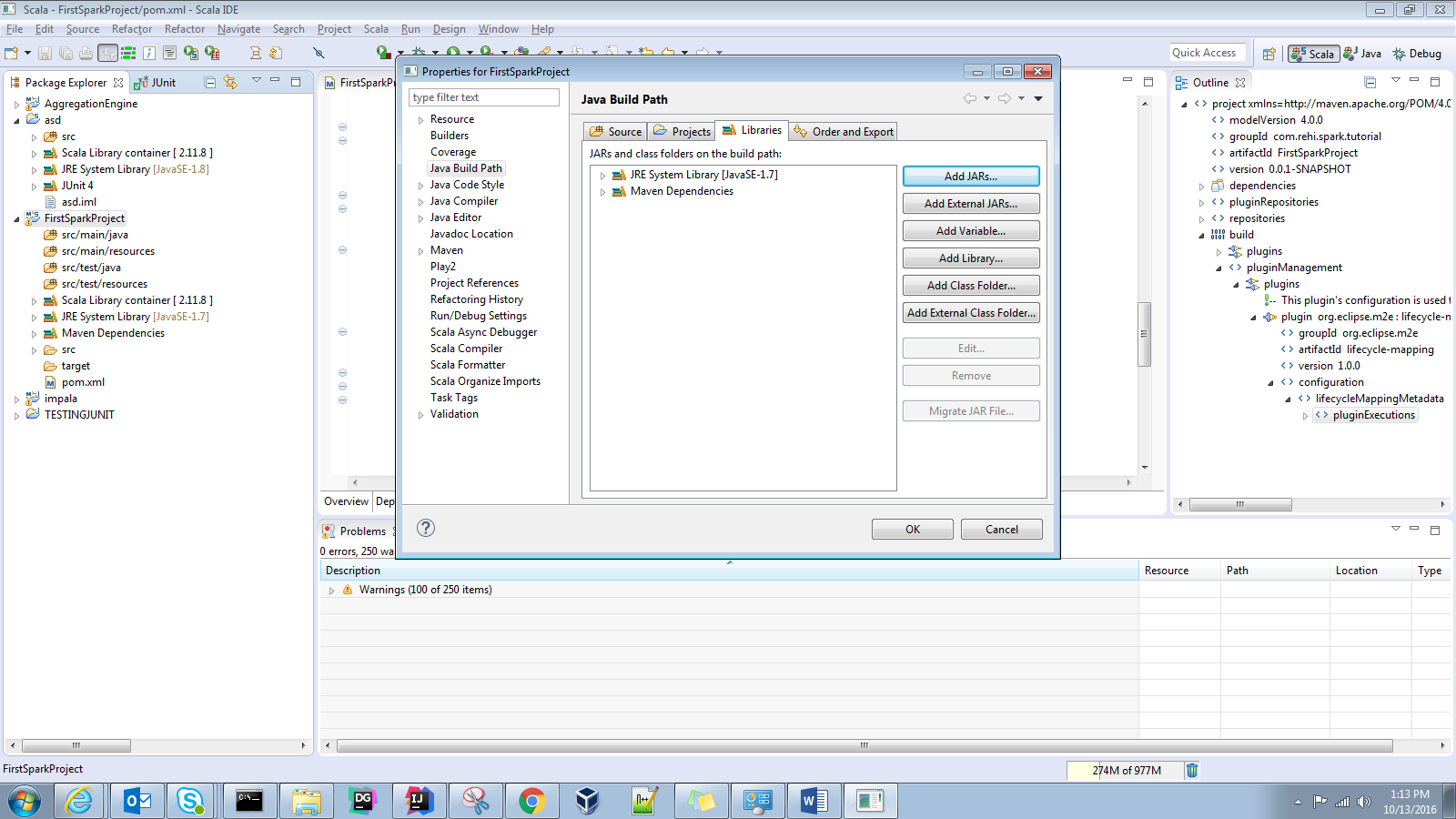
Go to configure Build Path:



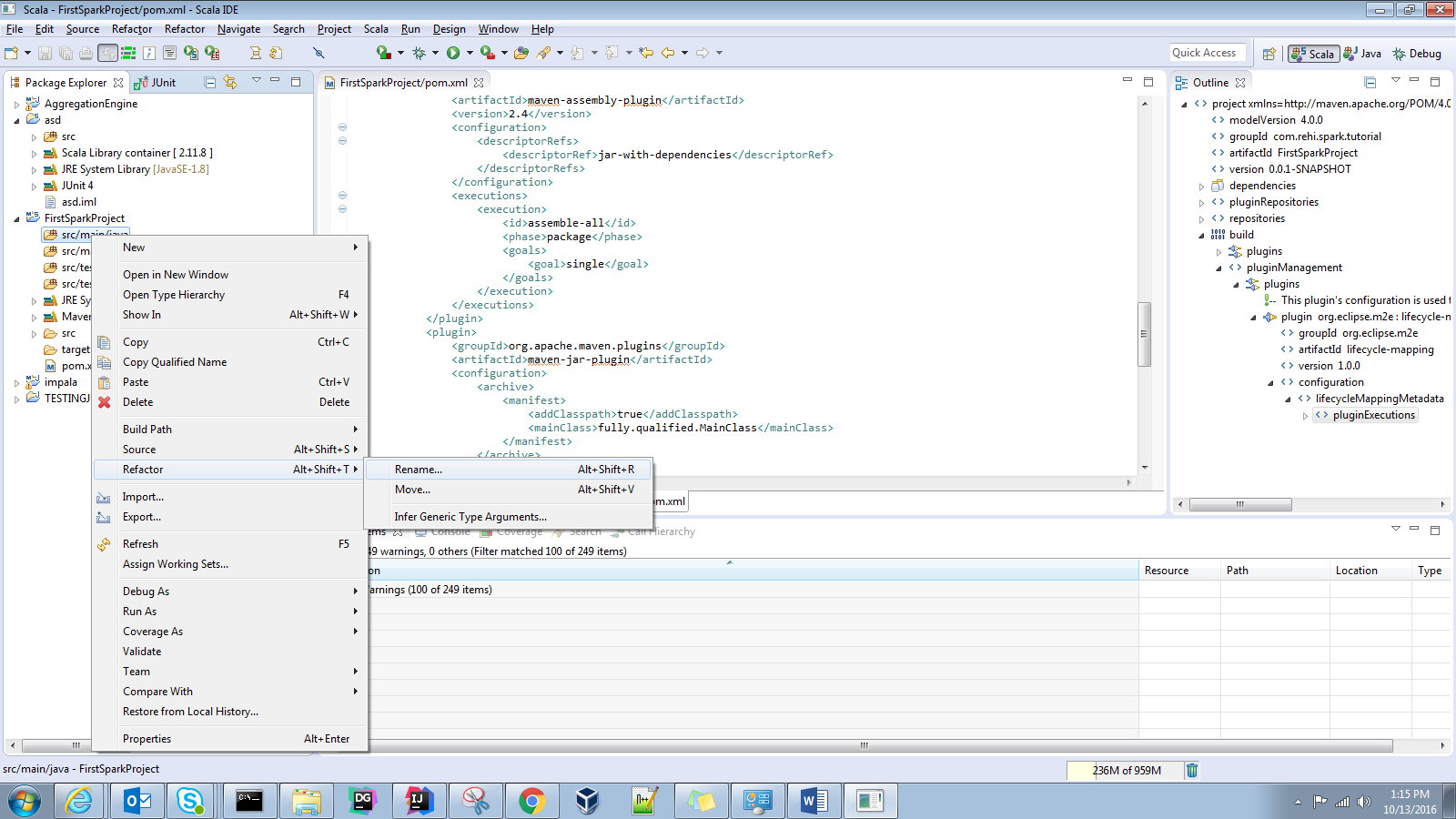
Remove the scala Library container:



After removing: (Press OK)

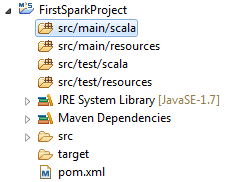


1. Rename the source folder from java to Scala using Refractor:

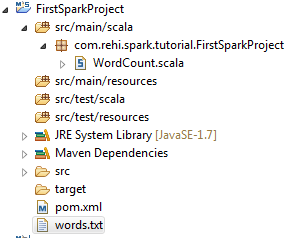


Do this for both src/main/java and src/test/java:

After renaming your project structure will look like this:



1. Create a package : Note the name of the project will be the same as the groupID and artifactID:



Now inside the package create a simple scalaObject that counts the words of the file words.txt



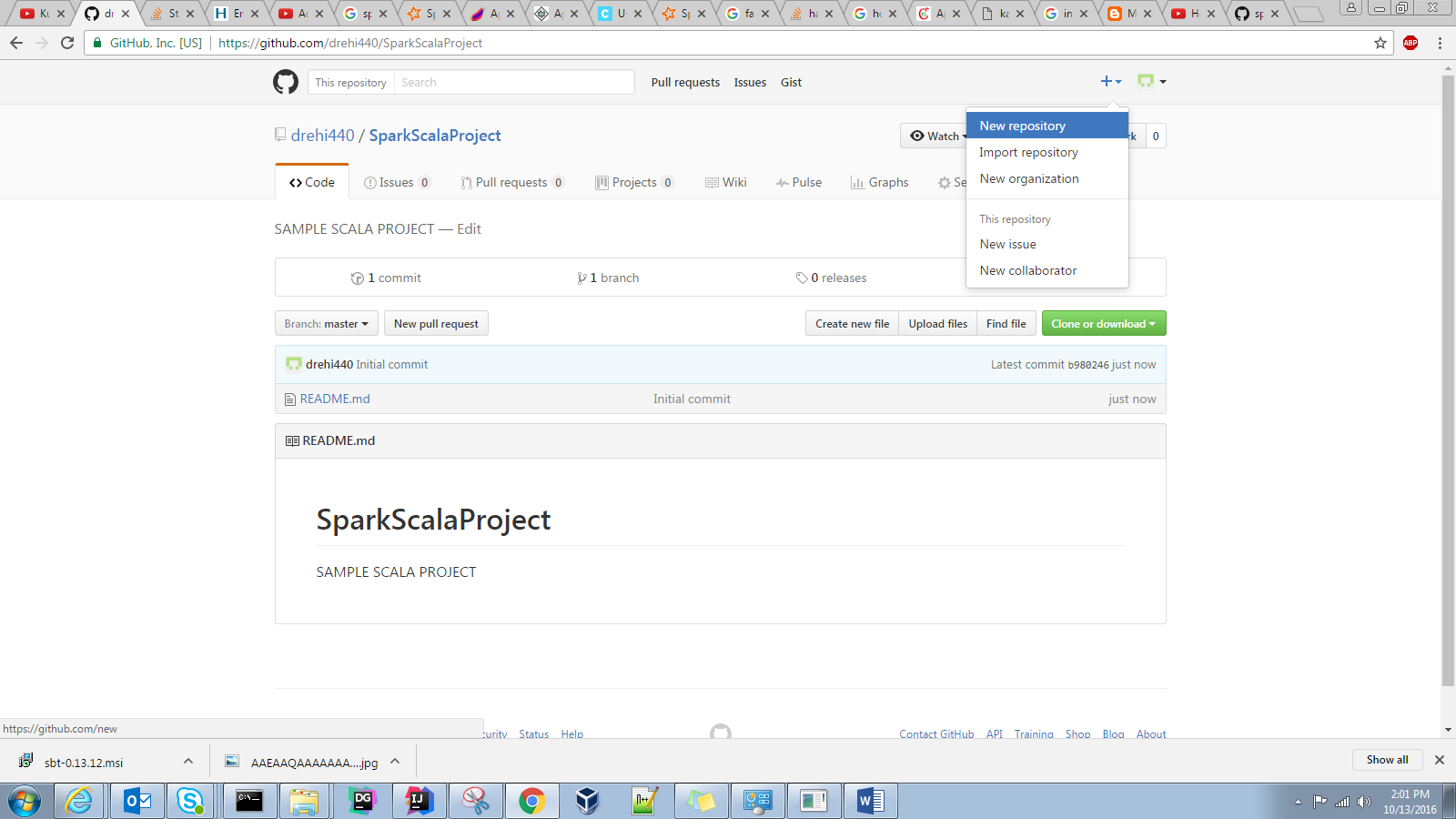
Refer to this sample Scala object



1. You can get Error of winutils.exe not available so you can configure and set the Hadoop\_home variable:

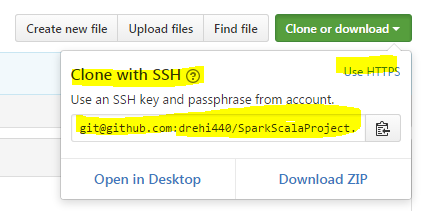
(See point no. 2)

1. You can create a new gitHub account or use your existing one and create a new repository:

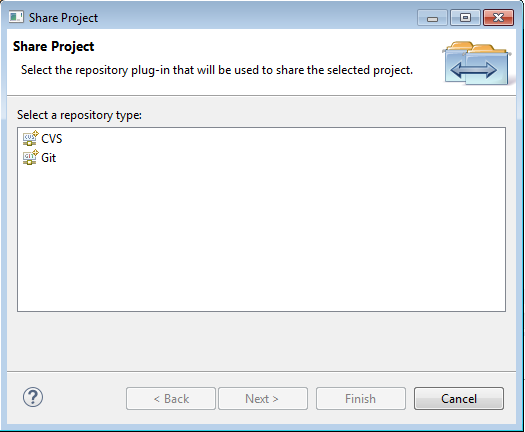


Fill the name and click on create repository: you will see the above screen

1. Go to clone or download and click on Use SSH you will get this below like link :

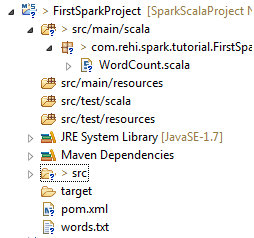


1. Go to the source folder right click -> Team-> Share Project



Click on git-> Create new Repository: Name the repository as your git repository (Best practice however you can name anything)

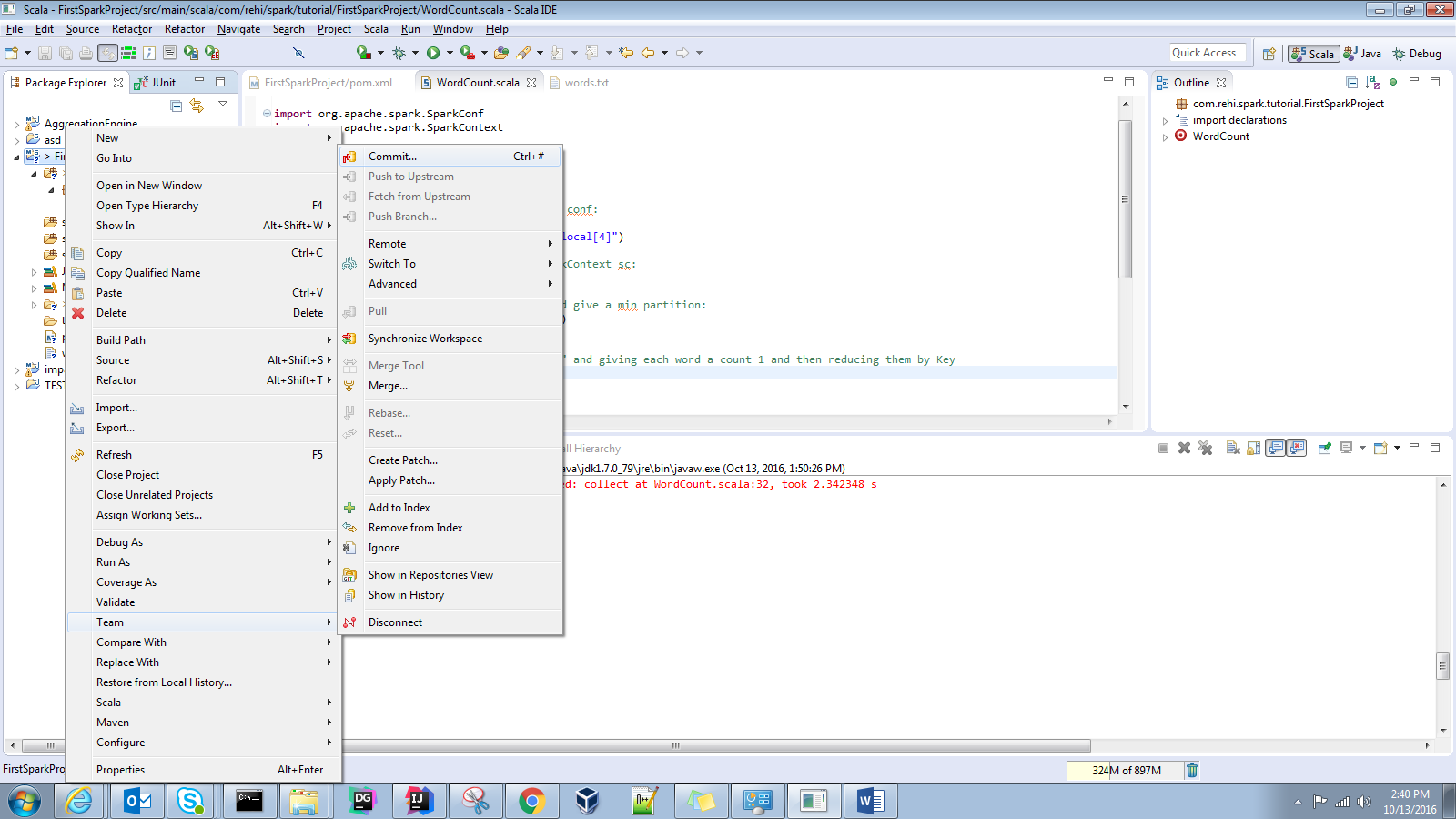
And click on finish:

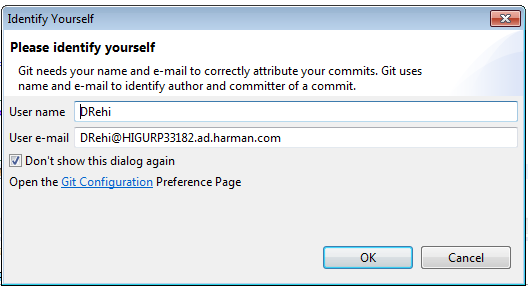


Also here I have deleted the output folder as It was not required!

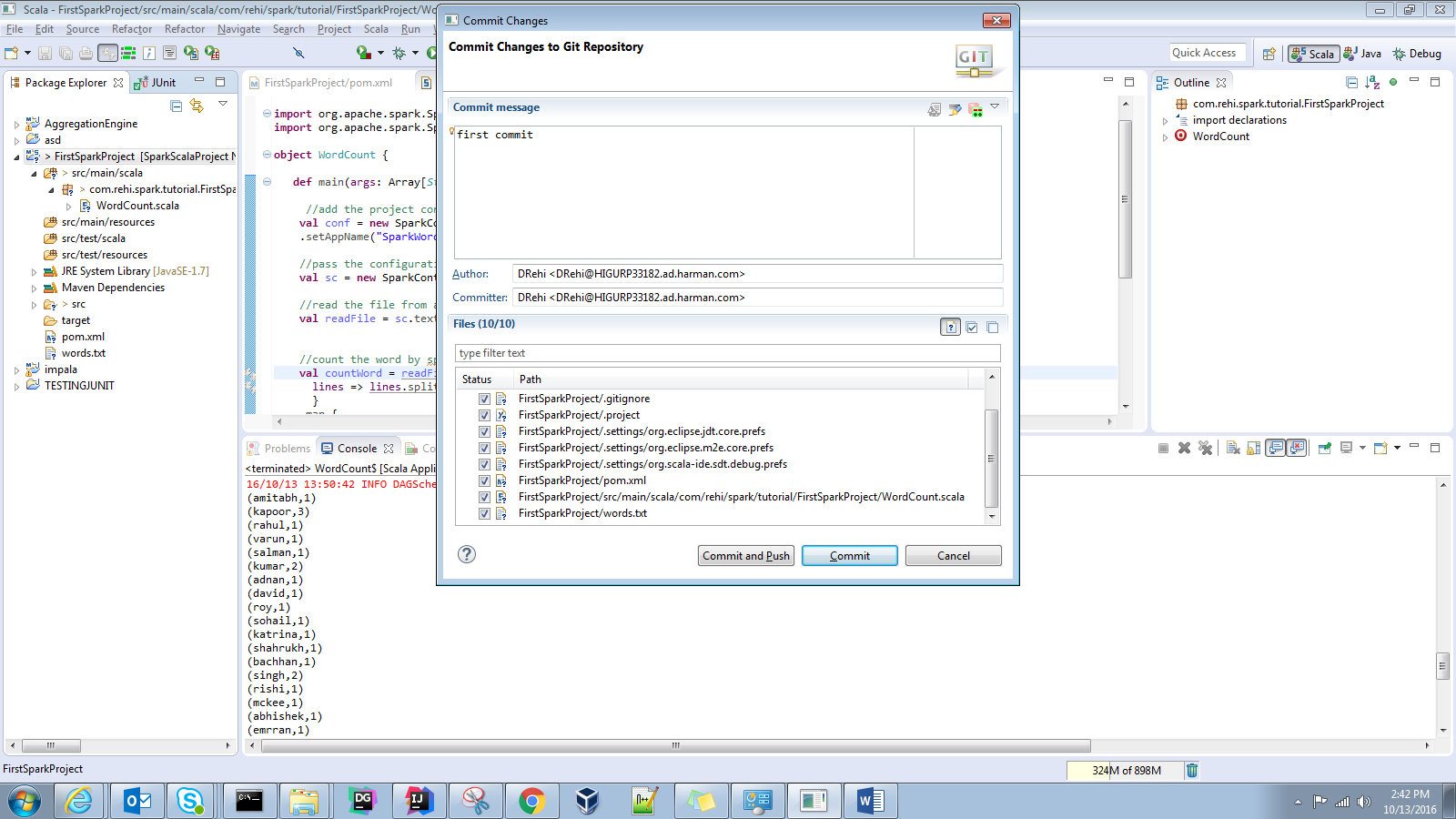
1. Now we have to do Commit:

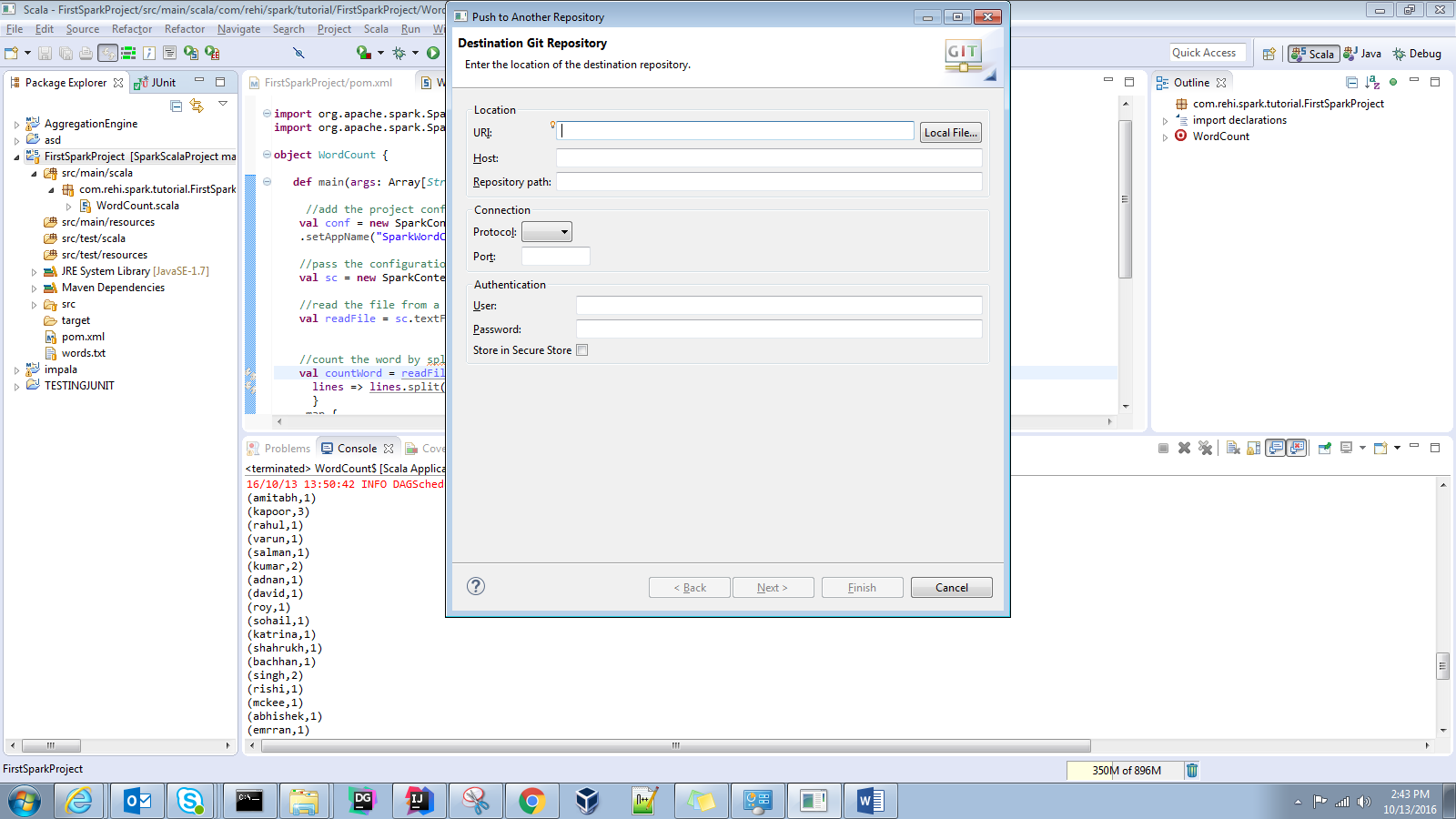
Right click on project name-> Team -> Commit:





Fill in the details:

Select all the files: type the necessary message and click on COMMIT AND PUSH:  


Enter the github repository URL (SSH):  


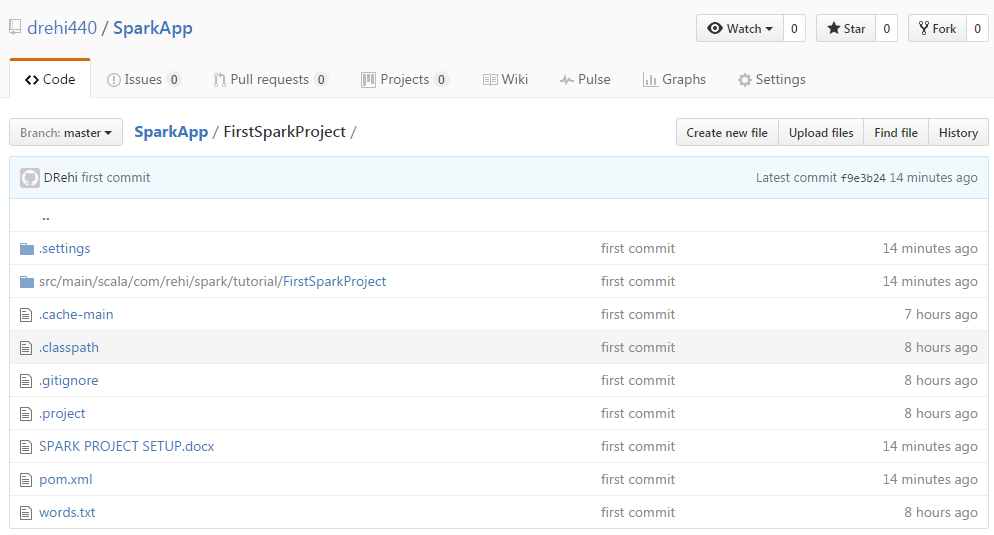
<https://www.youtube.com/watch?v=BH4OqYHoHC0>

this is complete video tutorial to set up to push pull to gitHub

after following this video you are ready to go:

I committed my push to this:

https://github.com/drehi440/SparkApp/tree/master/FirstSparkProject



In Connection: Protocol select ssh:

Now if you get the auth error for repositories refs:

You have to give your gitUsername and Password

And this will work:

You have to download gitBash and create a new gitssh:

<https://help.github.com/articles/generating-a-new-ssh-key-and-adding-it-to-the-ssh-agent/>