

Data Science Academy

www.datascienceacademy.com.br

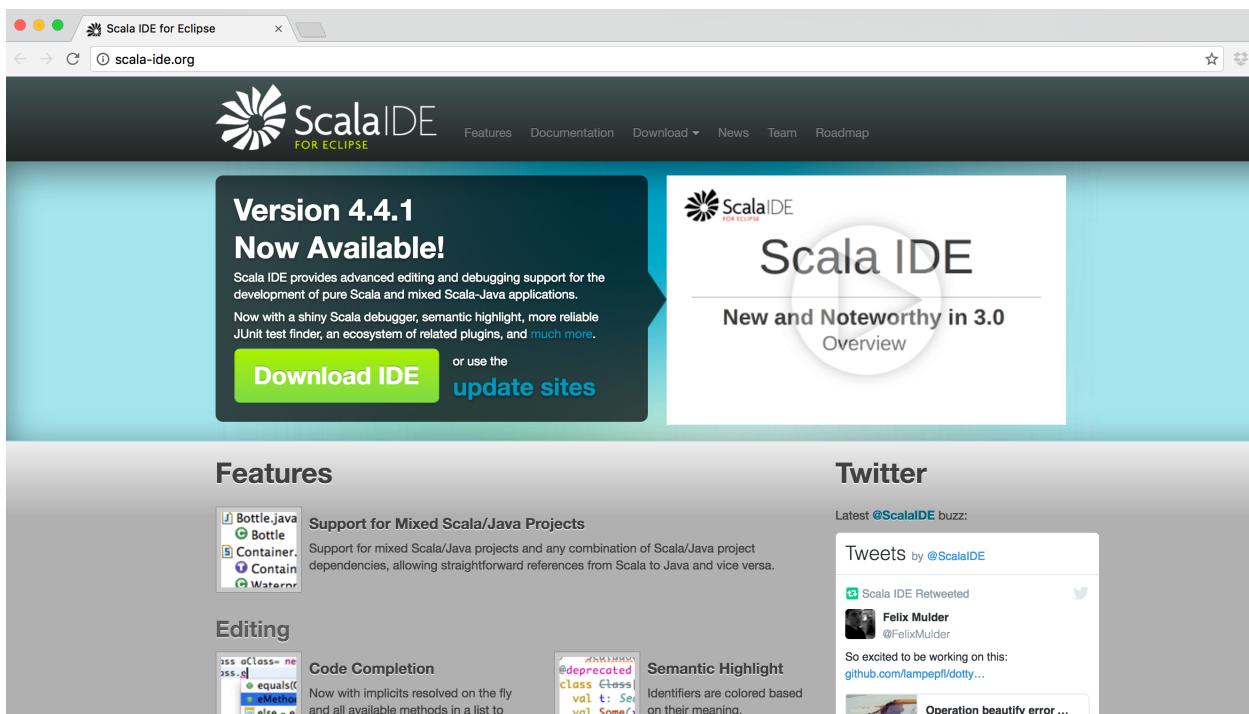
Big Data Real-Time Analytics com
Python e Spark

Bonus
Streaming de Dados do Twitter com Scala e
Spark Streaming

O Spark foi desenvolvido em linguagem Scala e embora ele suporte Java, Python e R, a linguagem Scala é a que oferece o maior poder sobre o framework. Neste bônus, veremos como usar Scala para coletar dados do Twitter em tempo real.

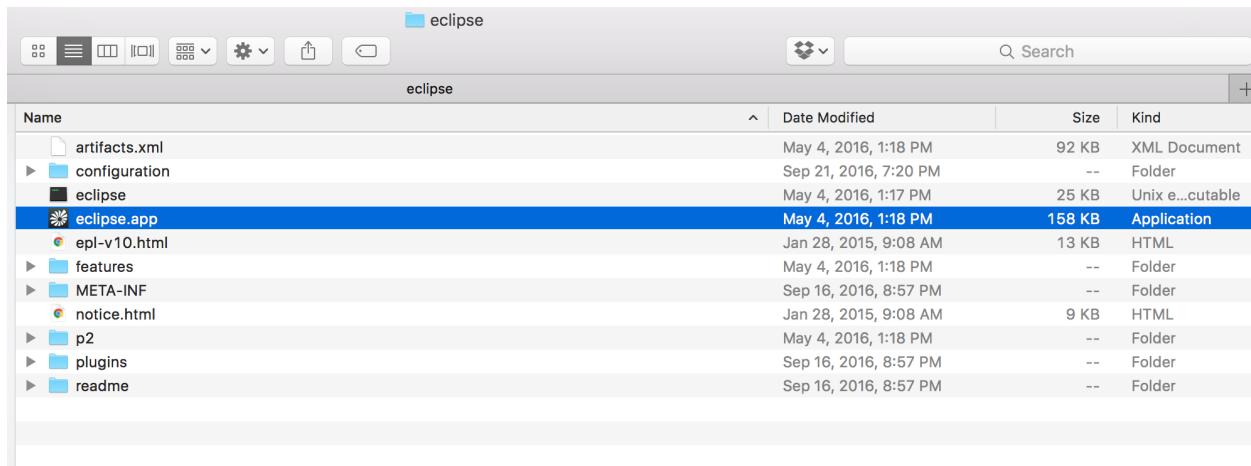
A configuração do seu ambiente requer alguns passos. Siga cuidadosamente o procedimento abaixo. Estamos considerando que o Spark já está devidamente configurado, de acordo com os capítulos anteriores.

1- Baixe a IDE Scala no seu computador, a partir do endereço: <http://scala-ide.org>. IDE é um ambiente de desenvolvimento e usaremos esta IDE para criar os scripts de conexão ao Twitter. O Java deve estar instalado no seu computador (mas como você instalou o Java para usar o Spark, a instalação já deve ter sido feita). Abra um prompt de comando e digite: **java –version** para confirmar que o Java está instalado.

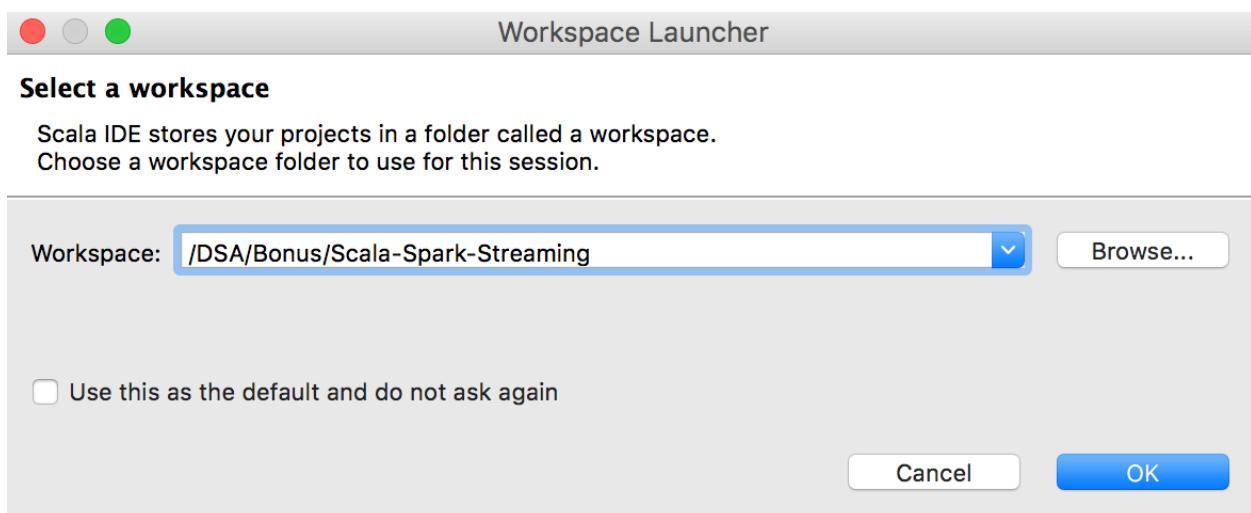


2- Ao concluir o download, descompacte o arquivo em um diretório na sua máquina, por exemplo c:\scala.

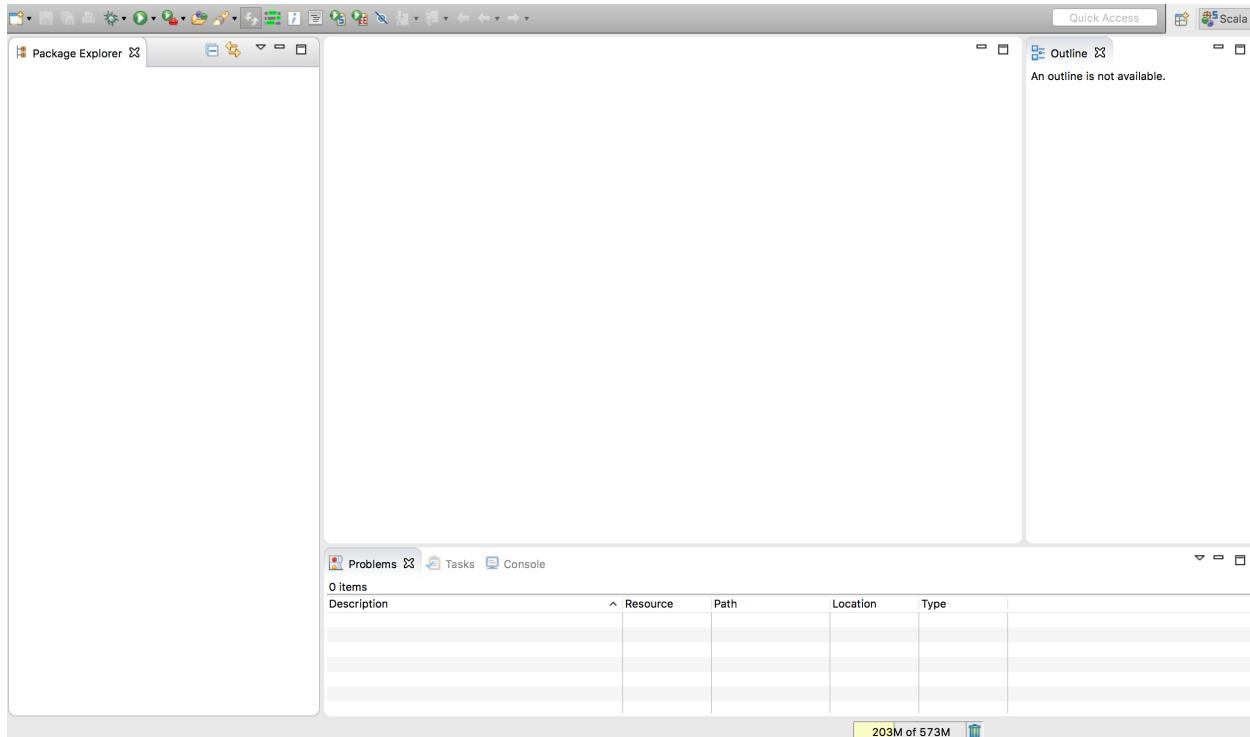
3- Navegue até o diretório e execute o Eclipse. O IDE Scala é na verdade um plugin para o Eclipse.



4- O Eclipse vai solicitar o diretório da workspace (onde ficará seu projeto). Selecione um diretório em sua máquina.



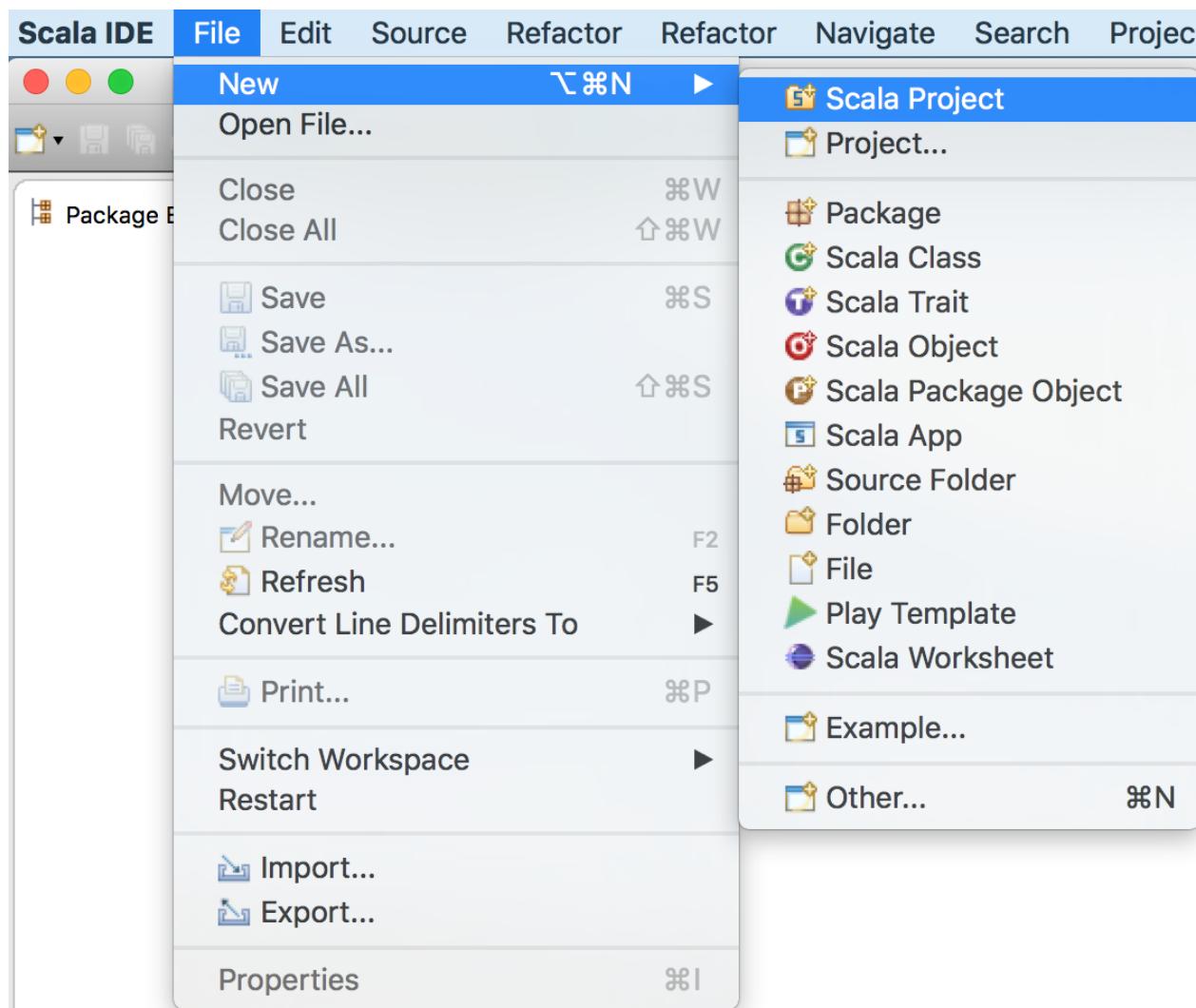
5- Você estará na IDE Scala:



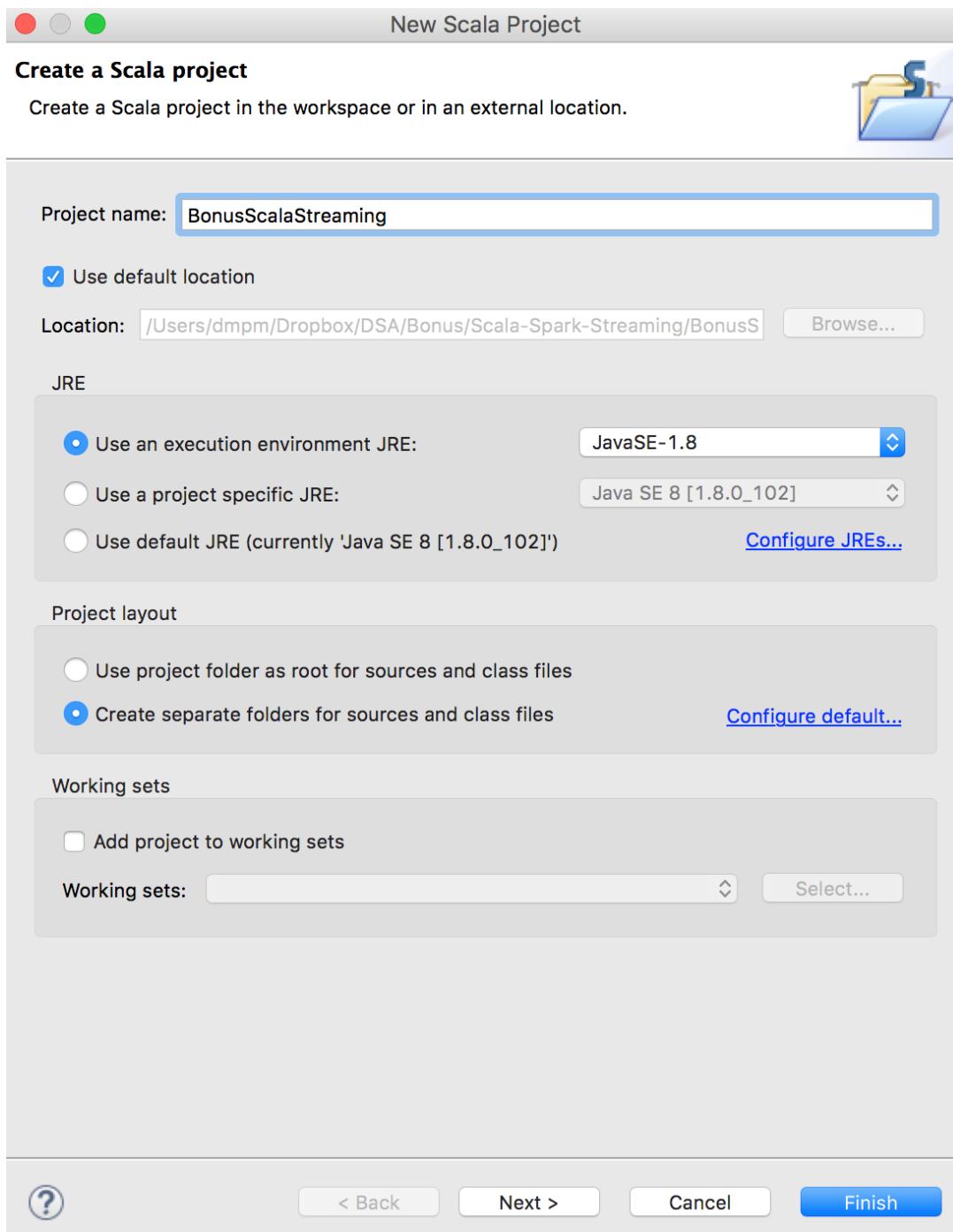
6- Retorne para seu computador e navegue até o diretório do workspace (que você definiu no item 4). Dentro desta pasta, crie um arquivo chamado **twitter.txt**. Esse arquivo deve conter as chaves e tokens da sua app do Twitter. Seu arquivo ficará assim (você deve usar as chaves da sua app do Twitter):

```
consumerKey jrcF2uMSMw5v0XAZzPb25foQB
consumerSecret uQT4iWnVyNxqVUFlsPrnceywAZXCUJwSzzQdDM5ywmXScnEmLI
access_token 703383646602981377-Ktxgd5k8yxjTsSywpnTaig0xT8pZqaG
access_secret EGw91XKccy1DPdWDL1LR1eAZ1dIPKd4n6mZiBbXcWGZnF
```

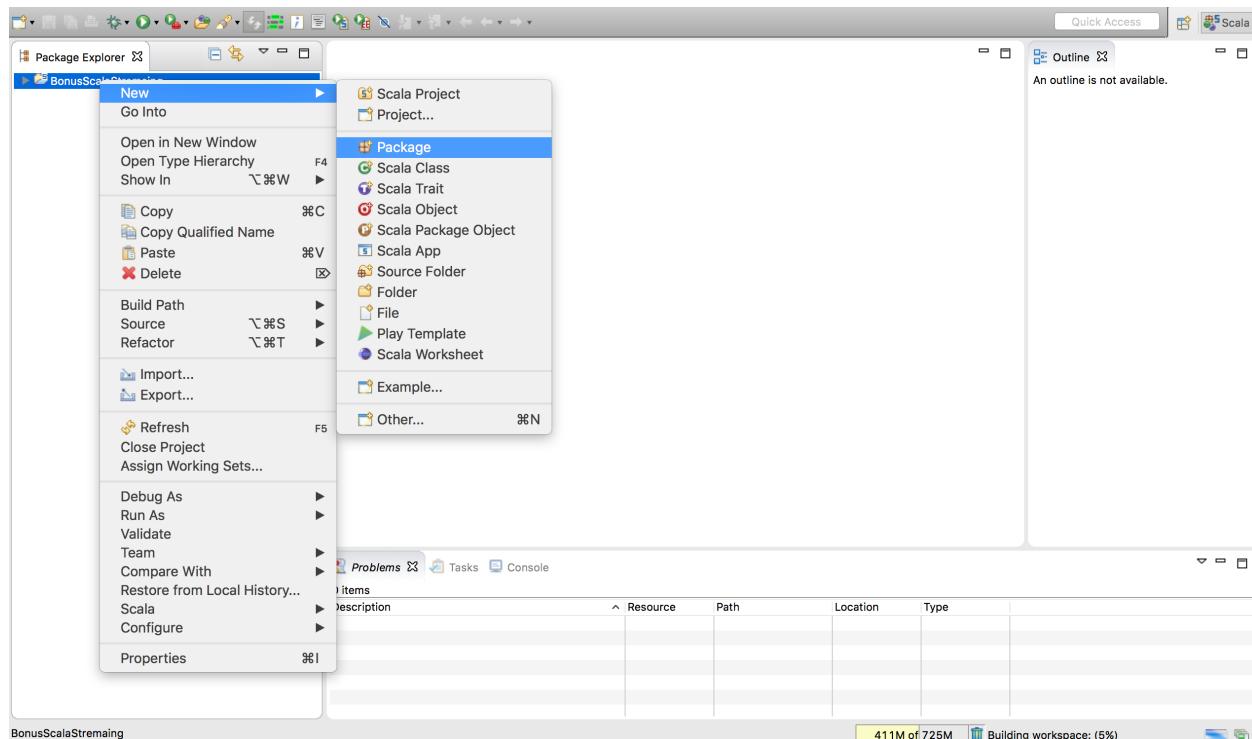
7- De volta ao IDE Scala, clique no menu New – Scala Project.



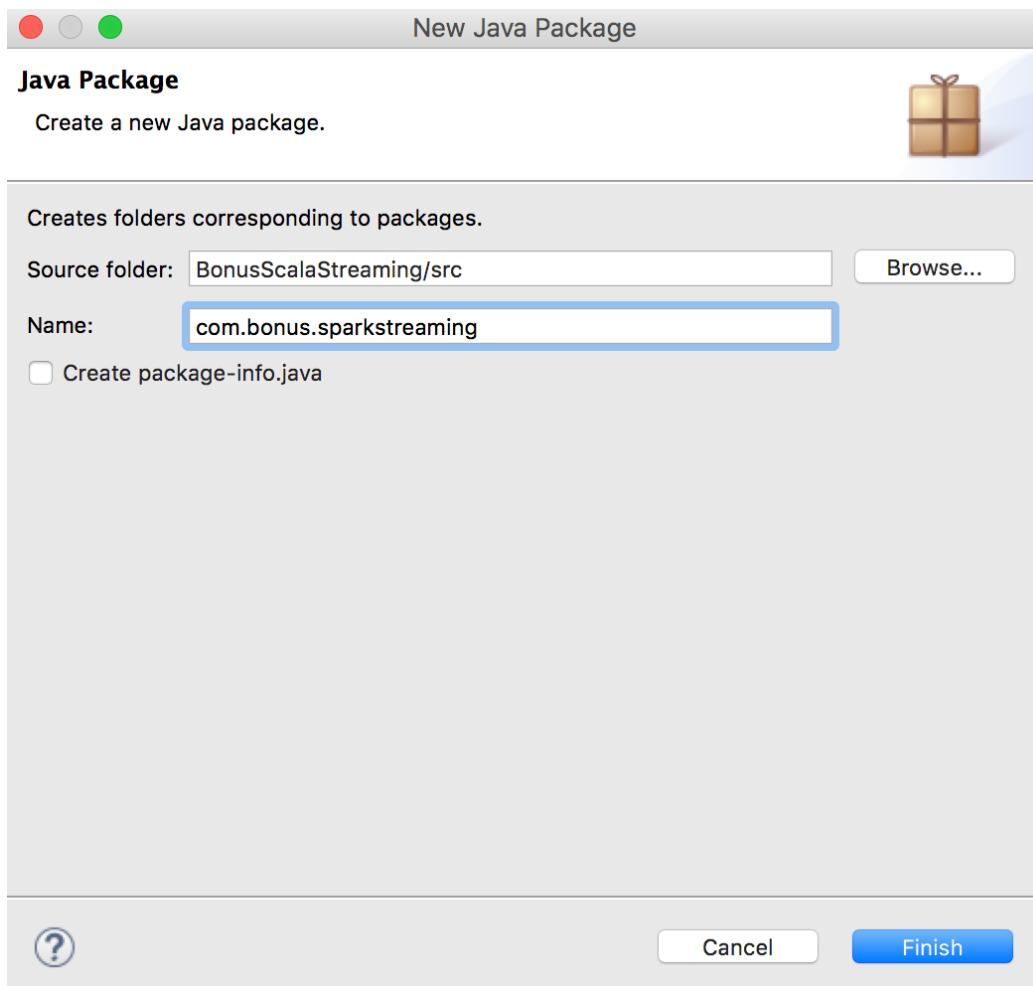
8- Defina o nome do seu projeto (pode ser o nome que você quiser) e clique no botão **Finish**:



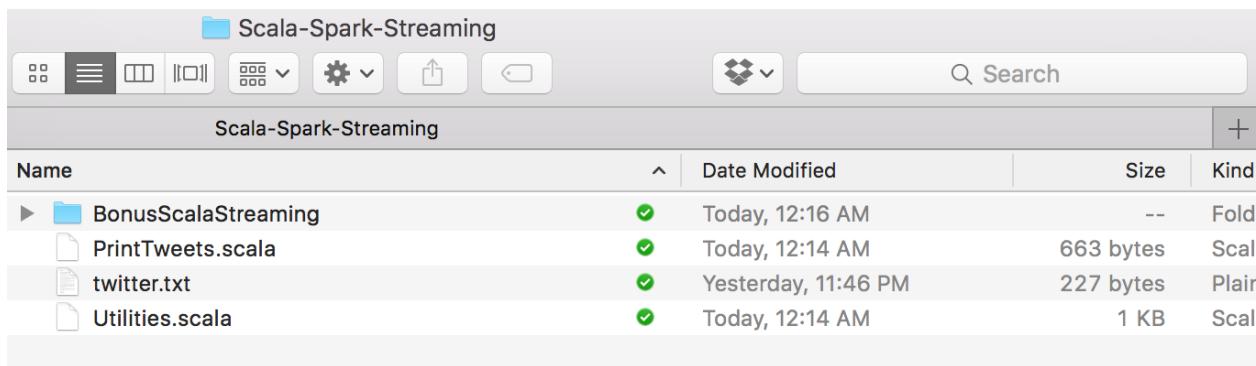
9- Clique no projeto que você criou com o botão direito do mouse e selecione a opção para criar um pacote (New – Package):



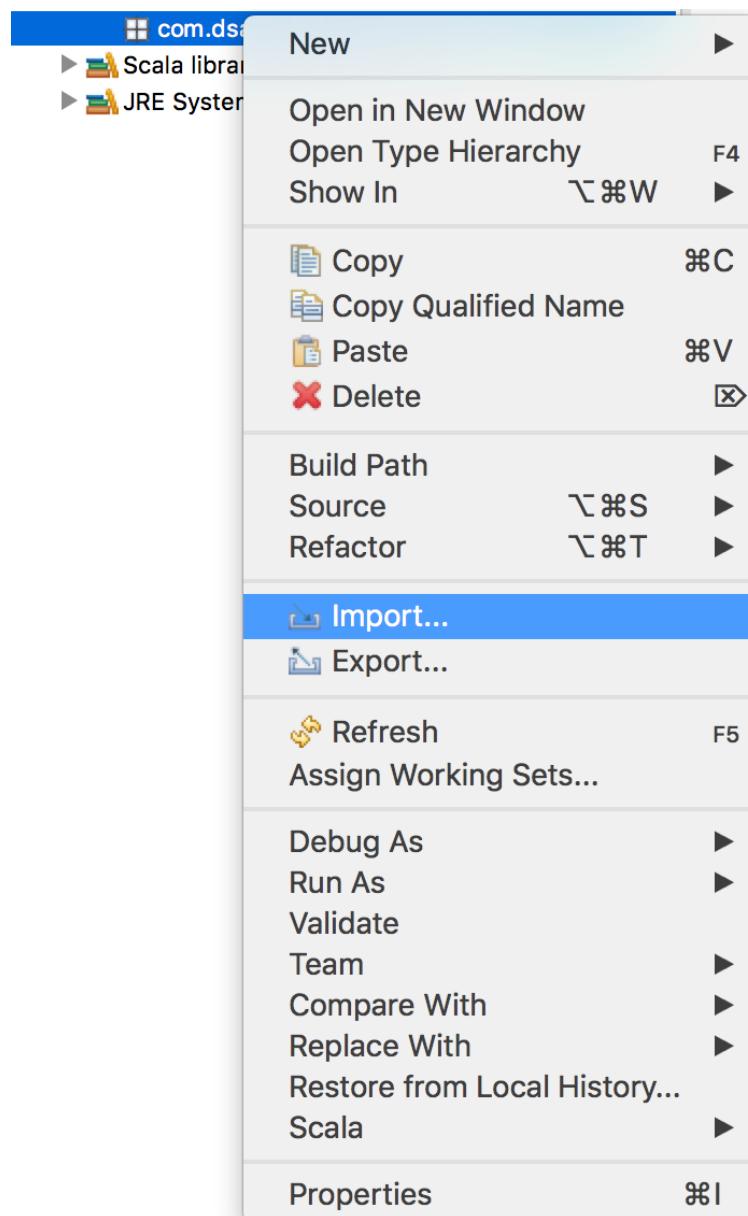
- 10- Digite o nome do pacote que será criado. Embora apareça “Java Package”, estamos criando um pacote Scala. A linguagem Scala é executada sobre uma máquina virtual Java. Clique no botão **Finish**.



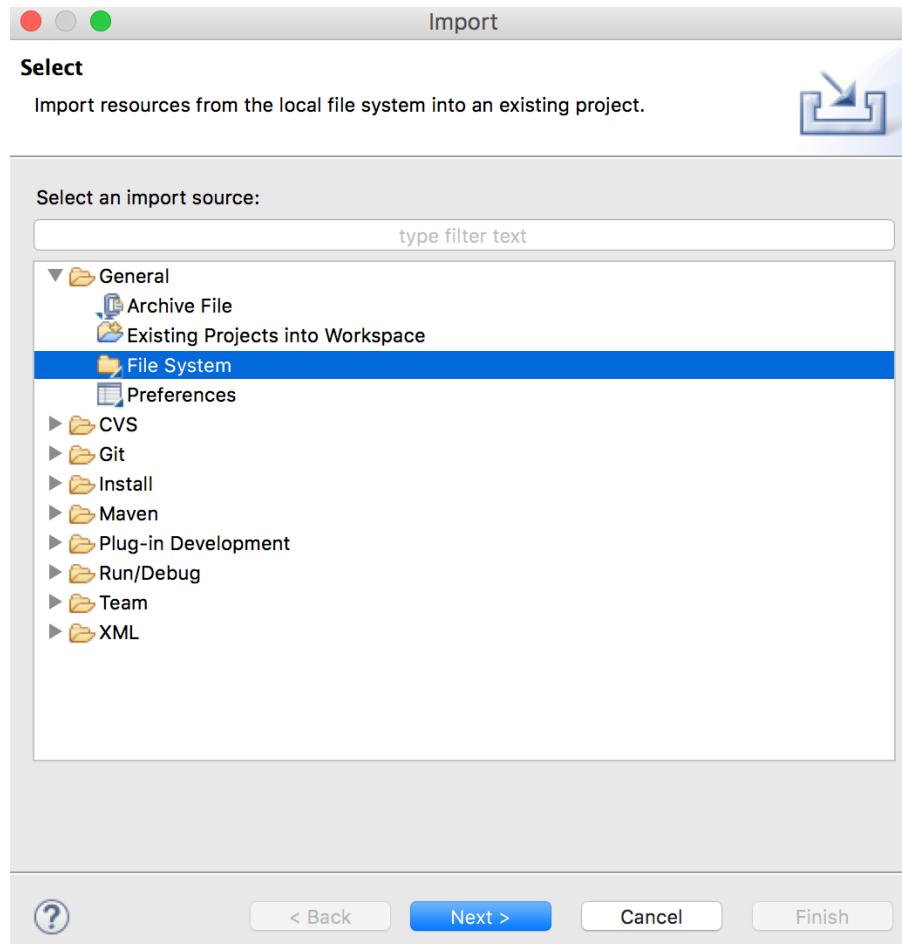
11- Junto com os demais arquivos deste bônus, você recebeu 2 scripts com extensão .scala: **PrintTweets.scala** e **Utilities.scala**. Copie os 2 arquivos para o diretório da sua workspace.



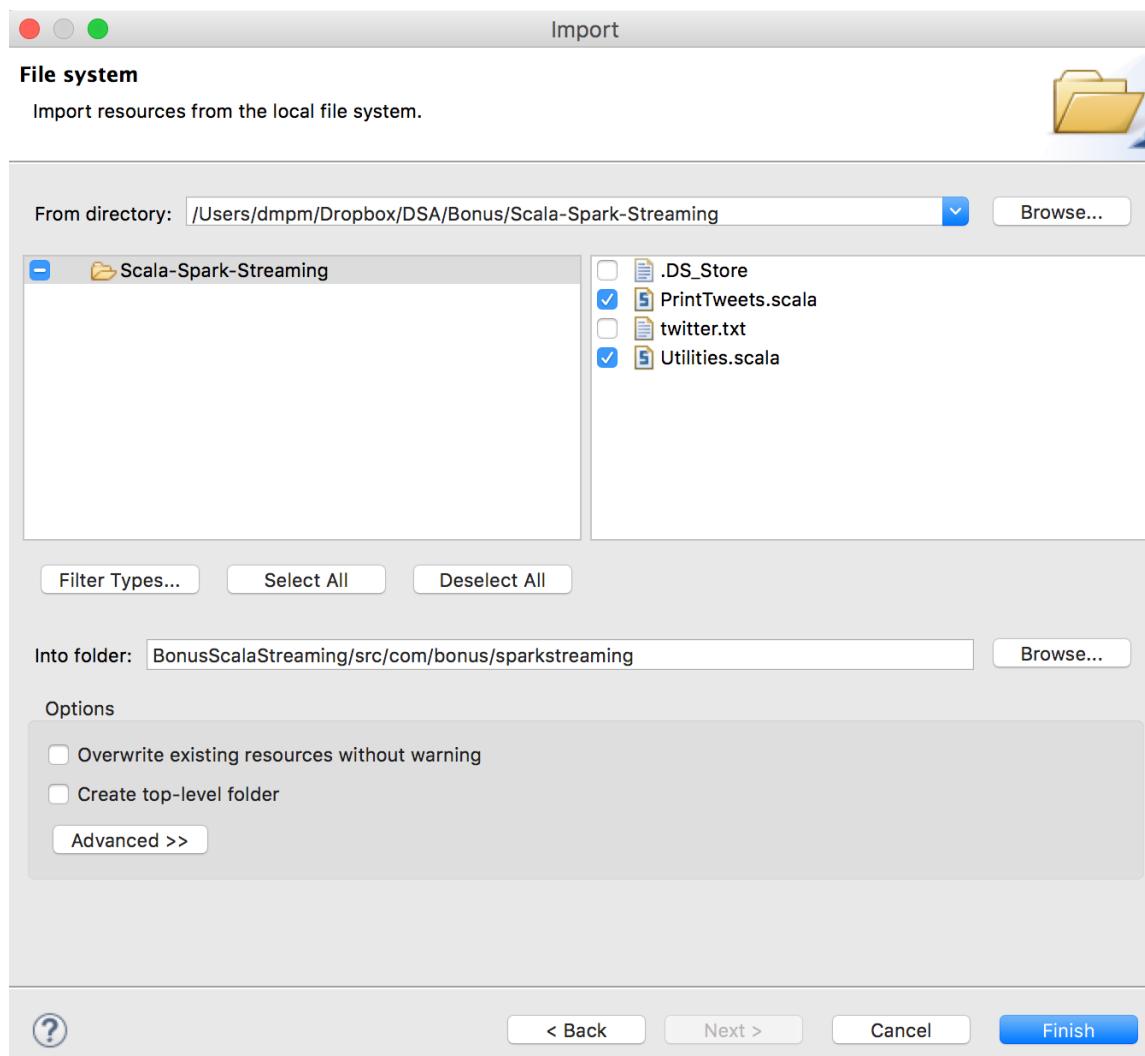
12- No IDE Scala, clique com o botão direito do mouse no pacote que você criou e selecione a opção Import.



13- Selecione os 2 scripts .scala que você copiou no diretório da workspace.

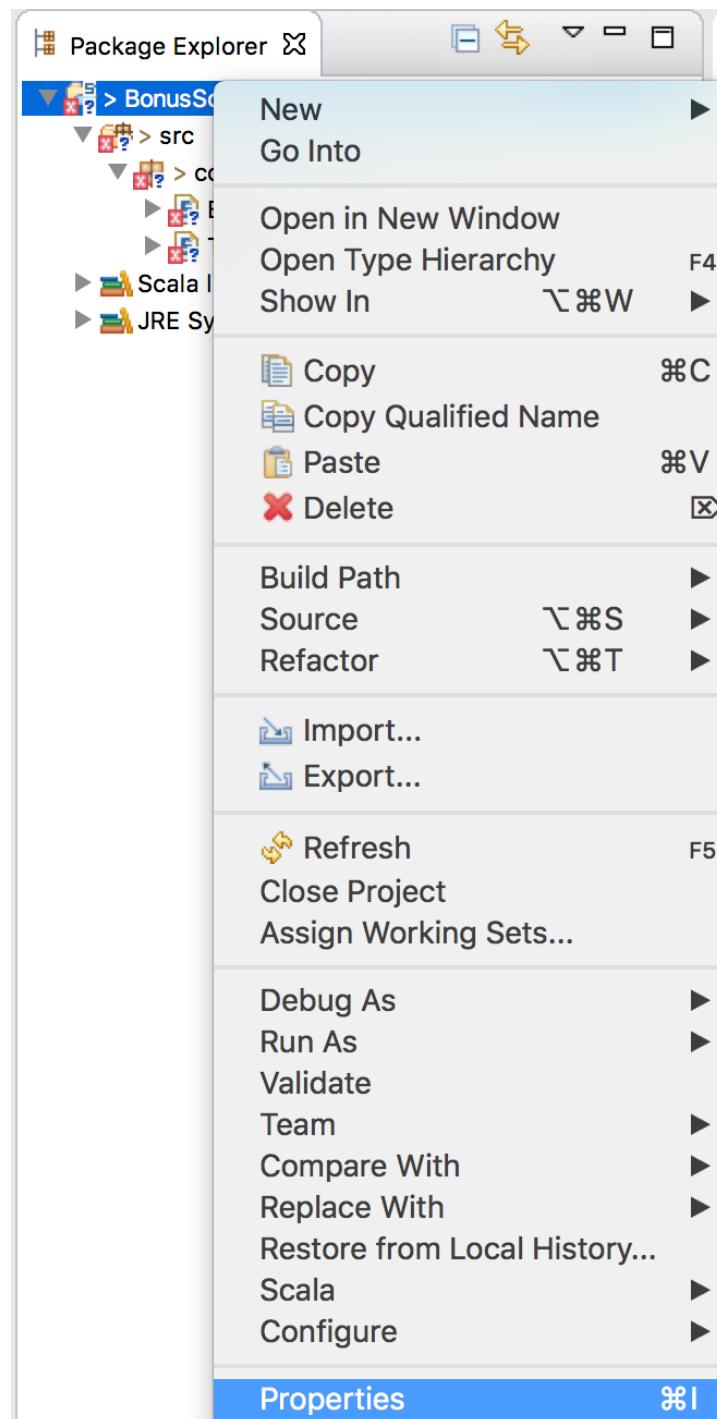


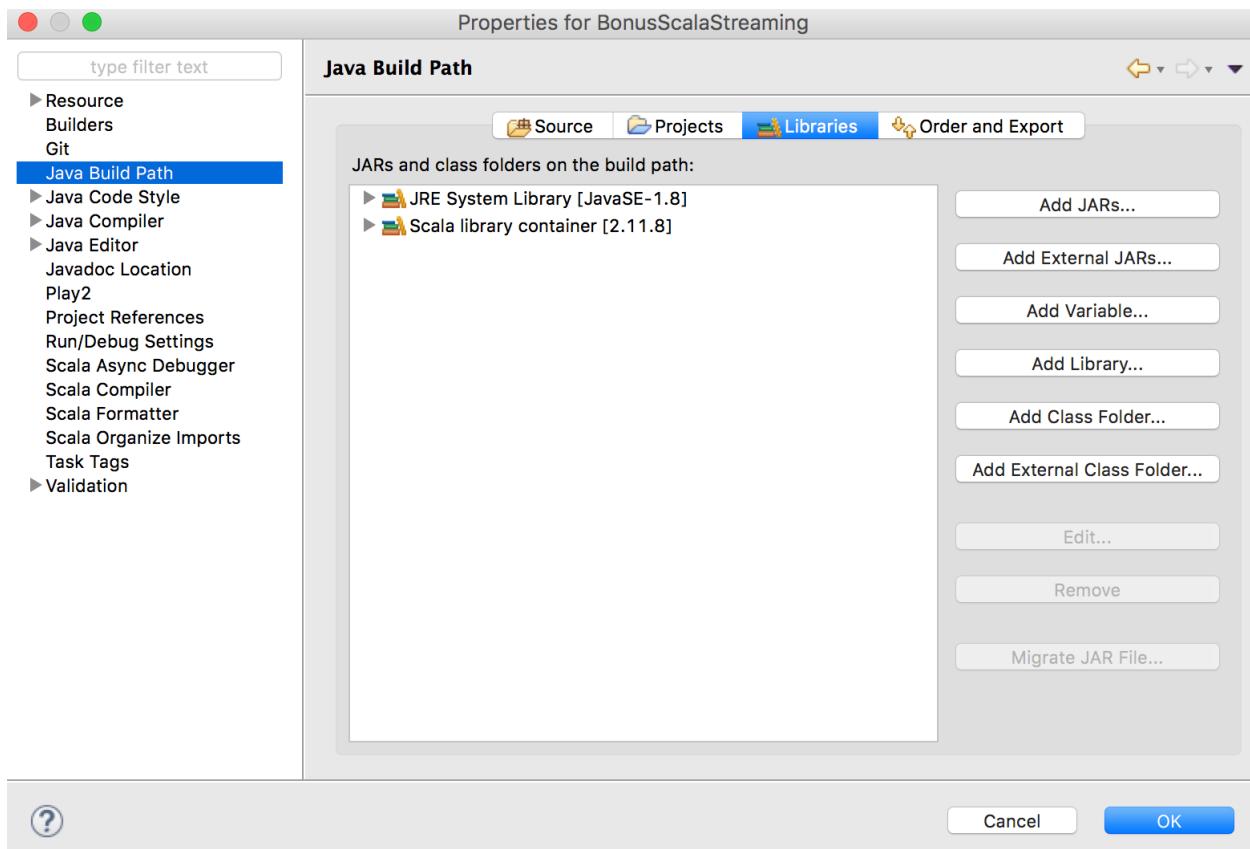
Selecione File System e clique Next



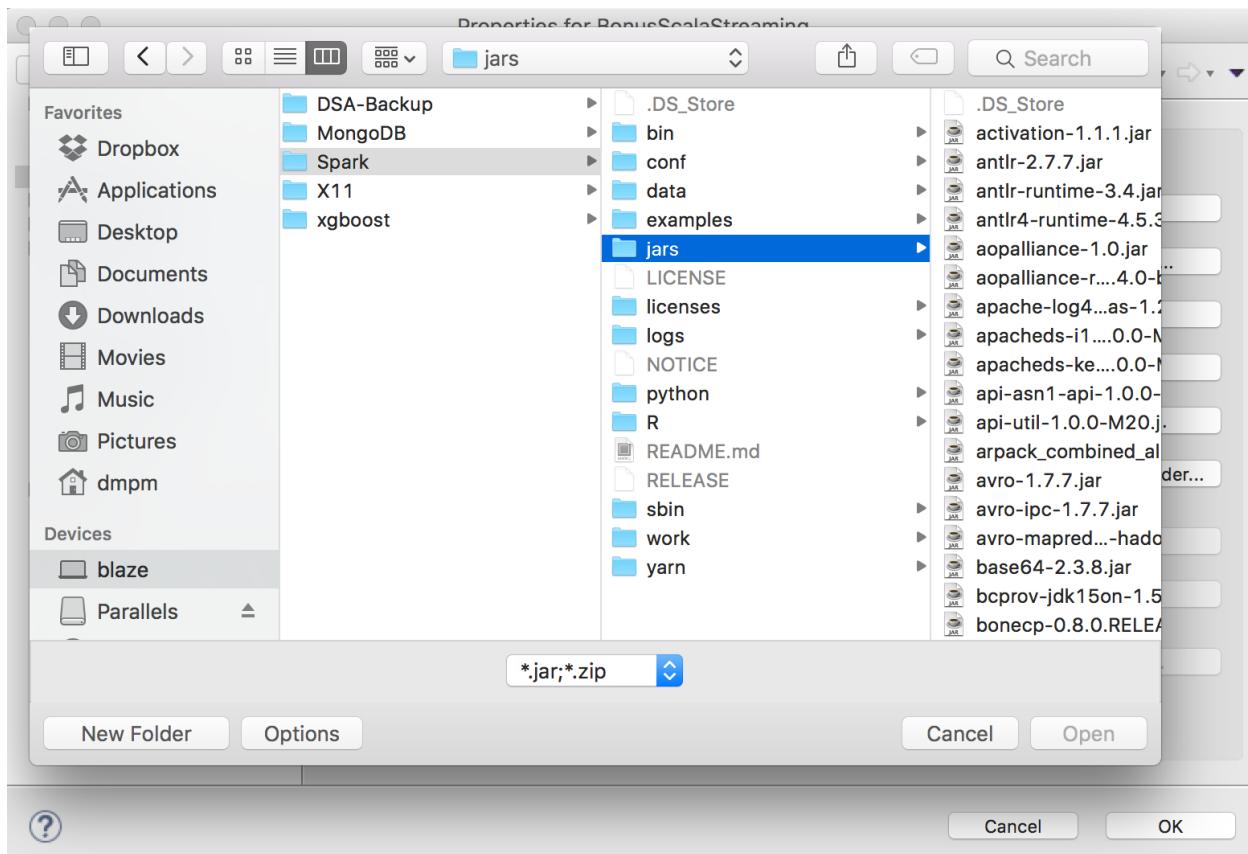
Selecione os 2 scripts .scala e clique em Finish

14- Clique com o botão direito do mouse no seu projeto e selecione **Properties**.

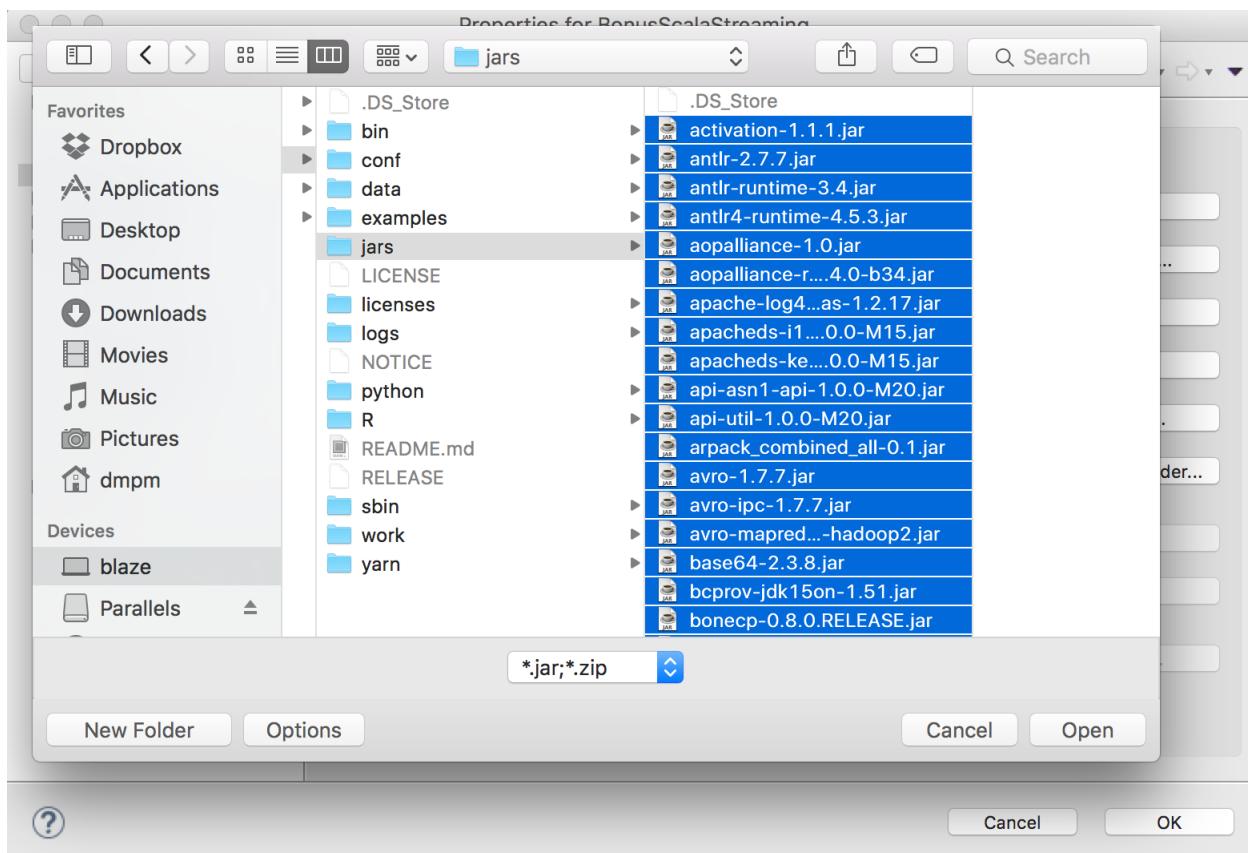




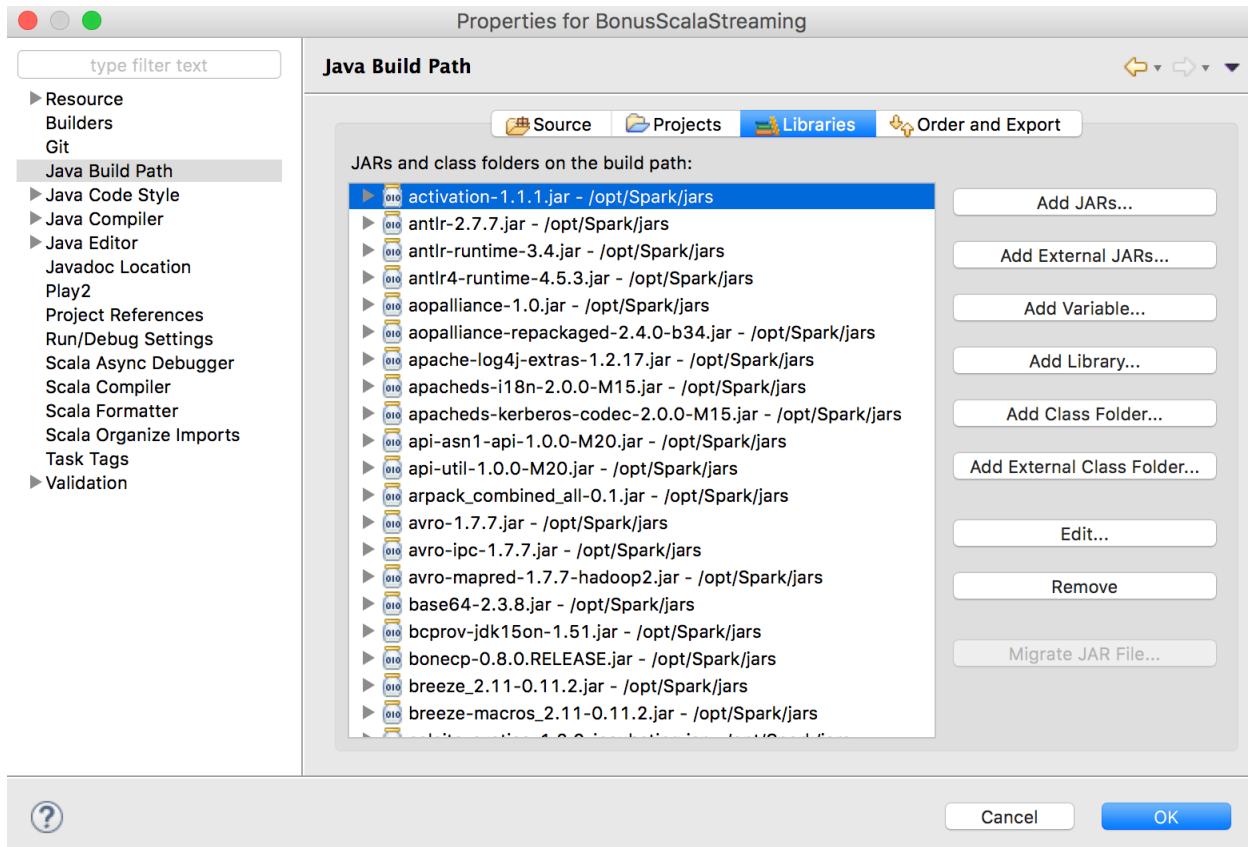
Selecione Java Build Path do lado esquerdo e depois a aba Libraries. No lado direito, clique no botão Add External JAR's



Selecione o diretório jars, dentro do diretório de instalação do Spark



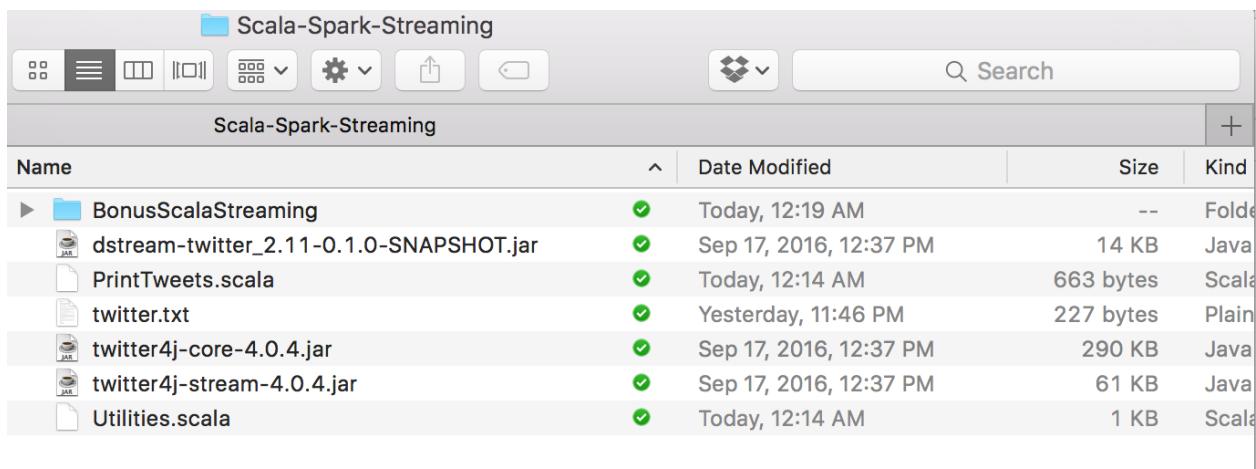
Selecione todos os arquivos .jar (use as teclas de atalho Ctrl + A) e clique open.



Clique Ok.

Calma. Ainda não acabou!!!! Lembre, o Cientista de Dados precisa saber muito, sobre muitas coisas!!

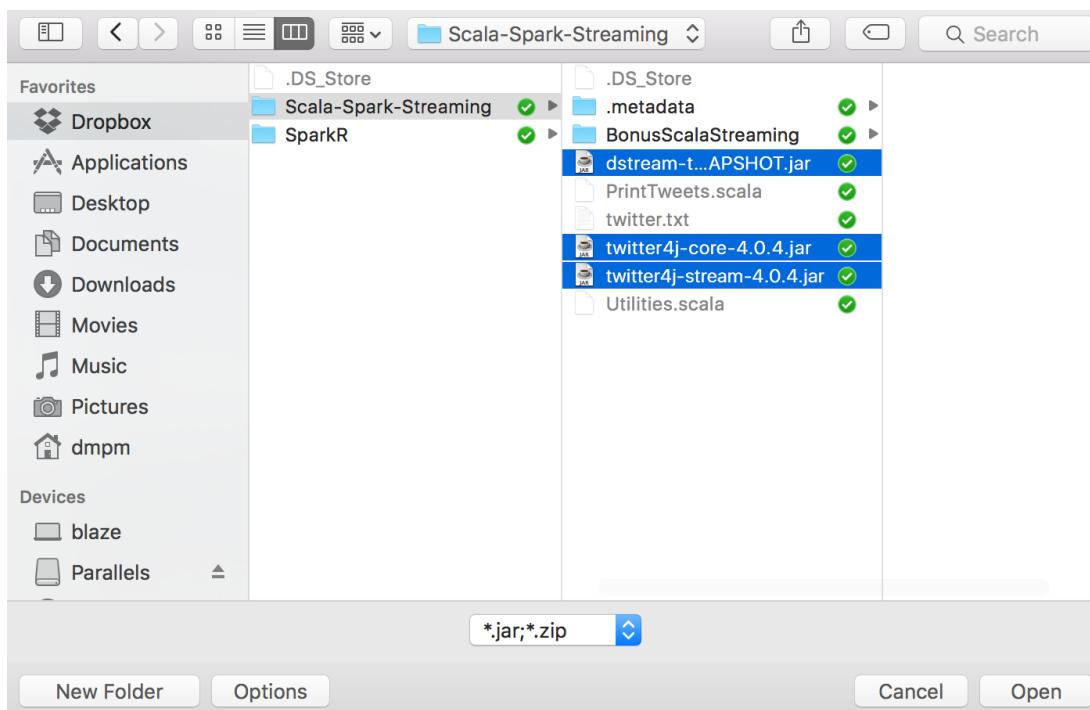
15- Junto com os demais arquivos deste bônus, você recebeu 3 arquivos .jar. Copie os 3 arquivos para o diretório da sua workspace.



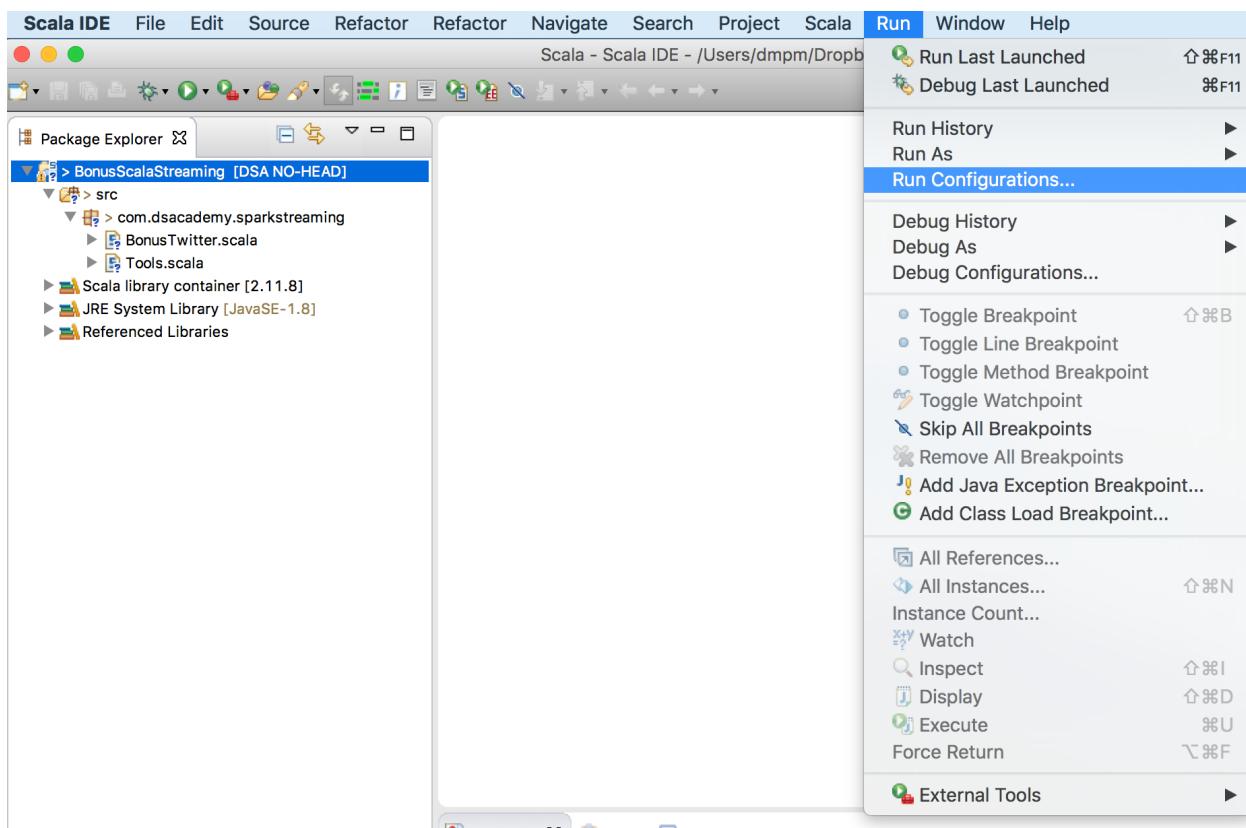
The screenshot shows a file browser window titled "Scala-Spark-Streaming". The directory structure is as follows:

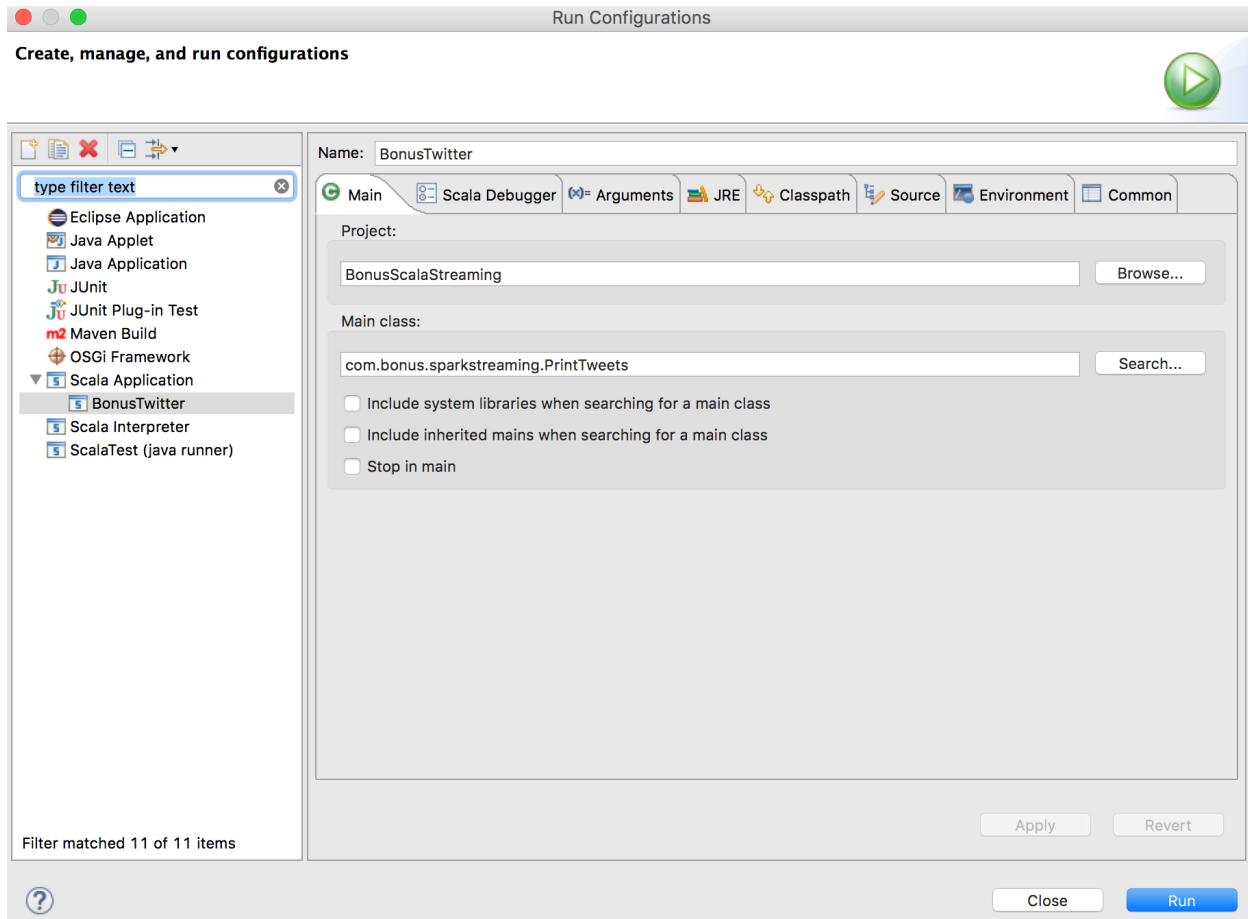
Name	Date Modified	Size	Kind
BonusScalaStreaming	Today, 12:19 AM	--	Folder
dstream-twitter_2.11-0.1.0-APSHOT.jar	Sep 17, 2016, 12:37 PM	14 KB	Java
PrintTweets.scala	Today, 12:14 AM	663 bytes	Scala
twitter.txt	Yesterday, 11:46 PM	227 bytes	Plain
twitter4j-core-4.0.4.jar	Sep 17, 2016, 12:37 PM	290 KB	Java
twitter4j-stream-4.0.4.jar	Sep 17, 2016, 12:37 PM	61 KB	Java
Utilities.scala	Today, 12:14 AM	1 KB	Scala

16- Repita os passos do item 14 e adicione esses 3 arquivos .jar ao seu projeto.



17- Estamos prontos para coletar tweets. Clique no menu **Run** e então **Run configurations**.





Dê um duplo clique em Scala Application no lado esquerdo. Dê um nome para sua aplicação e configure a Main Class.

Clique no botão Run e cole os tweets com Scala e Spark Streaming. Os tweets serão mostrados no console da IDE Scala.