Local Spark Instance

Download and decompress latest Spark binaries

Got to http://spark.apache.org/downloads.html and download the latest version of Spark binaries (click the link on item 4 to download on your local machine)

It will be a tgz file (mac should be able to open it automatically, window users will need something like free archiver: http://www.7-zip.org and un-compress the TAR file as well)

Download latest spark-ec2 from Github

Go to: https://github.com/amplab/spark-ec2
Download latest branch and unzip it on your local machine and add it to your main Spark binaries folder.

JAVA

Make sure you have Java installed on your machine (search for Install latest Java SE Downloads)

My MAC starting code (change paths to yours)

```
.libPaths(c(.libPaths(), '/Users/manuelamunategui/Downloads/spark-2.0.0-bin-hadoop2.7/R/lib'))
Sys.setenv(SPARK_HOME = '/Users/manuelamunategui/Downloads/spark-2.0.0-bin-hadoop2.7')
Sys.setenv(PATH = paste(Sys.getenv(c('PATH')),
       '/Users/manuelamunategui/Downloads/spark-2.0.0-bin-hadoop2.7/bin', sep=':'))

library(SparkR, lib.loc = c(file.path(Sys.getenv("SPARK_HOME"), "R", "lib")))
sparkR.session(master = "local[*]", enableHiveSupport = FALSE, sparkConfig =
list(spark.driver.memory = "1g",
spark.sql.warehouse.dir="/Users/manuelamunategui/Downloads/spark-2.0.0-bin-hadoop2.7/"))
```

My Windows starting code (change paths to yours)

```
.libPaths(c(.libPaths(), 'C:\\Users\\manuel\\Downloads\\spark-2.0.2-bin-hadoop2.7\\R\\lib'))
Sys.setenv(SPARK_HOME = 'C:\\Users\\manuel\\Downloads\\spark-2.0.2-bin-hadoop2.7')
Sys.setenv(PATH = paste(Sys.getenv(c('PATH')), 'C:\\Users\\manuel\\Downloads\\spark-2.0.2-bin-hadoop2.7\\bin', sep=':'))

library(SparkR, lib.loc = c(file.path(Sys.getenv("SPARK_HOME"), "R", "lib")))
sparkR.session(master = "local[*]", enableHiveSupport = FALSE, sparkConfig = list(spark.driver.memory = "1g", spark.sql.warehouse.dir="c:\\tmp\\"))
```

Some SparkR sample code

```
# now let's run thorugh a few local examples to confirm that we are working in Spark
df_spark<- as.DataFrame(faithful)
class(df_spark)
head(df_spark)

head(select(df_spark, df_spark$eruptions))
head(filter(df_spark, df_spark$waiting < 50))

# Grouping, Aggregation
head(summarize(groupBy(df_spark, df_spark$waiting), count = n(df_spark$waiting)))

# Operating on Columns
df_spark$waiting_secs <- df_spark$waiting * 60


# stop your session when finished
sparkR.session.stop()
```