Twitter Streaming 1

Set the packages. Vital information was hidden as well as warning and output.

library(twitteR)  
  
setup\_twitter\_oauth(key,secret,access,access\_secret)  
  
Sys.setenv(JAVA\_HOME="/usr/lib/jvm/default-java")  
Sys.setenv(HADOOP\_CMD="/home/bcarancibia/workspace/cuny\_msda\_is622/hadoop-2.7.1/bin/hadoop")  
Sys.setenv(HADOOP\_STREAMING="/home/bcarancibia/workspace/cuny\_msda\_is622/hadoop-2.7.1/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar")  
  
Sys.setenv(SPARK\_HOME = "/home/bcarancibia/workspace/cuny\_msda\_is622/spark-1.4.1-bin-hadoop2.6")  
.libPaths(c(file.path(Sys.getenv("SPARK\_HOME"), "R", "lib"), .libPaths()))  
library(SparkR)  
  
sc <- sparkR.init(master="local")  
sqlContext <- sparkRSQL.init(sc)

I am going to look at Twitter data, specifically looking at the hashtag #datarevolution. This is an important hashtag in the international development space because of the recent increase in desire for countries, companies, and aid organizations to integrate analytics into their everyday workflows. I am going to collect hashtags and then count the top ten screen names.

tweets <- searchTwitter("#datarevolution",n=9999)

## Warning in doRppAPICall("search/tweets", n, params = params,  
## retryOnRateLimit = retryOnRateLimit, : 9999 tweets were requested but the  
## API can only return 1648

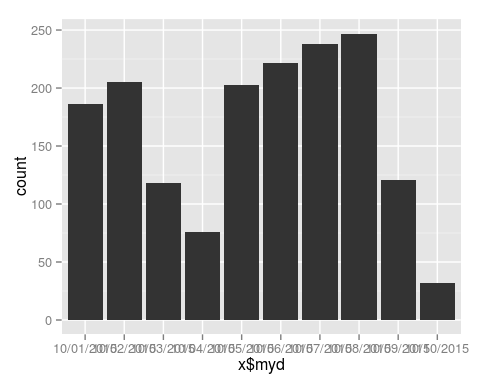
x <- twListToDF(tweets)  
  
sparkdf <- createDataFrame(sqlContext, x)  
  
group <- agg(group\_by(sparkdf, sparkdf$screenName), sum\_of\_screenname=(count(sparkdf$screenName)))  
head(group)

## screenName sum\_of\_screenname  
## 1 blondelena 7  
## 2 fpgil 1  
## 3 keyram10 3  
## 4 SadiQBichi 1  
## 5 alabriqu 1  
## 6 OpenDataService 2

library(ggplot2)  
library(lubridate)  
library(forecast)

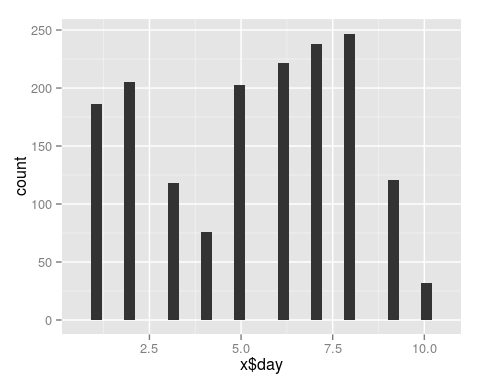
## Loading required package: zoo  
##   
## Attaching package: 'zoo'  
##   
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric  
##   
## Loading required package: timeDate  
## This is forecast 6.1

x$created <- parse\_date\_time(x$created, "%Y%m%d %H%M%S", truncated = 3)  
  
x$day <- day(x$created)  
x$month <- month(x$created)  
x$year <- year(x$created)  
x$hour <- hour(x$created)  
x$minute <- minute(x$created)  
x$time <- sprintf('%02d:%02d', x$hour, x$minute)  
x$myd <- sprintf('%02d/%02d/%02d', x$month, x$day, x$year)  
  
#Tweets by Month, Day, Year  
qplot(x$myd, data = x, geom="histogram")



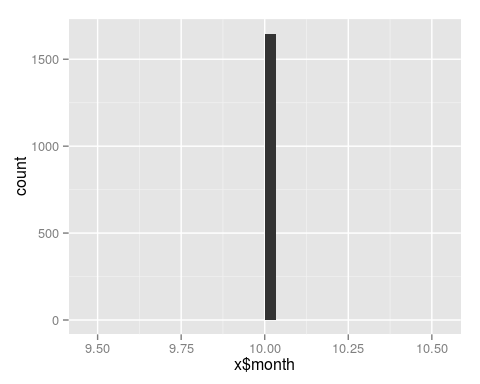
qplot(x$day, data=x, geom="histogram")

## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.



qplot(x$month, data=x, geom="histogram")

## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.



One thing to notice is that the R package to scrape tweets, only seems to take into account the past week or so.

Group by Screen and sum retweets.

group <- agg(group\_by(sparkdf, sparkdf$screenName), sum\_of\_retweets=(sum(sparkdf$retweetCount)))  
  
head(group,10)

## screenName sum\_of\_retweets  
## 1 blondelena 134  
## 2 fpgil 7  
## 3 keyram10 14  
## 4 SadiQBichi 25  
## 5 alabriqu 6  
## 6 OpenDataService 21  
## 7 bracken10011 10  
## 8 Cath\_Cand 3  
## 9 EvarMburu 1  
## 10 writeosahon 4