Clustering

Ben Arancibia

10/22/2015

This project will focus on taking Twitter data and performing a cluster analysis on the data. The first thing to do is setup the appropriate environment using the following.

Thing that are being setup. Twitter Package, Hadoop, and Spark just in case it's needed.

```
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/to

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 search Twitter data stream and get the hashtags of #datarevolution and #SDGs. These are useful to my field of work in international development. The main point of these hashtags are to disseminate information related to data analytics and the push for data analysis in the international development field.

```
tweets1 <- searchTwitter("#datarevolution",n=9999)
```

```
## Warning in doRppAPICall("search/tweets", n, params = params,
## retryOnRateLimit = retryOnRateLimit, : 9999 tweets were requested but the
## API can only return 1072
```

```
tweets2 <- searchTwitter("#SDGs",n=9999)

x <- twListToDF(tweets1)
x1 <- twListToDF(tweets2)

x3 <- rbind(x,x1)

n <- nrow(x3)
```

The first step is filter out data that I want to cluster on using kmeans clustering. First I am going to select the variables that I am interested in clustering. I will also determine the number of clusters by looking at the elbow graph of clusters.

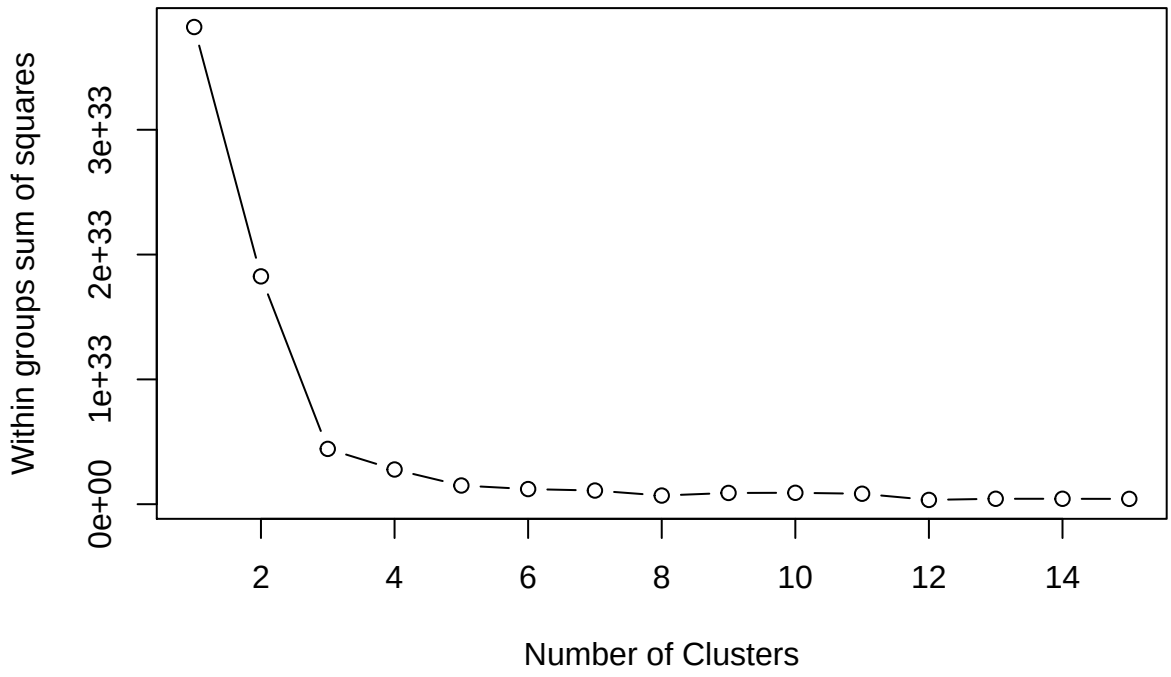
```
vars <- c("id", "favoriteCount", "retweetCount")

mydata <- x3[vars]
```

```
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata,
    centers=i)$withinss)
```

```
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 553550)
```

```
plot(1:15, wss, type="b", xlab="Number of Clusters",
  ylab="Within groups sum of squares")
```



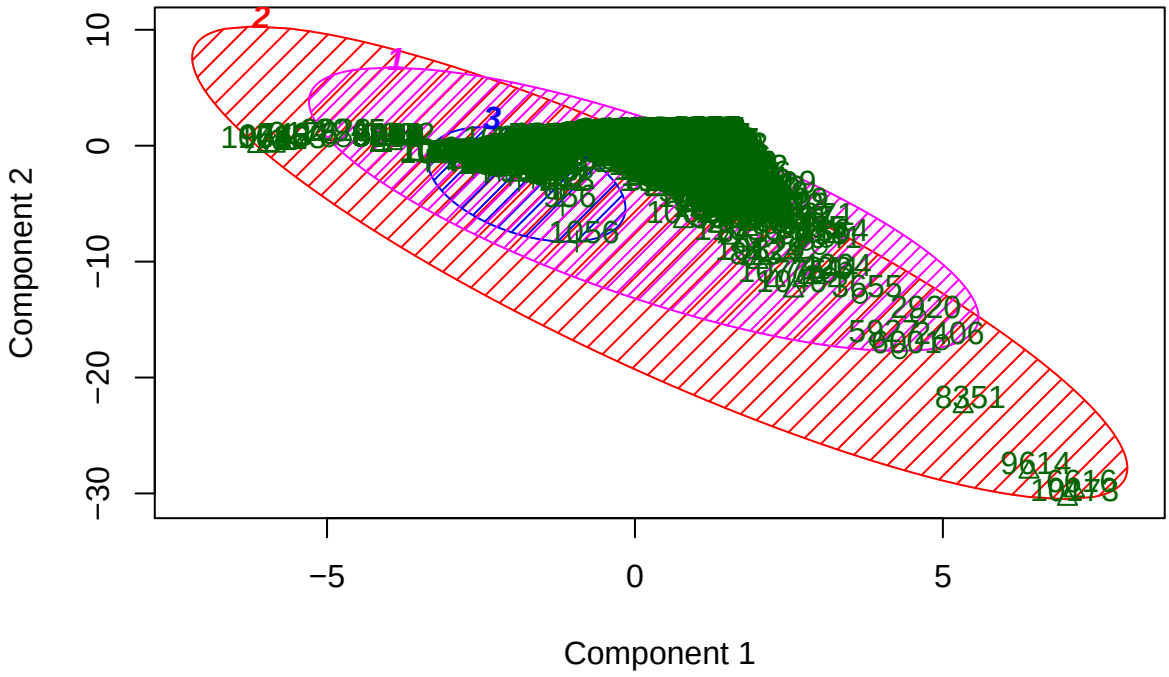
From this graphic it appears that 3 or four is the appropriate amount of clusters. Three will be used because it has the most change in slope.

Start Clustering.

```
library(cluster)
library(fpc)

clus <- kmeans(mydata, centers=3)

clusplot(mydata, clus$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```


CLUSPLOT(mydata)
Component 2
10
0
−10
−20
−30
−5
0
5
Component 1
These two components explain 68.91 % of the point variability.