Tarea#1 - Ciencia de los Datos

Leonardo Santella. Simon Saman. Eloy Toro

June 9, 2015

Introduccion

El objetivo de esta tarea es manipular el API de Twitter para extraer datos de una cuenta de usuario. En este documento no se ve reflejado los paso para la creacion de una aplicacion Twitter para generar las credenciales necesarias para extraer los datos.

Luego de extraer los datos, en principio, seran colocados en una estructura y seran eliminadas las caracteristicas poco influyentes para el analisis. El analisis estara basado en los resultados de originados a traves del algoritmo de clusterizacion (clustering) K-Medias (K-Means) con la eleccion previa de numero K (K es el numero de clusters) dependiente de la informacion generada por la tecnica del codo (Codo de Jambu).

Objetivos

Antes de empezar con los objetivos de la tarea, se deberan cargar las librerias respectivas para la ejecucion apropiada de las funciones de R. Es importante resaltar que en principio, varios de los siguientes paquetes deben ser instalados

```
library(twitteR)
library(RCurl)
```

```
## Loading required package: bitops
```

```
library(RJSONIO)
library(stringr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:twitteR':
##
##     id, location
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(devtools)
library(rCharts)
library(knitr)
```

1.Extraiga el listado de usuarios vinculados con su cuenta de Twitter(si la tiene, sino cree una), almacenando en un data frame tanto los usuarios que usted sigue (friends) como los que lo siguen a usted (followers). Se deben diferenciar ambos conjuntos con una nueva columna que contenga uno de los siguientes valores: 1 o 2 (1:friend,2:follower).

Para la obtenion de los datos, debemos autenticarnos. Asignamos a cada variable el string correspondiente al parametro requerido por la funcion setup_twitter_oauth(...) que es la encargada de procesar la autenticacion

```
api_key <- "8TduouVcjWi5YDnS2Z6SZxSnN"
api_secret <- "1YszOPvqOohFlBxBnFf2gz3zjDVScqiGuUOJM2zhWZvGYhYau2"
access_token <- "221541064-OmMDGnuFoYnqcVWufPfsJIucU8rwR5AC6vYi2xGB"
access_token_secret <- "f3bXhy1r1ZHfw7wahYFKnU3R5GiIQzZ78UaXeYcqfajsp"
setup_twitter_oauth(api_key,api_secret,access_token,access_token_secret)
```

```
## [1] "Using direct authentication"
```

Luego obtenemos el objeto usuario correspondiente y es almacenado en la variable "user". Acto seguido se forman los data frames con los flags requeridos (1:friend,2:follower).

2.Realice un estudio exploratorio de los datos para seleccionar los campos a utilizar y para determinar si es necesario algún tipo de pre-procesamiento o limpieza de los datos para su posterior análisis.

Observamos una pequeña muestra de los datos obtenidos hasta ahora

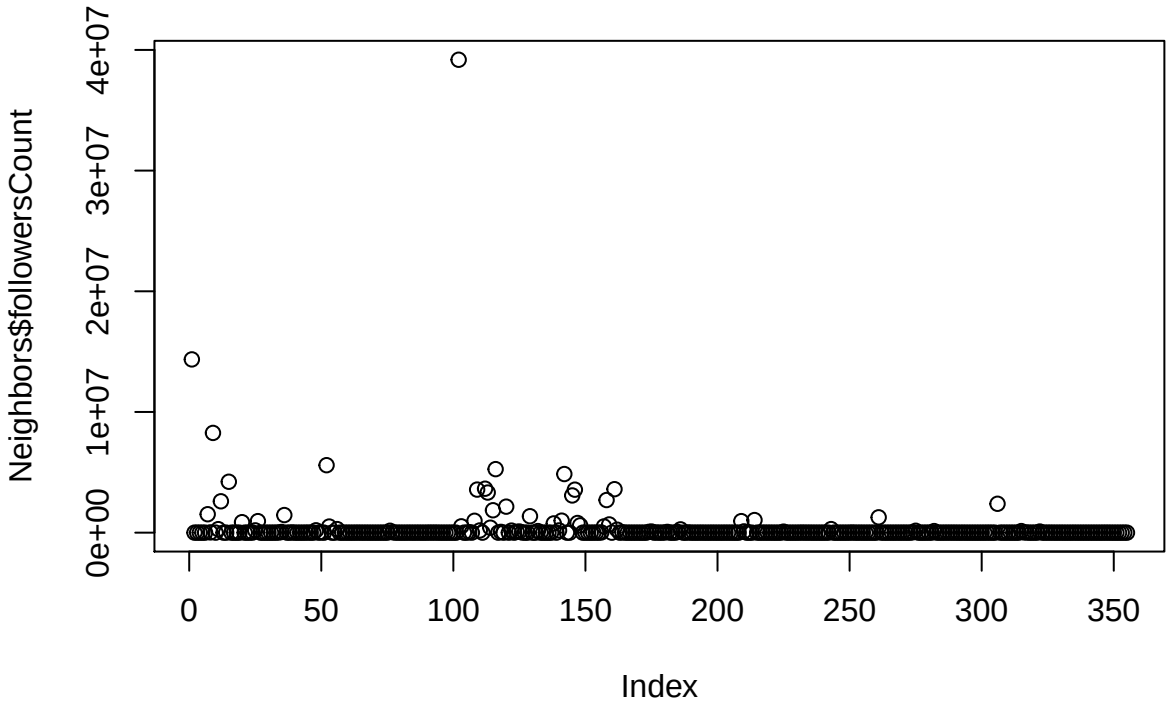
```
Neighbors[sample(nrow(Neighbors),3),]
```

```
##
## 308 No me rendiré ante ti ni ante nadie, solo ante mi porque conozco mi fuerza de voluntad,porque los
## 90
## 344
##     statusesCount followersCount favoritesCount friendsCount  url
## 308           159             69              0          374 <NA>
## 90             70            496              3          582 <NA>
## 344          4169            104             92          315 <NA>
##                      name             created protected verified
## 308 Cristian Andres Laya 2010-04-19 19:24:03     FALSE    FALSE
## 90       Alejandro Tovar 2011-05-15 19:34:37     FALSE    FALSE
## 344       Michelle Gomez 2011-09-17 07:51:34     FALSE    FALSE
##          screenName                  location lang        id listedCount
## 308           Toysk                   #Bogota   es 134895874           1
## 90       AleTovar92 Nueva Esparta, Venezuela    es 299261847           2
## 344 MichelleAgomezj                            en 374956271           0
##     followRequestSent
## 308             FALSE
## 90              FALSE
## 344             FALSE
##                                                                profileImageUrl
## 308 http://pbs.twimg.com/profile_images/450442247528787968/gbyQjyAi_normal.jpeg
## 90  http://pbs.twimg.com/profile_images/496322116485394432/EcRC3kOy_normal.jpeg
## 344 http://pbs.twimg.com/profile_images/532575416037683200/F9PDX6-9_normal.jpeg
##     flag
## 308    2
## 90     1
## 344    2
```

En este punto, pasamos a eliminar los datos que no son numericos, como tambien los datos no relevantes para nuestro estudio.

Luego de observar los datos en el siguiente grafico, nos damos cuenta que es necesario transformar estos datos, ya que se encuentran demasiado alejados uno del otro, en la escala por defecto

```
plot(Neighbors$followersCount, Neighbors$StatusesCount)
```



La transformacion que haremos en este caso sera aplicar la funcion logaritmo a cada entrada del conjunto de datos

```
Neighbors[Neighbors=="0"]<-1
Neighbors$logFollowersCount <-log(Neighbors$followersCount)
Neighbors$logStatusesCount<-log(Neighbors$statusesCount)
Neighbors$logFriendsCount<-log(Neighbors$friendsCount)
Neighbors$logFavoritesCount<-log(Neighbors$favoritesCount)
Neighbors$logListedCount<-log(Neighbors$listedCount)
data<-Neighbors[7:11]
```

Ahora vamos a realizar un analisis exploratorio de los datos a groso modo para definir que relacion vamos a estudiar

```
summary(data)
```

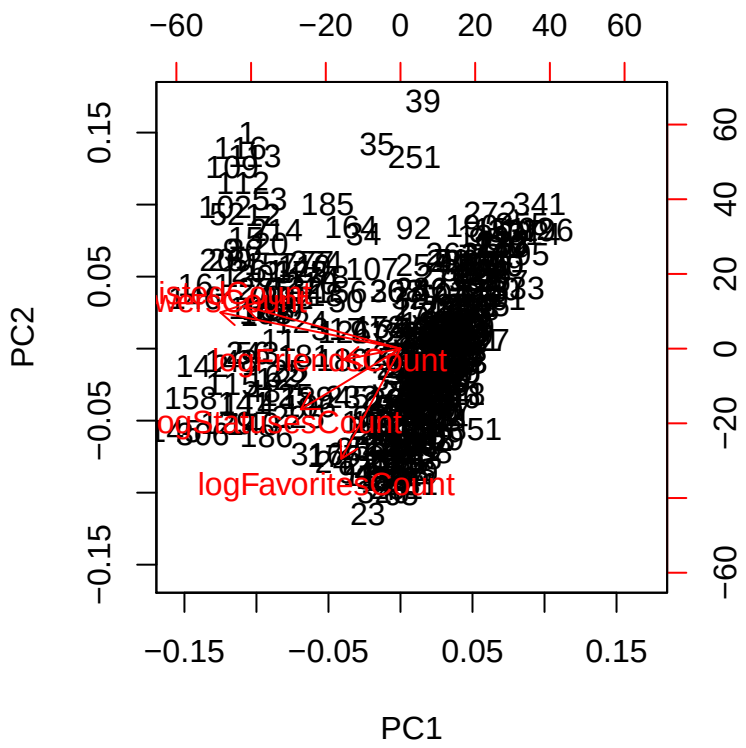
```
##  logFollowersCount logStatusesCount logFriendsCount  logFavoritesCount
##  Min.   : 1.099    Min.   : 0.000   Min.   : 0.000   Min.   :0.000
##  1st Qu.: 4.987    1st Qu.: 6.987   1st Qu.: 5.204   1st Qu.:1.386
##  Median : 5.964    Median : 8.609   Median : 5.953   Median :3.258
##  Mean   : 7.053    Mean   : 8.111   Mean   : 6.127   Mean   :3.370
##  3rd Qu.: 7.646    3rd Qu.: 9.543   3rd Qu.: 6.688   3rd Qu.:5.159
##  Max.   :17.484    Max.   :14.423   Max.   :12.811   Max.   :9.884
```

```
##  logListedCount
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 1.099
##  Mean   : 2.195
##  3rd Qu.: 2.944
##  Max.   :10.337
```

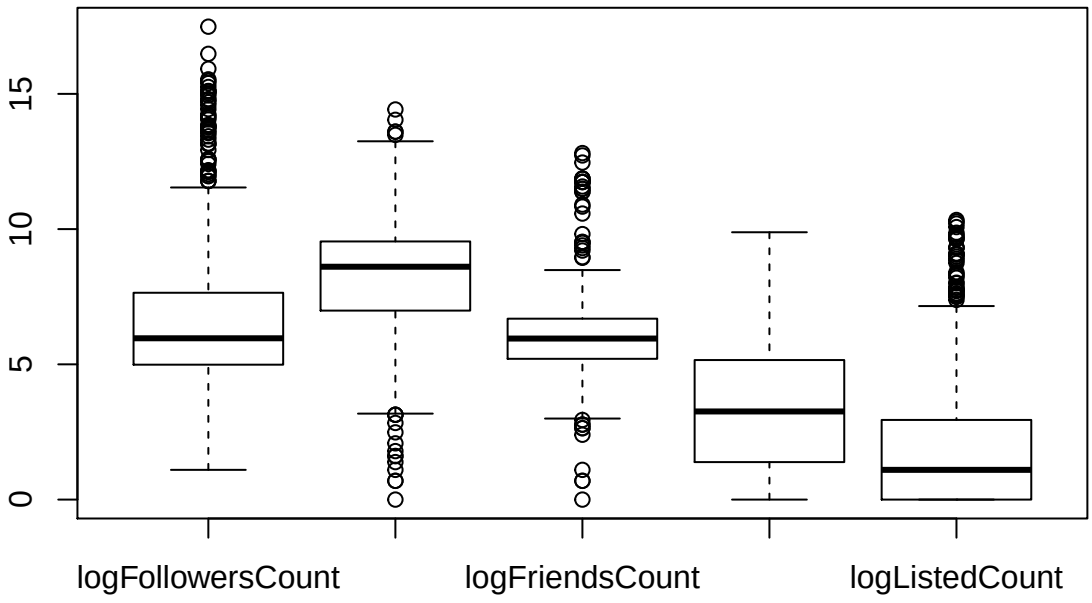
```
cor(data)
```

```
##                   logFollowersCount logStatusesCount logFriendsCount
## logFollowersCount         1.0000000        0.6133324       0.4601841
## logStatusesCount          0.6133324        1.0000000       0.5191689
## logFriendsCount           0.4601841        0.5191689       1.0000000
## logFavoritesCount         0.2905052        0.5356662       0.1866622
## logListedCount            0.9630735        0.5404713       0.4210719
##                   logFavoritesCount logListedCount
## logFollowersCount         0.2905052      0.9630735
## logStatusesCount          0.5356662      0.5404713
## logFriendsCount           0.1866622      0.4210719
## logFavoritesCount         1.0000000      0.2533786
## logListedCount            0.2533786      1.0000000
```

```
biplot(prcomp(data))
```




```
boxplot(data)
```



Los datos a estudiar seran el numero de tweets y retweets y el numero de seguidores.

```
Neighbors$favoritesCount<-NULL
Neighbors$listedCount<-NULL
Neighbors$logFavoritesCount<-NULL
Neighbors$logListedCount<-NULL
```

Ahora podemos observar el cambio en el grafico de puntos

```
plot(Neighbors$logFollowersCount, Neighbors$logFriendsCount)
```

3.Este nuevo data frame conformará el nuevo conjunto de datos de entrada y se debe guardar en un archivo llamado CI1_CI2_C3_twitter_usuario.csv. Donde CI1 y CI2 son las cédulas de los participantes del proyecto y usuario es el nombre del usuario delquese extrajo la información.

```
write.csv(Neighbors, file = "21014872_22022441_23194702_LeoSantella.csv")
```

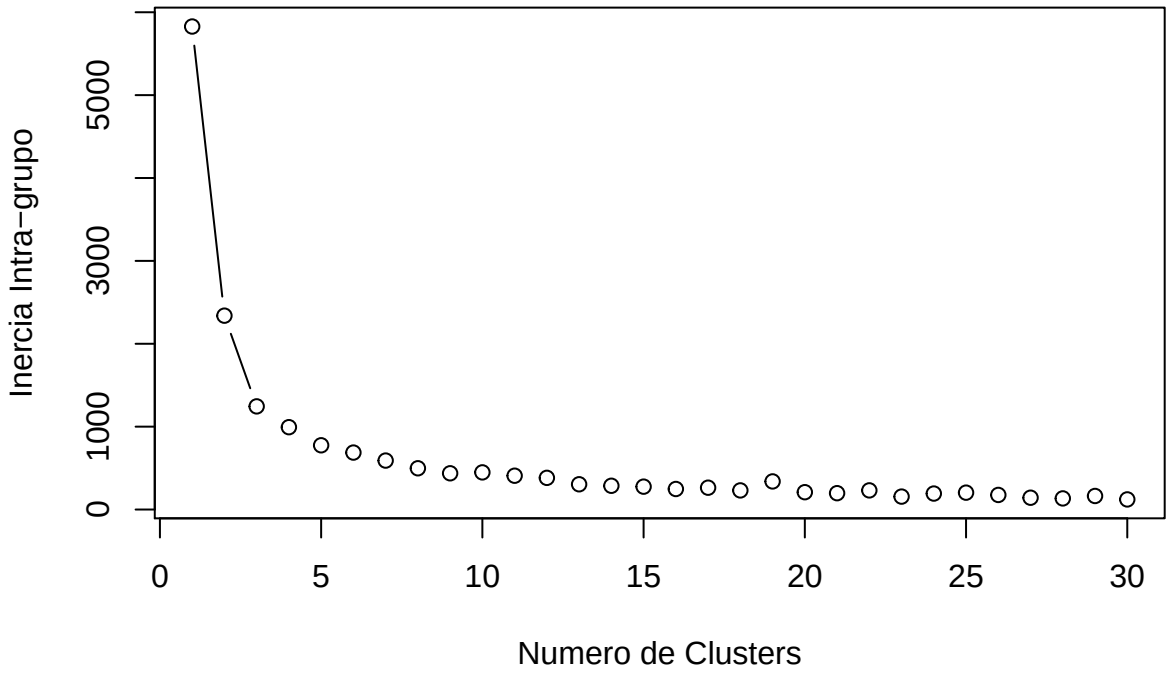
4.Proponga un algoritmo de K-medias implementado en R que, dado este conjunto de datos, retorne el listado de los grupos asignados a cada registro de dicho conjunto de datos de entrada

En este punto de la tarea decidimos que el algoritmo que vamos a utilizar sera el algoritmo K-Medias (K-means) de Hartigan-Wong. En este caso, es el por defecto de la funcion kmeans(...) de R.

5.Aplique el método del "Codo de Jambú" para determinar el número de grupos óptimo para este conjunto de datos.

El grafico deja en evidencia que la inercia Intra-grupo se estabiliza apartir de 5 clusters. Por lo tanto el numero de clusters optimo determinado en este caso es 5.

```
Kobject <- data.frame(Neighbors$logStatusesCount,Neighbors$logFollowersCount)
mydata<-Kobject
wss <- rep(0,30)
for (i in 1:30) wss[i] <- (kmeans(mydata,i))$tot.withinss
plot(1:30, wss, type= "b", xlab="Numero de Clusters", ylab= "Inercia Intra-grupo")
```



6.Aplique el algoritmo propuesto por usted al conjunto de entrada e incorpore los grupos asignados a cada registro como una nueva columna del data frame, almacenándolo en un nuevo archivo con el nombre CI1_CI2_CI3_twitter_usuario_grupos.csv

```
NeighborsMeans <- kmeans(Kobject, centers=5, iter.max=10, nstart=100)
Neighbors$cluster <- NeighborsMeans$cluster
write.csv(Neighbors, file = "21014872_22022441_23194702_LeoSantella_grupos.csv")
```

7.Caracterice los grupos encontrados, identificando por cada uno: a.Número de usuarios que lo componen y su porcentaje del total b.Características más resaltantes y diferenciadoras de cada grupo

```
counts<-count(Neighbors,cluster)
count1<-counts[[1,2]]
count2<-counts[[2,2]]
count3<-counts[[3,2]]
count4<-counts[[4,2]]
count5<-counts[[5,2]]
porcentaje1 <- (count1*100)/355
porcentaje2 <- (count2*100)/355
porcentaje3 <- (count3*100)/355
porcentaje4 <- (count4*100)/355
porcentaje5 <- (count5*100)/355
porcentaje1
```

```
## [1] 11.5493
```

```
porcentaje2
```

```
## [1] 9.295775
```

```
porcentaje3
```

```
## [1] 41.12676
```

```
porcentaje4
```

```
## [1] 11.5493
```

```
porcentaje5
```

```
## [1] 26.47887
```

```
p2 <- nPlot(logFollowersCount ~ logStatusesCount, group = 'cluster', data = Neighbors, type = 'scatterCl
p2$xAxis(axisLabel = 'Statuses Count')
p2$yAxis(axisLabel = 'Followers Count')
p2$chart(tooltipContent = "#! function(key, x, y, e){
return e.point.screenName + ' Followers: ' + e.point.followersCount +' Friends: ' + e.point.statusesCou
} !#")
p2
```

```
## <iframe src=' Tarea1_files/figure-latex/unnamed-chunk-14-1.html ' scrolling='no' frameBorder='0' sea
```

b. Las caracteristicas en las cuales se diferencia el grupo 1 del resto de los grupos es que los elementos que lo conforman tienen una cantidad de Tweets y seguidores en el cual podria decirse que tienen mucha relacion, ademas en el grafico se pude observar que la inercia intra clase en este cluster es notablemente menor en comparacion a otros clusters. Componen un 41% aproximadamente de la poblacion total.

Las caracteristicas en las cuales se diferencia el grupo 2 del resto de los grupos es que los elementos que lo conforman tienen una cantidad de tweets y retweets y seguidores relacionados de una manera parecida a los

elementos del grupo 1. Componen un 26.4% aproximadamente de la poblacion total, a traves del grafico se puede observar que es uno de los grupos con menor inercia intra clase. Es el 2do grupo con mas integrantes.

Las caracteristicas en las cuales se diferencia el grupo 3 del resto de los grupos es que los elementos que lo conforman tienen una cantidad de Tweets y Retweets y seguidores bastante parecidas. Componen un 11.5% respectivamente, aproximadamente de la poblacion total, a traves del grafico se puede observar que es ambos grupos tienen una inercia intra clase notable. Esta compuesto por 41 individuos.

Las caracteristicas en las cuales se diferencia el grupo 4 del resto de los grupos es que los elementos que lo conforman tienen una cantidad de tweets y retweets y seguidores que esta relacionado, es el grupo en el cual los elementos que lo componen tienen una gran cantidad de seguidores como de tweets y retweets. Componen un 11% aproximadamente de la poblacion total, a traves del grafico se puede observar que es uno de los grupos con una inercia intra clase relativamente media en comparacion a los demas grupos. Tambien esta compuesto por 41 elementos

Las caracteristicas en las cuales se diferencia el grupo 5 del resto de los grupos es que los elementos que lo conforman tienen una cantidad de tweets y retweets y seguidores que esta relacionado, es el grupo en el cual los elementos que lo componen tienen una escueta cantidad de seguidores como de tweets y retweets. Componen un 9% aproximadamente de la poblacion total, a traves del grafico se puede observar que es uno de los grupos con una inercia intra clase relativamente media en comparacion a los demas grupos. Tambien esta compuesto por 33 elementos

Claramente nos damos cuenta que el mayor porcentaje de la poblacion lo acumulan los grupos 1 y 2, en los cuales se encuentran personas que tienen numeros no muy altos de seguidores y tweets, y ademas es casi proporcional el numero de tweets y seguidores.

Cabe destacar que esta evidenciado a traves del grafico, que existen individuos con muchos seguidores, mucho mas seguidores que tweets o retweets. Esto es observable en el grafico.