**CURSO DE FORMACIÓN EN DATA MINING CON "R"**

## TAREA 6

***1) Lea el archivo de datos “titanic”, del archivo “titanic.RData”, que contiene 2201 casos relativos a las personas que viajaban en el Titanic con cuatro variables indicando sexo, edad, clase en la que viajaban y si sobrevivieron o no. Aplique el método Apriori (sustituya en las órdenes del ejemplo de la unidad “compra” por “titanic”) para intentar relacionar a los supervivientes con alguna de las características indicadas. Ordene las reglas obtenidas por el estadístico de interés “lift” en lugar de “confidence” e interprete únicamente las reglas con lift > 1,20 que considere que parecen relevantes.***

Establecemos directorio de trabajo:

setwd("/Users/jlsovaz/Desktop/curso\_data\_mining/TEORIA/UNIDAD 6/")

Cargamos el paquete necesario para poder utilizar el método Apriori:

library(arules)

Cargamos el dataset que vamos a utilizar:

load("titanic.RData")

Es necesario transformar el dataset a analizar en un objeto de clase "transactions". Para ello:

transacciones <- as(titanic, "transactions")

Aplicamos el método apriori con los parámetros por defecto:

reglas <- apriori(transacciones, parameter = list(support = 0.1, confidence = 0.80))

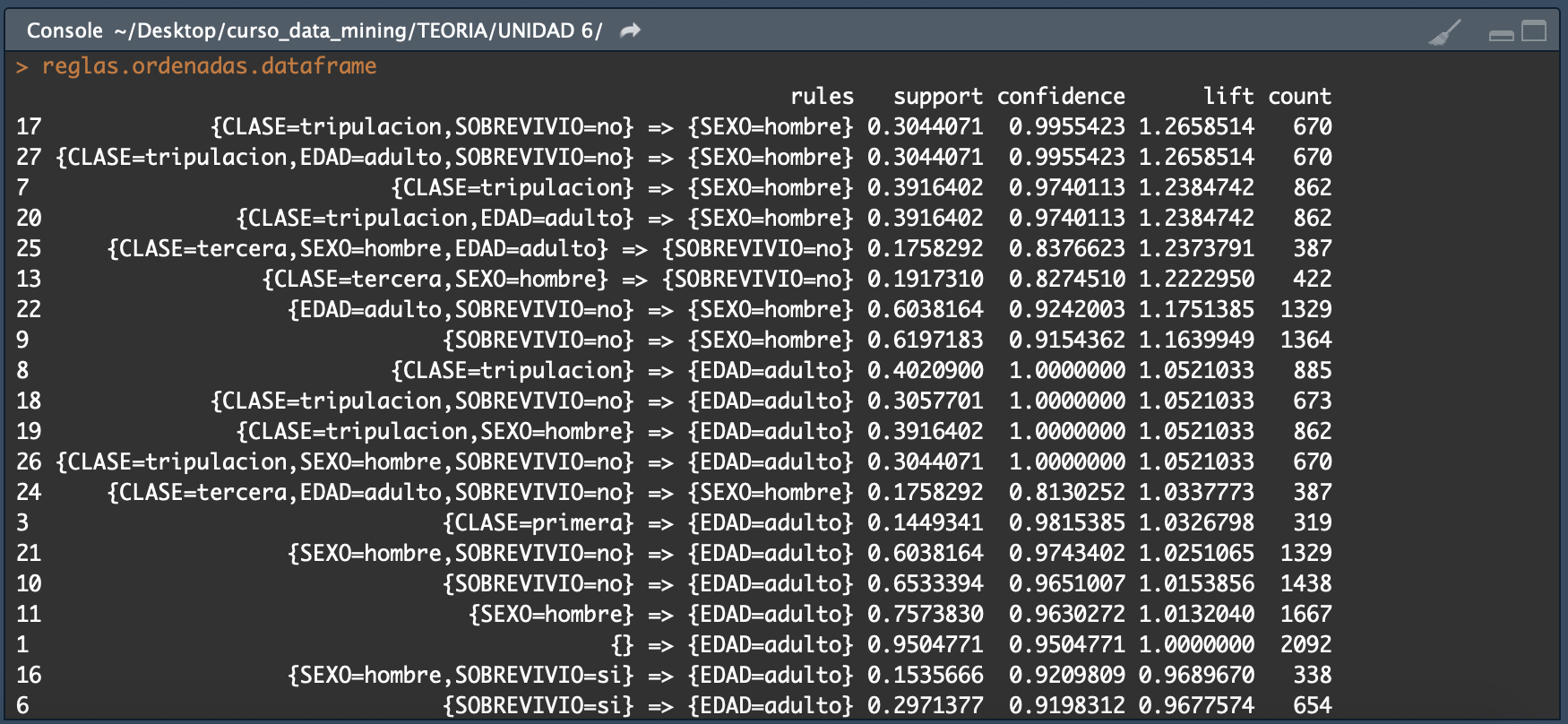
Ordenamos las reglas de mayor a menor en función de la variable “lift”:

reglas.ordenadas <- sort(reglas, by = "lift")

Transformamos el objeto reglas ordenadas en data.frame:

reglas.ordenadas.dataframe <- as(reglas.ordenadas, "data.frame")

Exploramos el resultado:

reglas.ordenadas.dataframe

El algoritmo ha encontrado 6 subconjuntos frecuentes que presentan un valor de lift mayor que 1.20:

El dataframe presenta 4 tipos de variables: CLASE (tripulación, primera, segunda, tercera); SEXO (hombre, mujer); EDAD (adulto, niño); SOBREVIVIÓ (Si, No). La interpretación de los subconjuntos ordenados de mayor a menor lift son:

1. El subconjunto definido por: Las personas de la clase **tripulación** **que no han sobrevivido** son mayoritariamente **hombres** (la frecuencia absoluta es de 670 casos, con una confidencia de 0.99 y una frecuencia de 0.30).
2. El subconjunto definido por: Las personas de la clase **tripulación** **adultas** que **no han sobrevivido** son mayoritariamente **hombres** (la frecuencia absoluta es de 670 casos, con una confidencia de 0.99 y una frecuencia de 0.30).
3. El subconjunto definido por: Las personas de la clase **tripulación** en general son mayoritariamente **hombres** (la frecuencia absoluta es de 862 casos, con una confidencia de 0.97 y una frecuencia de 0.39).
4. El subconjunto definido por: Las personas de la clase **tripulación** **adultas** son mayoritariamente **hombres** (la frecuencia absoluta es de 862 casos, con una confidencia de 0.97 y una frecuencia de 0.39).
5. El subconjunto definido por: Los **hombres** **adultos** de la clase **tercera** mayoritariamente **no han sobrevivido** (la frecuencia absoluta es de 387 casos, con una confidencia de 0.83 y una frecuencia de 0.17).
6. El subconjunto definido por: Los **hombres** **adultos** de la clase **tercera** mayoritariamente **no han sobrevivido** (la frecuencia absoluta es de 422 casos, con una confidencia de 0.82 y una frecuencia de 0.19).

***2) Lea el conjunto de datos "futbol.csv", que contiene un conjunto de 75 partidos de futbol (datos reales tomados del principio de la liga actual) jugados por 15 equipos, con 2 columnas o variables (PIERDE, GANA) indicando los equipos perdedor/ganador en cada partido (se han omitido los empates). Aceptando que tiene más mérito ganar a un equipo que suele ganar que a otro que suele perder, considerando que el equipo que pierde emite un voto a favor del que gana y que ese voto tiene más valor cuando viene de un equipo más meritorio, construya un índice (pagerank) que permita ordenar a los equipos por “merito deportivo”.***

***{      Dataset <- read.csv2("futbol.csv", header=TRUE,encoding="latin1")      }***

Cargamos el paquete necesario para poder realizar el algoritmo de PageRank:

library(igraph)

Establecemos directorio de trabajo:

setwd("/Users/jlsovaz/Desktop/curso\_data\_mining/TEORIA/UNIDAD 6/")

Cargamos el dataset con el que trabajaremos:

Dataset <- read.csv2("futbol.csv", header=TRUE,encoding="latin1")

Es neceario convertir el dataframe que queremos analizar en un objeto "grafo"

grafo <- graph.data.frame(Dataset)

Establecemos una semilla para la reproducibilidad de los resultados:

set.seed(123)

Ploteamos el grafo para visualizar todas las relaciones de i conoce a j

plot(grafo)

Calcular el índice de PageRank:

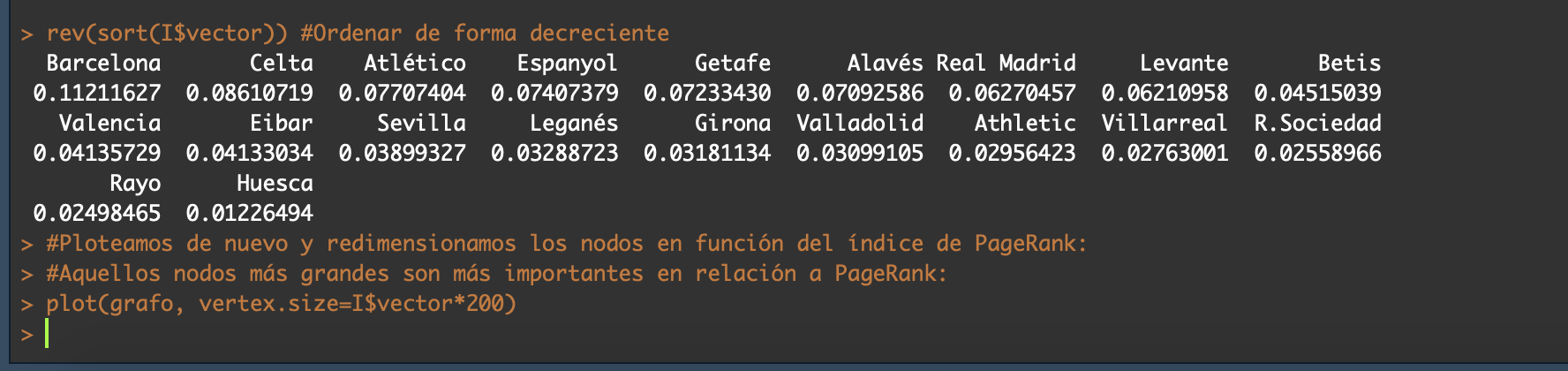
I <- page.rank(grafo)

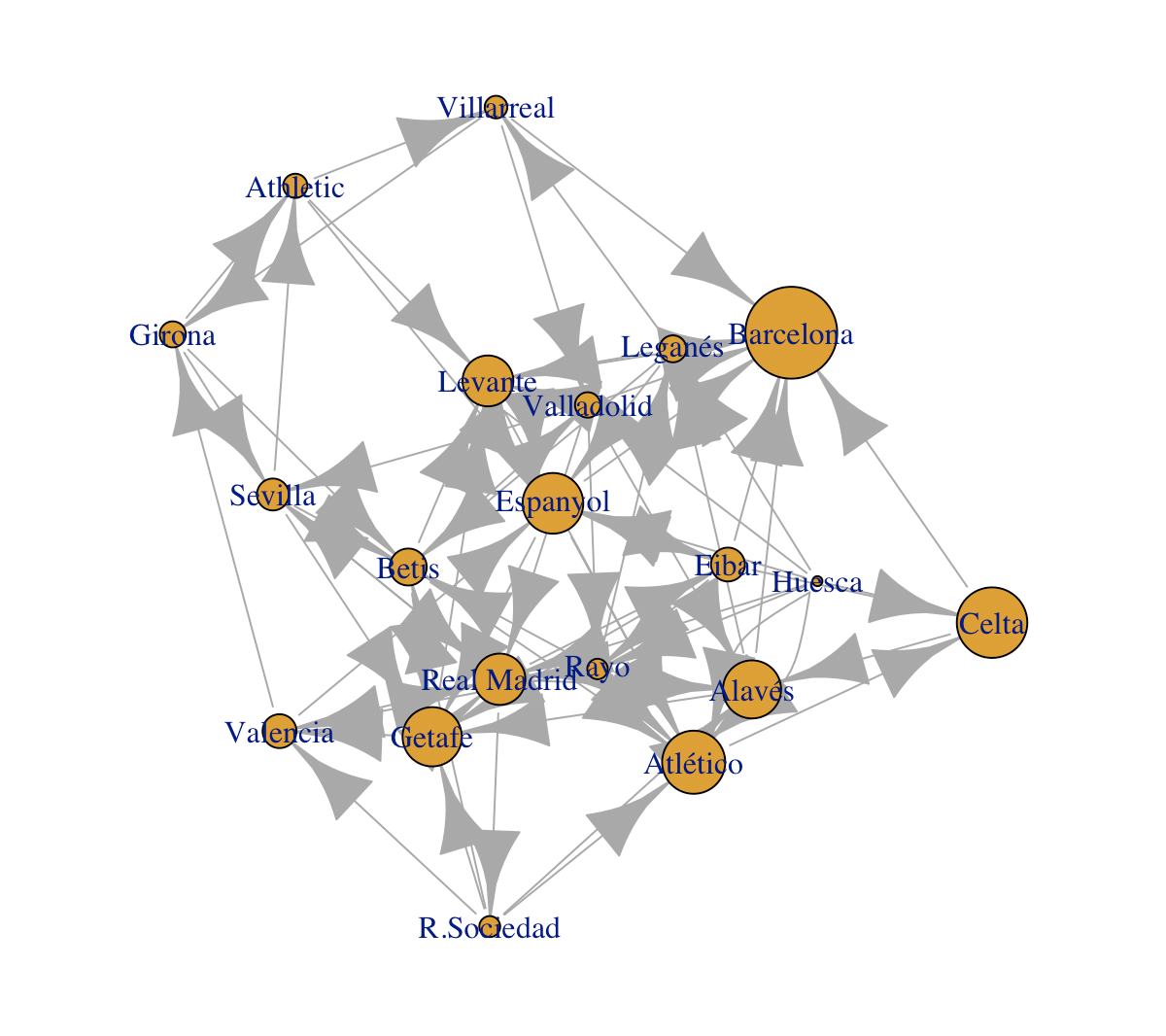
I

rev(sort(I$vector)) #Ordenar de forma decreciente

Ploteamos de nuevo y redimensionamos los nodos en función del índice de PageRank:

Aquellos nodos más grandes son más importantes en relación a PageRank:

plot(grafo, vertex.size=I$vector\*200)



Según el algoritmo de PageRank, en este ejemplo decimos que los equipos con mayor mérito Deportivo no son solo aquellos que ganan más partidos, sino los que además ganan a los equipos que suelen ganar más que perder.

Con los resultados del algoritmo podemos ver que los equipos con mayor índice de PageRank son El Barcelona, el Celta y el Atlético, etc. (y no necesariamente son los equipos que más victorias han obtenido, sino que es más importante para PageRank que hayan ganado a equipos que también suelen ganar).

Por ejemplo, el Celta sólo ganó tres partidos, pero uno de ellos ha sido contra el Atlético, que a su vez ha ganado 8 partidos. Por esta razón el Celta obtiene importancia (mérito) según PageRank