Tarea R #6 (sol)

Diego Alonso Alfaro Bergueiro

6 de noviembre de 2016

Nota Aclaratoria

Esta representa solo una posible solución. No hay expectativa alguna de que su trabajo haya sido exáctamente igual a este, lo importante es que hayan desarrollado el código para hacer lo que se les pidió y que hayan podido describir lo que estaban haciendo y los resultados.

Análisis del Problema

La agrupación y clasificación automática de documentos, en una época donde casi todo contenido se puede digitalizar, son tareas de suma relevancia en diferentes contextos. Por ejemplo, se pueden utilizar para agrupar documentos con contenidos similares para presentarle sugerencias a usuarios acerca de qué otros documentos les pueden ser relevantes. Adicionalmente, para detección de plagios, el poder agrupar documentos de acuerdo a su similitud puede ayudar a editoriales o a autores individuales a encontrar documentos que sean muy similares a los suyos, para luego determinar acciones potenciales a tomar.

Entendimiento de los Datos

En este caso, se tienen 6 documentos que datan de la época de la independencia de los Estados Unidos de América. Tres documentos son atribuídos al autor Jay, uno a Hamilton, uno a Madisson y otro se atribuye a una colaboración. Sin embargo, hay dudas con respecto a si en el documento que está marcado como "colaboración" predomina el estilo de uno de los tres autores.

Exploración de los Datos

```
library(cluster)
library(flexclust)
```

```
## Loading required package: grid
```

```
## Loading required package: lattice
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
library(caTools)
library(tm)
```

```
## Loading required package: NLP
```

```r
library(SnowballC)
library(RWeka)
```

```
##
## Attaching package: 'RWeka'

## The following object is masked from 'package:caTools':
##
##     LogitBoost
```

```r
library(qcc)
```

```
## Package 'qcc', version 2.6

## Type 'citation("qcc")' for citing this R package in publications.
```

```r
setwd('E:/CENFOTEC/Maestría - Introducción a Minería de Datos/Materiales de las Clases/Semana 12/dataset

texto1 <- readChar('Chapter12_Federalist03_Jay.txt', file.info('Chapter12_Federalist03_Jay.txt')$size)
texto1 <- gsub('\r\n\r\n', ' ', texto1)

texto2 <- readChar('Chapter12_Federalist04_Jay.txt', file.info('Chapter12_Federalist04_Jay.txt')$size)
texto2 <- gsub('\r\n\r\n', ' ', texto2)

texto3 <- readChar('Chapter12_Federalist05_Jay.txt', file.info('Chapter12_Federalist05_Jay.txt')$size)
texto3 <- gsub('\r\n\r\n', ' ', texto3)

texto4 <- readChar('Chapter12_Federalist14_Madison.txt', file.info('Chapter12_Federalist14_Madison.txt'
texto4 <- gsub('\r\n\r\n', ' ', texto4)

texto5 <- readChar('Chapter12_Federalist17_Hamilton.txt', file.info('Chapter12_Federalist17_Hamilton.tx
texto5 <- gsub('\r\n\r\n', ' ', texto5)

texto6 <- readChar('Chapter12_Federalist18_Collaboration.txt', file.info('Chapter12_Federalist18_Collab
texto6 <- gsub('\r\n\r\n', ' ', texto6)

##Consolidar en un solo dataset
datos <- data.frame(rbind(cbind(Autor = 'Jay', Nombre = 'Documento 1', Texto = texto1),
                          cbind(Autor = 'Jay', Nombre = 'Documento 2', Texto = texto2),
                          cbind(Autor = 'Jay', Nombre = 'Documento 3', Texto = texto3),
                          cbind(Autor = 'Madison', Nombre = 'Documento 4', Texto = texto4),
                          cbind(Autor = 'Hamilton', Nombre = 'Documento 5', Texto = texto5),
                          cbind(Autor = 'Varios', Nombre = 'Documento 6', Texto = texto6)))
```

Los 6 textos se pueden consolidar en un solo data frame, para luego aplicar las técnicas de pre-procesamiento de texto:

```r
datos
```

```
##      Autor      Nombre
## 1      Jay Documento 1
```

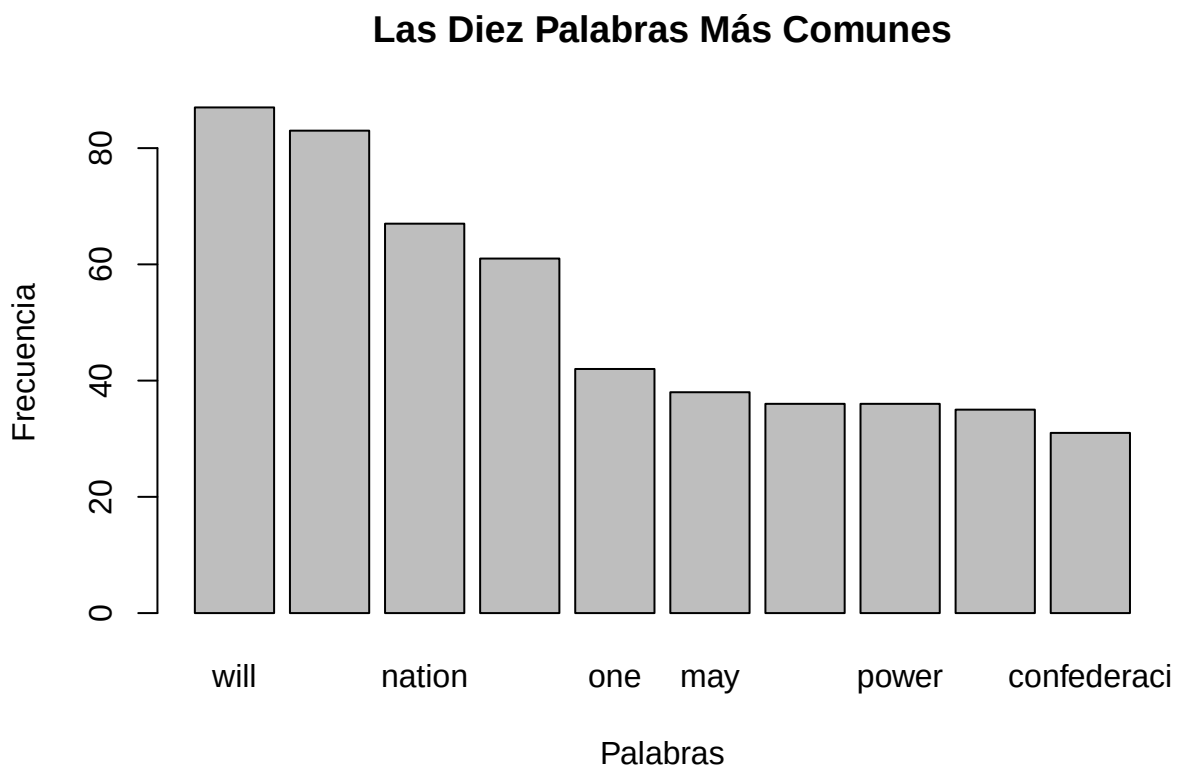
```
## 2      Jay Documento 2
## 3      Jay Documento 3
## 4  Madison Documento 4
## 5 Hamilton Documento 5
## 6   Varios Documento 6
##
## 1
## 2
## 3
```

```
corpus.tarea <- Corpus(VectorSource(datos$Texto))

corpus.tarea <- tm_map(corpus.tarea, tolower)
corpus.tarea <- tm_map(corpus.tarea, PlainTextDocument)

corpus.tarea <- tm_map(corpus.tarea, removeNumbers)

corpus.tarea <- tm_map(corpus.tarea, removePunctuation)

corpus.tarea <- tm_map(corpus.tarea, stemDocument)

corpus.tarea <- tm_map(corpus.tarea, stripWhitespace)

corpus.tarea <- tm_map(corpus.tarea, removeWords, stopwords("english"))

ngramTokenizer <- function(x) NGramTokenizer(x, Weka_control(min = 1, max = 1))

dtm <- DocumentTermMatrix(corpus.tarea, control = list(tokenize = ngramTokenizer))

dtm <- data.frame(as.matrix(dtm))
```

```
## Warning in data.row.names(row.names, rowsi, i): some row.names duplicated:
## 2,3,4,5,6 --> row.names NOT used
```

Y luego de consolidar los textos en un solo corpus, se puede visualizar cuáles son las palabras más comunes:

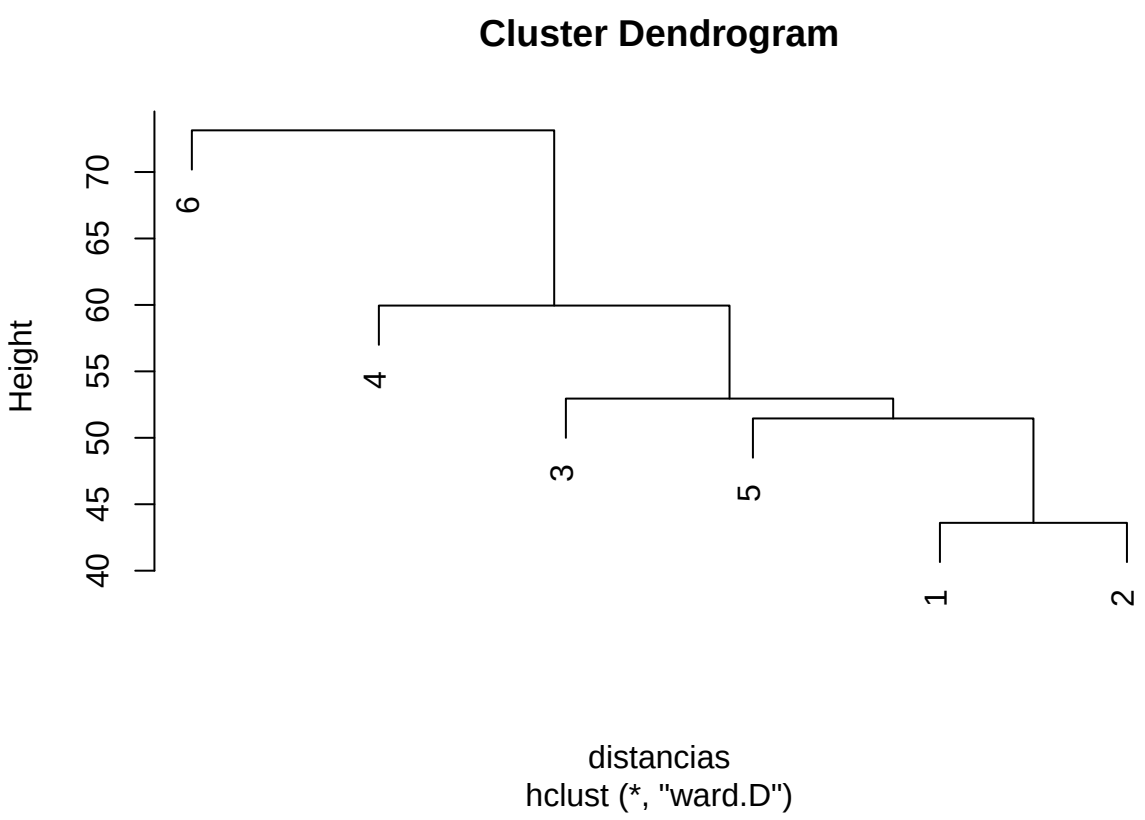
```
barplot(head(sort(colSums(dtm), decreasing = T), 10),
        main = 'Las Diez Palabras Más Comunes',
        xlab = 'Palabras',
        ylab = 'Frecuencia')
```



Creación del Modelo

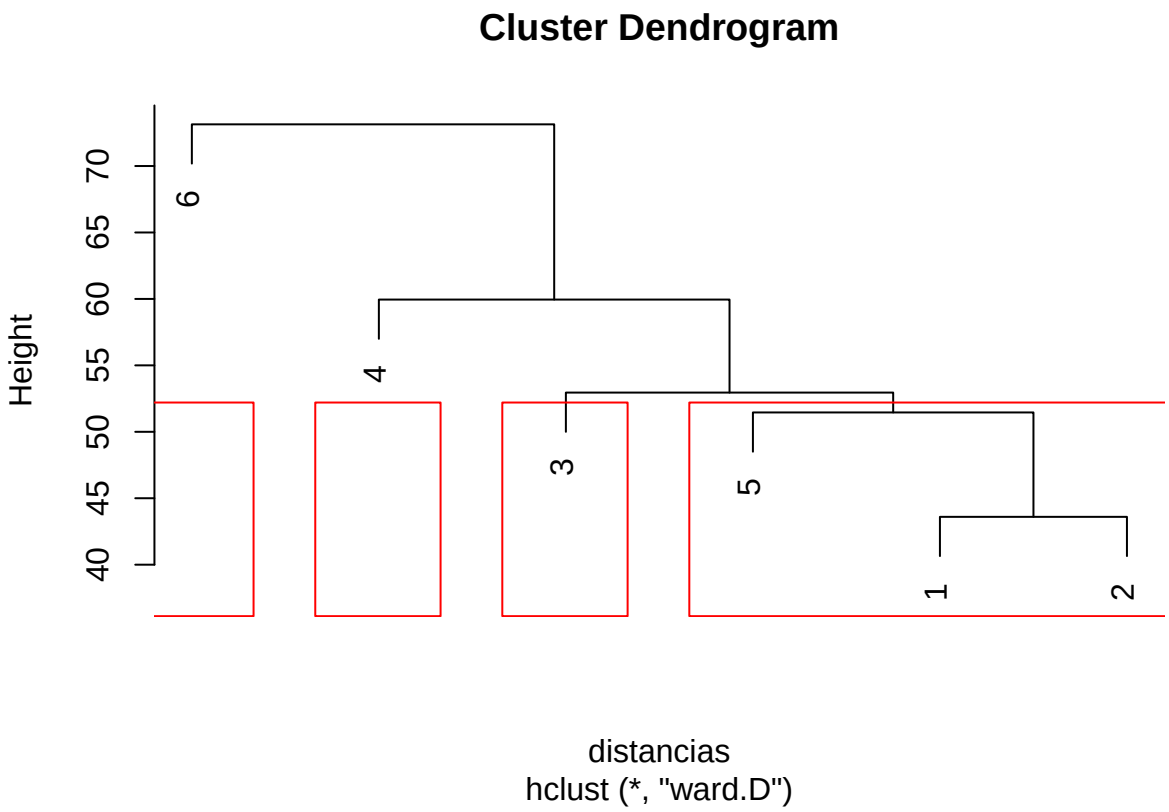
Luego de procesar el texto y consolidar el conteo de palabras en una matriz documentos-términos, se procede a calcular la distancia entre los diferentes documentos. A partir de esta matriz de distancia se puede hacer un agrupamiento jerárquico que nos permite determinar la cantidad de grupos que se pueden generar.

```
#calcular distancias entre documentos
distancias <- dist(dtm, method="euclidean") #calcular distancias usando distancia euclidiana

####utilizar agrupación jerárquica en los datos
datos.jerarquico <- hclust(distancias, method="ward.D")

##ver el reultado del agrupamiento jerárquico
plot(datos.jerarquico)
```



De acuerdo con el agrupamiento jerárquico, se puede comenzar con 2 grupos donde queda 1 documento solo (el 6) y el resto en un segundo grupo. Poco a poco, se pueden crear más grupos, y se van creando grupos donde quedan documentos solos más un solo grupo con varios documentos. Para ilustrar esto, se pueden crear 4 grupos, y se aprecia como quedan 3 grupos con un documento cada uno, y un grupo con 3 documentos.

```
##ver el reultado del agrupamiento jerárquico
plot(datos.jerarquico)

##Marcar 2 clústeres
rect.hclust(datos.jerarquico, k = 4, border = "red")
```




```
##Asignar # de grupo
datos$cluster.jerarquico <- factor(cutree(datos.jerarquico, k=4))
```

Trabajando con 4 grupos, podemos utilizar también agrupamiento por particiones.

```
##Agrupación con KMeans
set.seed(3235) #necesario para replicabilidad
km <- kmeans(dtm, centers = 4)

##Observaciones por clúster
table(km$cluster)
```

```
##
## 1 2 3 4
## 1 3 1 1
```

```
##Asignar # de clúster a los datos originales
datos$cluster.particionamiento <- factor(km$cluster)
```

Evaluación

Al asignar los números de grupo al conjunto de datos, se puede ver claramente como los dos métodos de agrupamiento, tanto el jerárquico como el de particionamiento, crearon los mismos grupos aún si asignaron números de grupos diferente:

```
datos[, c(1, 2, 4, 5)]
```

```
##      Autor      Nombre cluster.jerarquico cluster.particionamiento
## 1      Jay Documento 1                  1                        2
## 2      Jay Documento 2                  1                        2
## 3      Jay Documento 3                  2                        1
## 4  Madison Documento 4                  3                        3
## 5 Hamilton Documento 5                  1                        2
## 6   Varios Documento 6                  4                        4
```

En general, un documento de Jay, uno de Madison y la colaboración forman grupos por sí solos, mientras dos documentos de Jay y el de Hamilton están en el mismo grupo. Esto nos permitiría concluir que en la colaboración no predomina el estilo de un solo escritor, y que Madison tiene una estilo similar a Jay en cuanto a selección de palabras.

Podemos reducir la cantidad de grupos a dos, y al realizar el agrupamiento por particiones, tenemos esta distribución de documentos:

```
##Agrupación con KMeans
set.seed(3235) #necesario para replicabilidad
km <- kmeans(dtm, centers = 2)

##Observaciones por clúster
table(km$cluster)
```

```
##
## 1 2
## 1 5
```

```
##Asignar # de clúster a los datos originales
datos$cluster.particionamiento <- factor(km$cluster)

datos[, c(1, 2, 4, 5)]
```

```
##      Autor      Nombre cluster.jerarquico cluster.particionamiento
## 1      Jay Documento 1                  1                        2
## 2      Jay Documento 2                  1                        2
## 3      Jay Documento 3                  2                        2
## 4  Madison Documento 4                  3                        2
## 5 Hamilton Documento 5                  1                        2
## 6   Varios Documento 6                  4                        1
```

Al reducir la cantidad de grupos a 2, se queda el documento que está marcado como colaboración en un grupo por sí solo, mientras que el resto de documentos están en un mismo grupo. Nuevamente, esto refuerza la conclusión de que en el documento colaborativo no predomina el estilo de ninguno de los escritores, por lo menos en lo que respecta a selección de palabras.

Resultados

En general, el pre-procesamiento de los diferentes textos nos permitió pasar de tener 6 documentos a tener 6 observaciones con variables claramente definidas. Con este conjunto de datos, se pudo realizar un análisis de agrupamiento con dos técnicas diferentes, el cual nos ayudó a sacar la conclusión de que en el documento colaborativo no predomina el estilo de alguno de los 3 autores de los cuales se tienen documentos.