ReporteKmeansOnline

Jessic Vega

31/8/2020

1) Explicar la elección del número de clusters (gráfica si es necesario)

Con base en nuestra implementación del algoritmo de kmeans online, elegimos el número de clusters como aquel que minimice la distancia intra clusters (lo cual sabemos que es equivalente a maximizar la distancia entre clusters).

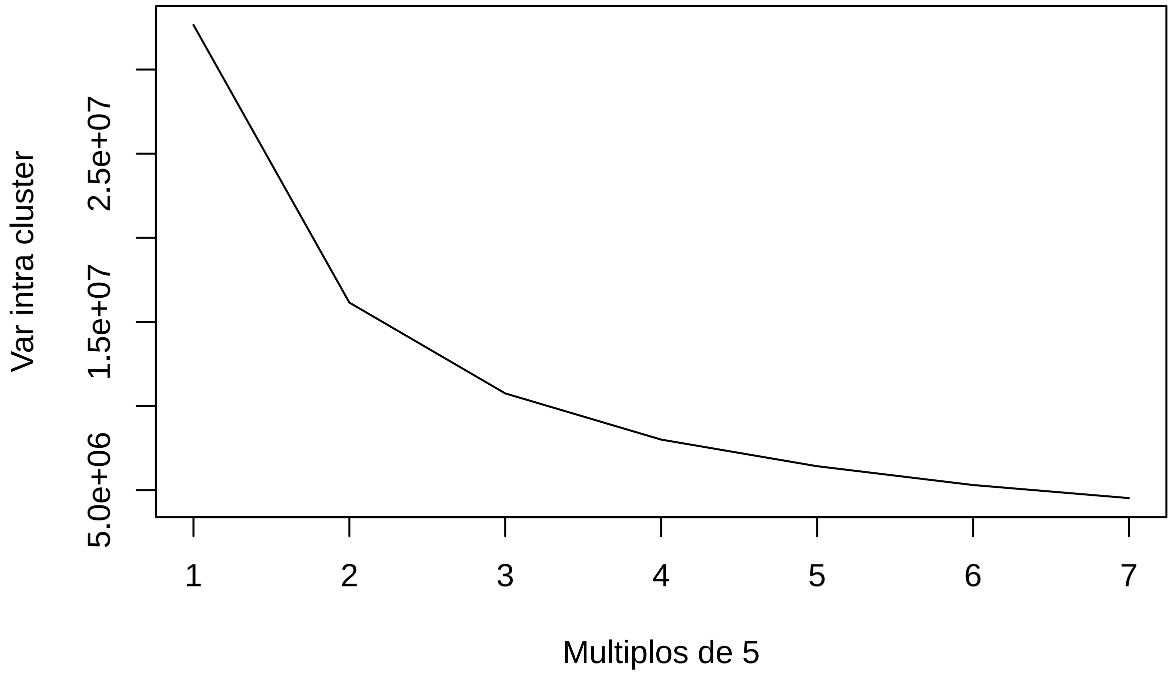
En la gráfica siguiente podemos ver que conforme se incrementa el número de clusters la distancia intra clusters disminuye, sin embargo por el criterio de codo, elegimos 25 como un número de clusters prudente.

```r
data <- read.csv(file = 'docword.nips.txt', header = FALSE, sep=' ', skip = 3)
names(data) <- c('Id.Doc', 'Id.Word', 'freq')
require(reshape2)
```

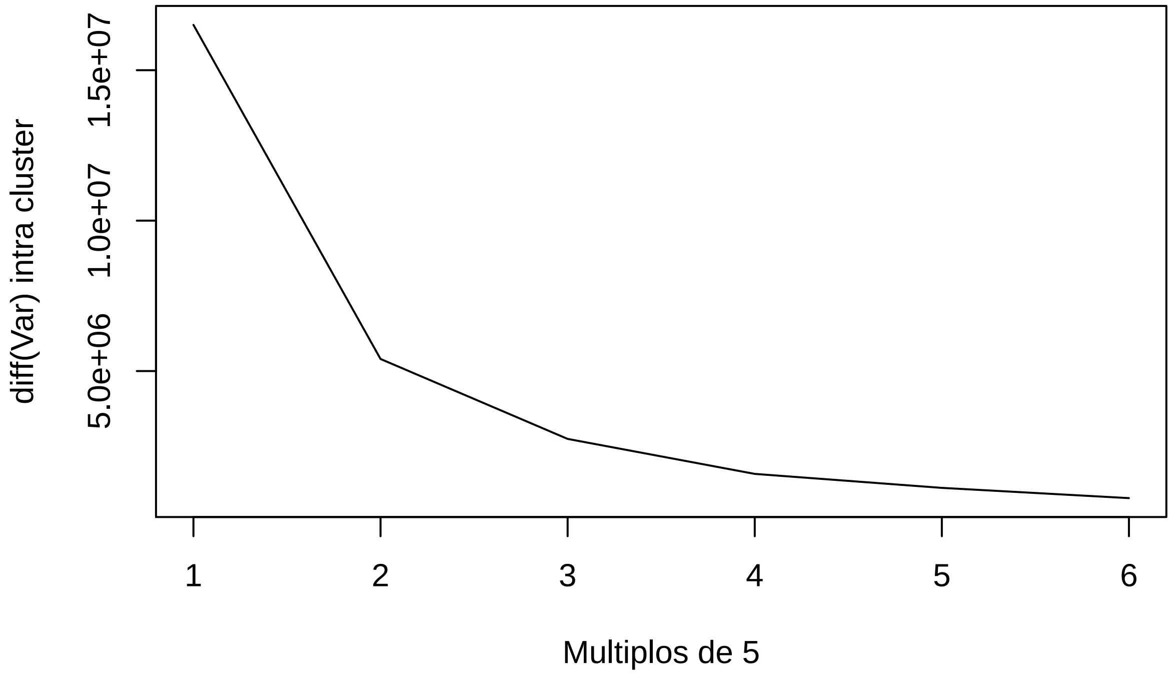
```
## Loading required package: reshape2
```

```r
docs.vector <- dcast(data, Id.Doc ~Id.Word, value.var = 'freq', fill=0)
docs.vector$Id.Doc <- NULL
#docs.vector <- head(docs.vector, 100)
#docs.vector[, 1:dim(docs.vector)[2]] <- scale(docs.vector)

cluster <- 1:7
for( i in 1:7)
{
  #print(i)
  res <- kmeans.online.b(k=5*i)
  cluster[i] <- sum(res$statas.intra)
  #print(cluster)
}
plot((cluster),type = 'l', xlab = 'Multiplos de 5', ylab = 'Var intra cluster')
```



```r
plot(abs(diff(cluster)),type = 'l',  xlab = 'Multiplos de 5', ylab = 'diff(Var) intra cluster')
```



2) Mostrar las 10 palabras más comunes de cada cluster en una Tabla.

```
k <- 25
res <- kmeans.online.b(k=k)
words <- read.csv('vocab.nips.txt', header = FALSE)
names(words) <- 'Palabra'
words$Id.Word <- as.numeric(row.names(words))
data <- merge(data, words, all.x=TRUE)
cluster.palabras.top10 <- data.frame(Cluster=1:25, Palabras='')
require(dplyr)
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
for (i in 1:25)
{
  index <- which(res$tabla.master$Cluster==i)
  data.subset <- subset(data, Id.Doc %in% index)
  data.subset %>% select(Id.Doc, freq, Palabra) %>% group_by(Palabra) %>% summarise(freq=sum(freq)) %>%
    arrange(-freq) %>% head(10) -> temp
  string <- paste0(temp$Palabra, collapse = '', sep=', ')
  cluster.palabras.top10$Palabras[i] <- string
}
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
write.csv(res$tabla.master,file='tabla.master.csv', row.names = FALSE)
cluster.palabras.top10
```

```
##    Cluster
## 1        1
## 2        2
## 3        3
## 4        4
## 5        5
## 6        6
## 7        7
## 8        8
## 9        9
## 10      10
## 11      11
## 12      12
## 13      13
## 14      14
## 15      15
## 16      16
## 17      17
## 18      18
## 19      19
## 20      20
## 21      21
## 22      22
## 23      23
## 24      24
## 25      25
##                                                                                     Palabras
## 1     network, unit, neural, system, learning, function, algorithm, error, hidden, weight,
## 2    network, model, neural, learning, neuron, input, algorithm, function, system, output,
## 3             network, unit, function, input, model, neural, output, training, layer, set,
## 4  model, function, data, set, network, distribution, learning, vector, neural, algorithm,
## 5       network, unit, neural, input, learning, function, output, weight, system, problem,
## 6      function, model, algorithm, input, network, learning, set, neural, vector, problem,
## 7        network, learning, weight, function, algorithm, input, neural, set, data, output,
## 8        model, network, algorithm, learning, data, unit, hidden, function, set, training,
## 9     network, neural, training, set, input, classifier, system, data, algorithm, problem,
## 10        function, data, system, network, algorithm, set, neural, model, neuron, problem,
## 11   learning, action, system, control, function, model, task, algorithm, network, result,
## 12           network, learning, training, error, set, model, data, unit, input, algorithm,
## 13            model, cell, network, data, input, motion, neural, learning, system, visual,
## 14     learning, algorithm, function, model, set, vector, data, system, problem, training,
## 15         network, neuron, neural, learning, input, model, set, pattern, function, error,
## 16       cell, learning, network, algorithm, input, model, vector, output, system, weight,
## 17        network, unit, input, output, layer, hidden, learning, set, algorithm, function,
## 18             model, data, set, cell, function, point, network, input, parameter, number,
## 19   learning, model, function, network, neuron, system, input, neural, object, algorithm,
## 20          network, input, neural, algorithm, set, system, function, point, output, data,
## 21           network, function, model, data, set, training, input, weight, method, neural,
## 22          model, network, system, neural, input, neuron, data, cell, information, field,
## 23           network, model, input, error, object, training, set, function, neural, image,
## 24          network, system, function, point, error, input, learning, unit, model, neural,
## 25           network, input, neural, model, set, system, training, data, learning, output,
```

```
require(xtable)
```

```
## Loading required package: xtable
```

3) Breve intuición acerca del tipo de documento que cada cluster representa.

A grandes rasgos podemos englobar a los clusters en 4 meta clusters, hacemos hincapié en que nuestra implementación puede, por ejemplo, estar detectando papers (documentos) en años parecidos en lugar de grandes temas que se distribuyen en el contenido de los mismos.

El primer meta cluster corresponde a papers de contenido meramente teórico que no hace referencia a la aplicación de un método en particular ni de su implementación por ello las palabras 'algorithm', 'input' , 'output', etc. no aparecen en su top 10. Este meta cluster está integrado por los clusters: 1, 8, 10, 11, 20, 21 y 24.

El segundo meta cluster está formado por pares que hacen referencia a implementaciones de métodos por lo cual contienen pseudocódigos y las palabras 'algorithm', 'input' , 'output' están presentes en ellos (en su top 10). Ese meta cluster está formado por los clusters: 2, 3, 4, 5, 6, 7, 9, 12, 15 y 17.

El tercer meta cluster está formado por los papers de aplicación no médica en donde se utiliza un conjunto de datos y se cuantifica el error de una o varias técnicas de aprendizaje máquina. Este mega cluster está formado por los clusters: 14, 18, 19 y 25

Cluster	Palabras
1	network, unit, neural, system, learning, function, algorithm, error, hidden, weight
2	network, model, neural, learning, neuron, input, algorithm, function, system, output
3	network, unit, function, input, model, neural, output, training, layer, set
4	model, function, data, set, network, distribution, learning, vector, neural, algorithm
5	network, unit, neural, input, learning, function, output, weight, system, problem
6	function, model, algorithm, input, network, learning, set, neural, vector, problem
7	network, learning, weight, function, algorithm, input, neural, set, data, output
8	model, network, algorithm, learning, data, unit, hidden, function, set, training
9	network, neural, training, set, input, classifier, system, data, algorithm, problem
10	function, data, system, network, algorithm, set, neural, model, neuron, problem
11	learning, action, system, control, function, model, task, algorithm, network, result
12	network, learning, training, error, set, model, data, unit, input, algorithm
13	model, cell, network, data, input, motion, neural, learning, system, visual
14	learning, algorithm, function, model, set, vector, data, system, problem, training
15	network, neuron, neural, learning, input, model, set, pattern, function, error
16	cell, learning, network, algorithm, input, model, vector, output, system, weight
17	network, unit, input, output, layer, hidden, learning, set, algorithm, function
18	model, data, set, cell, function, point, network, input, parameter, number
19	learning, model, function, network, neuron, system, input, neural, object, algorithm
20	network, input, neural, algorithm, set, system, function, point, output, data
21	network, function, model, data, set, training, input, weight, method, neural
22	model, network, system, neural, input, neuron, data, cell, information, field
23	network, model, input, error, object, training, set, function, neural, image
24	network, system, function, point, error, input, learning, unit, model, neural
25	network, input, neural, model, set, system, training, data, learning, output

El último cluster se forma por los papers de aplicación médica, los cuales se distinguen por las palabras ‘cell’, ‘data’ e ‘image’. Está formado por los clusters: 13, 16, 22 y 23.

4) Código del algoritmo: solamente la sección donde se realiza el algoritmo k-means.

Se anexa en el archivo ‘KmeansOnlineJessVega.R’