Análisis de texto

##Búsqueda de artículos de PUBMED##

Instalación del paquete RISmed

```r
options(stringsAsFactors = F)
#install.packages(RISmed)
library(RISmed)
```

Se crea un archivo query para buscar artículos de PUBMED.

En este caso buscaremos query relacionados al área de medicina regenerativa, especialmente de andamios porosos para regeneración ósea; bone, scaffold, hydrogel, tissue, geometry.

```r
query_scaffold <- "\"bone\"[TIAB] AND \"scaffold\"[TIAB] AND \"hydrogel\"[TIAB] OR (\"polymer\"[TIAB] AND \"tissu
e\"[TIAB] AND \"geometry\"[TIAB])"
```

```r
search_query <-EUtilsSummary(query_scaffold)
```

```r
summary(search_query)
```

```
Query:
"bone"[TIAB] AND "scaffold"[TIAB] AND "hydrogel"[TIAB] OR ("polymer"[TIAB] AND "tissue"[TIAB] AND "geometry"[TIA
B])

Result count:  765
```

```r
records<-EUtilsGet(search_query)
pubmed_data<-data.frame('Title'=ArticleTitle(records), 'Abstract'=AbstractText(records), 'PID'=ArticleId(record
s))
pubmed_data[1:3,c('Title','PID')]
```

1
2
3

3 rows | 1-1 of 2 columns

```r
pubmed_data[1:3,'Title']
```

```
[1] "a biomimetic biphasic osteochondral scaffold with layer-specific release of stem cell differentiation induce
rs for the reconstruction of osteochondral defects"
[2] "injectable and crosslinkable plga-based microribbons as 3d macroporous stem cell niche"
[3] "fabrication of three-dimensional alginate porous scaffold incorporated with decellularized cornu cervi panto
trichum particle for bone tissue engineering"
```

```r
pubmed_data[1:3,'PID']
```

```
[1] "32338462" "32338432" "32331103"
```

Se realiza un pre-procesamiento para quitar caracteres (. : , ; []) tanto en el título como en el abstract.

```r
pubmed_data$Title<-gsub(pattern = "\\.|:|,|;|\\[|\\]", replacement="", pubmed_data$Title)

pubmed_data$Abstract <- gsub(pattern="\\.|:|,|;|\\[|\\]", replacement="", pubmed_data$Abstract)
```

Y se pasa todo a minúsculas

```r
pubmed_data$Title <- tolower(pubmed_data$Title)
pubmed_data$Abstract <- tolower(pubmed_data$Abstract)
pubmed_data[1,]
```

1

1 row | 1-1 of 3 columns

```r
pubmed_data[1,1]
```

```
[1] "a biomimetic biphasic osteochondral scaffold with layer-specific release of stem cell differentiation induce
rs for the reconstruction of osteochondral defects"
```

```r
pubmed_data[1,2]
```

```
[1] "there is a great challenge in regenerating osteochondral defects because they involve lesions of both cartil
age and subchondral bone which have remarkable differences in their chemical compositions and biological lineages
thus considering the complicated requirements in osteochondral reconstruction a biomimetic biphasic osteochondral
scaffold (bbos) with the layer-specific release of stem cell differentiation inducers are developed the cartilage
regeneration layer (cartilage scaffold cs) in the bbos contains a hyaluronic acid hydrogel to mimic the compositi
on of cartilage which is mechanically enhanced by host-guest supramolecular units to control the release of karto
genin (kgn) additionally a 3d-printed hydroxyapatite (hap) scaffold releasing alendronate (aln) is employed as th
e bone-regeneration layer (bone scaffold bs) the two layers are bound by semi-immersion and could regulate the hi
erarchical targeted differentiation behavior of the stem cells compared to the drug-free scaffold the mscs in the
bbos could be promoted to differentiate into both chondrocytes and osteoblasts the in vivo results demonstrate th
e strong promotion of cartilage or bone regeneration in their respective layers it is expected that this bbos wit
h layer-specific inducer release can become a new strategy for osteochondral regeneration"
```

```r
pubmed_data[1,3]
```

```
[1] "32338462"
```

Después se obtienen las palabras contenidas en el abstract mediante la función strsplit.

```r
unlist(strsplit(pubmed_data$Abstract[1], " "))[1:10]
```

```
 [1] "there"         "is"            "a"             "great"
 [5] "challenge"     "in"            "regenerating"  "osteochondral"
 [9] "defects"       "because"
```

Se descartan los artículos con abstracts vacíos:

```r
which(pubmed_data$Abstract == "")
```

```
 [1]   9  12  13  20  21  31  36  43  44  48  53  56  64  65  72  77  78  80  88
[20]  94  96 102 103 106 112 124 126 129 131 134 145 153 155 156 168 170 181 189
[39] 199 201 219 224 227 229 245 261 265 268 279 281 293 301 323 325 341 342 376
[58] 406 473 486 595
```

Se hace un data frame para guardar las palabras más importantes del abstract:

```r

word_list <- c()
#Ciclo para todos los abstracts
for(i in 1:length(pubmed_data$Abstract)){
    aux_word <- unlist(strsplit(pubmed_data$Abstract[i], " "))
    if(length(aux_word) > 0){
        aux_list <- cbind(pubmed_data$PID[i], aux_word)
        word_list <- rbind(word_list, aux_list)
    }
}
colnames(word_list) <- c("PID","Word")
dim(word_list)
```

```
[1] 161329      2
```

```r
word_list[1:8,]
```

```
     PID        Word
[1,] "32338462" "there"
[2,] "32338462" "is"
[3,] "32338462" "a"
[4,] "32338462" "great"
[5,] "32338462" "challenge"
[6,] "32338462" "in"
[7,] "32338462" "regenerating"
[8,] "32338462" "osteochondral"
```

Luego se procura obtener las palabras más frecuentes después de eliminar las “stopwords” (palabras auxiliares como adverbios, pronombres, etc).

```r
install.packages("tm")
library(tm)
```

```r
stop_words <-stopwords(kind = "en")
stop_words
```

```
  [1] "i"          "me"         "my"         "myself"     "we"
  [6] "our"        "ours"       "ourselves"  "you"        "your"
 [11] "yours"      "yourself"   "yourselves" "he"         "him"
 [16] "his"        "himself"    "she"        "her"        "hers"
 [21] "herself"    "it"         "its"        "itself"     "they"
 [26] "them"       "their"      "theirs"     "themselves" "what"
 [31] "which"      "who"        "whom"       "this"       "that"
 [36] "these"      "those"      "am"         "is"         "are"
 [41] "was"        "were"       "be"         "been"       "being"
 [46] "have"       "has"        "had"        "having"     "do"
 [51] "does"       "did"        "doing"      "would"      "should"
 [56] "could"      "ought"      "i'm"        "you're"     "he's"
 [61] "she's"      "it's"       "we're"      "they're"    "i've"
 [66] "you've"     "we've"      "they've"    "i'd"        "you'd"
 [71] "he'd"       "she'd"      "we'd"       "they'd"     "i'll"
 [76] "you'll"     "he'll"      "she'll"     "we'll"      "they'll"
 [81] "isn't"      "aren't"     "wasn't"     "weren't"    "hasn't"
 [86] "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"
 [91] "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"
 [96] "cannot"     "couldn't"   "mustn't"    "let's"      "that's"
[101] "who's"      "what's"     "here's"     "there's"    "when's"
[106] "where's"    "why's"      "how's"      "a"          "an"
[111] "the"        "and"        "but"        "if"         "or"
[116] "because"    "as"         "until"      "while"      "of"
[121] "at"         "by"         "for"        "with"       "about"
[126] "against"    "between"    "into"       "through"    "during"
[131] "before"     "after"      "above"      "below"      "to"
[136] "from"       "up"         "down"       "in"         "out"
[141] "on"         "off"        "over"       "under"      "again"
[146] "further"    "then"       "once"       "here"       "there"
[151] "when"       "where"      "why"        "how"        "all"
[156] "any"        "both"       "each"       "few"        "more"
[161] "most"       "other"      "some"       "such"       "no"
[166] "nor"        "not"        "only"       "own"        "same"
[171] "so"         "than"       "too"        "very"
```

Se guardan los índices de estas palabras y que deben ser removidas para un mejor análisis:

```r
index_stop_word <- which(word_list[,2] %in% stop_words)
length(index_stop_word)
```

```
[1] 54817
```

```r
dim(word_list)
```

```
[1] 161329      2
```

```r
word_list <- word_list[-index_stop_word,]
dim(word_list)
```

```
[1] 106512      2
```

Entonces, las palabras más frecuentes son:

```r
sort(table(word_list[,2]), decreasing=T)[1:10]
```

```

       bone    hydrogel    scaffold       cells      tissue   scaffolds
       1789        1259        1200        1152        1143         931
       cell   hydrogels       study engineering
        904         529         527         520
```

Después queda quitar las palabras repetidas en un mismo abstract:

```r
word_df <- data.frame(PID=as.numeric(word_list[,1]), Word=word_list[,2], PIDWord=as.character(apply(word_list, 1,
paste, collapse="_")))

word_df[1:5,]
```

	PID <dbl>	Word <chr>	PIDWord <chr>
1	32338462	great	32338462_great
2	32338462	challenge	32338462_challenge
3	32338462	regenerating	32338462_regenerating
4	32338462	osteochondral	32338462_osteochondral
5	32338462	defects	32338462_defects

5 rows

```r
NA
```

```r
dup_index <- duplicated(word_df$PIDWord)
word_df$PIDWord[1:10]
```

```
 [1] "32338462_great"         "32338462_challenge"     "32338462_regenerating"
 [4] "32338462_osteochondral" "32338462_defects"       "32338462_involve"
 [7] "32338462_lesions"       "32338462_cartilage"     "32338462_subchondral"
[10] "32338462_bone"
```

```r
dup_index[1:10]
```

```
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```r
length(which(dup_index))
```

```
[1] 30344
```

```r
dim(word_df)
```

```
[1] 106512      3
```

```r
word_df <- word_df[-which(dup_index),]
dim(word_df)
```

```
[1] 76168     3
```

```r
sort(table(word_df$Word), decreasing=T)[1:50]
```

```

           bone        scaffold          tissue        hydrogel           cells
            548             536             529             503             451
           cell           study     engineering       scaffolds           using
            373             371             327             323             279
        results             can            stem    regeneration            used
            266             260             253             252             246
      potential      properties           vitro          growth differentiation
            240             238             237             221             219
         showed       formation     mesenchymal          matrix      mechanical
            214             207             198             192             192
          human       hydrogels        compared            vivo            also
            188             185             184             181             177
        polymer   proliferation   significantly          within             new
            169             169             165             165             163
     osteogenic        collagen    demonstrated        geometry         however
            163             160             158             154             149
          weeks         culture       promising          repair       developed
            149             147             144             143             141
      different           novel       increased           model          system
            141             139             138             137             137
```

Tratando de filtrar un poco más la información, se puede optar por buscar palabras adicionales, por ejemplo, “printing”, refiriendose a los andamios que se imprimen mediante bioimpresoras 3D:

```r
word_df <- word_df[order(word_df$PID, decreasing=T),]
index_printing <- which(word_df$Word %in% c("printing"))
length(index_printing)
```

```
[1] 35
```

```r
word_df[index_printing[6:10], c("PID","Word")]
```

	PID <dbl>	Word <chr>
10581	31081613	printing
11612	31026858	printing
12178	30921360	printing
14001	30791603	printing
35994	30603476	printing

5 rows

```r
pubmed_data$Title[which(pubmed_data$PID == "31081613")]
```

```
[1] "multiscale porosity in compressible cryogenically 3d printed gels for bone tissue engineering"
```

De esta manera encontramos un artículo que podría ser de mucha utilidad para el tema relacionado, que de otra manera, ni siquiera nos enteraríamos de que existe.

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Cmd+Option+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Cmd+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.