Actividad_3

Análisis de biología computacional

Dra. Yocanxóchitl Perfecto Avalos

Integrantes: José de Jesús Gutiérrez Aldrete, Oscar Miranda Escalante, Luis Humberto Sánchez Vaca, José Gerardo Villanueva Barceló, Ana Ivette Farías Rodríguez

Parte 1

Primero hay que instalar y cargar el paquete "RISmed". También definimos la opción para no se lean los strings como factores.

```
options(stringsAsFactors = F)
#Cargar paquete
install.packages("RISmed")
```

```
##
## The downloaded binary packages are in
##  /var/folders/j3/89_pjybx40s_g_w578pj92hw0000gn/T//Rtmpq35YSe/downloaded_packages
```

```
library(RISmed)
```

Después, hacer query ""colon"[TIAB] AND "cancer"[TIAB] AND "young"[TIAB] AND "incidence"[TIAB] AND ("mutation"[TIAB] OR "treatment"[TIAB] AND "hereditary"[TIAB] OR "apc"[TIAB] OR "TP53"[TIAB] OR "KRAS"[TIAB] OR "pik3ca"[TIAB] OR "fbxw7"[TIAB] OR "smaD4"[TIAB] OR "tcf7l2"[TIAB]OR "nras"[TIAB] OR "early"[TIAB])"

```
#Hacer query
query_colon <- "\"colon\"[TIAB] AND \"cancer\"[TIAB] AND \"young\"[TIAB] AND \"incidence\"[TIAB] AND (\

#Instrucción para crear un query
search_query <- EUtilsSummary(query_colon)

#Hacer un resumen de la búsqueda
summary(search_query)
```

```
## Query:
## "colon"[TIAB] AND "cancer"[TIAB] AND "young"[TIAB] AND "incidence"[TIAB] AND ("mutation"[TIAB] OR "t
##
## Result count:  78
```

Obtener datos de los artículos

```r
#Ejecutar la búsqueda para obtener los datos de PUBMED
records <- EUtilsGet(search_query)

#Con esta instrucción obtenemos el título, abstract y el PUBMED ID
pubmed_data <-data.frame('Title'=ArticleTitle(records), 'Abstract'=AbstractText(records), 'PID'=Article
dim(pubmed_data)
```

```
## [1] 78  3
```

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

```
##
## 1 Molecular Aspects of Colorectal Adenomas: The Interplay among Microenvironment, Oxidative Stress, a
##   Abstract      PID
## 1          32258104
```

```r
pubmed_data[1:3,c("Title","PID")]
```

```
##
## 1                     Molecular Aspects of Colorectal Adenomas: The Interplay among Microenvironment
## 2 Clinical characteristics and a rising incidence of early-onset colorectal cancer in a nationwide co
## 3                                        International incidence trends in early- and late-onset colore
##        PID
## 1 32258104
## 2 32234586
## 3 32173775
```

Preprocesar datos

```r
#Eliminar algunos caracteres del título y del abstract
pubmed_data$Title <- gsub(pattern="\\.|:|,|;|\\[|\\]", replacement="", pubmed_data$Title)
pubmed_data$Abstract <- gsub(pattern="\\.|:|,|;|\\[|\\]", replacement="", pubmed_data$Abstract)

#Pasar todo a minúsculas
pubmed_data$Title <- tolower(pubmed_data$Title)
pubmed_data$Abstract <- tolower(pubmed_data$Abstract)
pubmed_data[1,]
```

```
##
## 1 molecular aspects of colorectal adenomas the interplay among microenvironment oxidative stress and
##   Abstract      PID
## 1          32258104
```

```r
length(pubmed_data$Title)
```

```
## [1] 78
```

Obtener las palabras del abstract

```r
#Con strsplit podemos separar las palabras por espacio
unlist(strsplit(pubmed_data$Abstract[1], " "))[1:10]
```

```
##  [1] NA NA NA NA NA NA NA NA NA NA
```

Problemas con algunos abstracts

```r
#Artículos de los cuales no se obtiene el abstract
which(pubmed_data$Abstract == "")
```

```
## [1]  1  5 21
```

Obtener las palabras del abstract

```r
#Obtener las palabras en un data frame junto con el PUBMED ID
# data frame para guardar las palabras
word_list <- c()
#Ciclo para todos los abstracts
for(i in 1:length(pubmed_data$Abstract)){
    #Obtener las palabras como vector en lugar de lista
    aux_word <- unlist(strsplit(pubmed_data$Abstract[i], " "))
    #Si el abstract tiene palabras
    if(length(aux_word) > 0){
        #Se juntan las palabras y el PUBMED ID
        aux_list <- cbind(pubmed_data$PID[i], aux_word)
        #Se pega este data frame auxiliar al que guarda todo
        word_list <- rbind(word_list, aux_list)
    }}
colnames(word_list) <- c("PID","Word")
dim(word_list)
```

```
## [1] 22091     2
```

```r
word_list[1:5,]
```

```
##      PID        Word
## [1,] "32234586" "background"
## [2,] "32234586" "the"
## [3,] "32234586" "incidence"
## [4,] "32234586" "of"
## [5,] "32234586" "early-onset"
```

Quitar “stopwords”

```r
#Cargar paquete tm
#install.packages("tm")
library(tm)
```

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

```r
#Obtener stopwords en ingles
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"
```

```r
#Palabras que son stopwords
index_stop_word <- which(word_list[,2] %in% stop_words)
length(index_stop_word)
```

```
## [1] 7640
```

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

```
## [1] 14451     2
```

Palabras más frecuentes

```
#Instrucción table y sort para obtener una tabla y ordenarla
sort(table(word_list[,2]), decreasing=T)[1:10]
```

```
##
##     cancer   patients  incidence        crc colorectal        age      young
##        403        313        232        205        185        180        168
##      colon      years      among
##        147        133        128
```

Dejar una combinación palabra-documento

```
#Crear un data frame y agregar una columna con la combinación PID_palabra
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        Word              PIDWord
## 1 32234586  background  32234586_background
## 2 32234586   incidence   32234586_incidence
## 3 32234586 early-onset 32234586_early-onset
## 4 32234586  colorectal  32234586_colorectal
## 5 32234586      cancer      32234586_cancer
```

Quitar repetidos usando duplicated

```
#Obtener duplicados
dup_index <- duplicated(word_df$PIDWord)
word_df$PIDWord[1:10]
```

```
##  [1] "32234586_background"  "32234586_incidence"   "32234586_early-onset"
##  [4] "32234586_colorectal"  "32234586_cancer"      "32234586_(eocrc)"
##  [7] "32234586_reported"    "32234586_increase"    "32234586_patients"
## [10] "32234586_eocrc"
```

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

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

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

```
## [1] 5842
```

Quitar duplicados

```
#Quitar duplicados
dim(word_df)
```

```
## [1] 14451     3
```

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

```
## [1] 8609    3
```

Palabras más frecuentes

```r
#Instrucción table y sort para obtener una tabla y ordenarla
sort(table(word_df$Word), decreasing=T)[1:10]
```

```
##
##  incidence     cancer      young      colon colorectal        age   patients
##         73         71         66         64         57         55         53
##      early      years      study
##         42         42         35
```

-Revisar que hayas obtenido la misma cantidad de artículos en la búsqueda manual como en la hecha con R: *La búsqueda manual en PubliMed arrojo una cantidad de 81 artículos. La diferencia se debe a que la información no ha sido actualizada para RISmed. Probablemente se hayan omitido los 3 artículos más recientes.*

-Usando los resultados en R, obtén las 10 palabras más frecuentes y da una interpretación de la enfermedad. Sólo considera palabras relacionadas a medicina, es decir evita artículos, adverbios, adjetivos, etc. *Estudios demuestran que la incidencia en edades tempranas del cáncer de colon/colorrectal ha aumentado, acrecentando la cantidad de pacientes jóvenes.*