**INTEGRANTES:**

**PROYECTO DE PROCESOS**

**CASTRO ALEXANDER**

**CHEA BETHSAIDA**

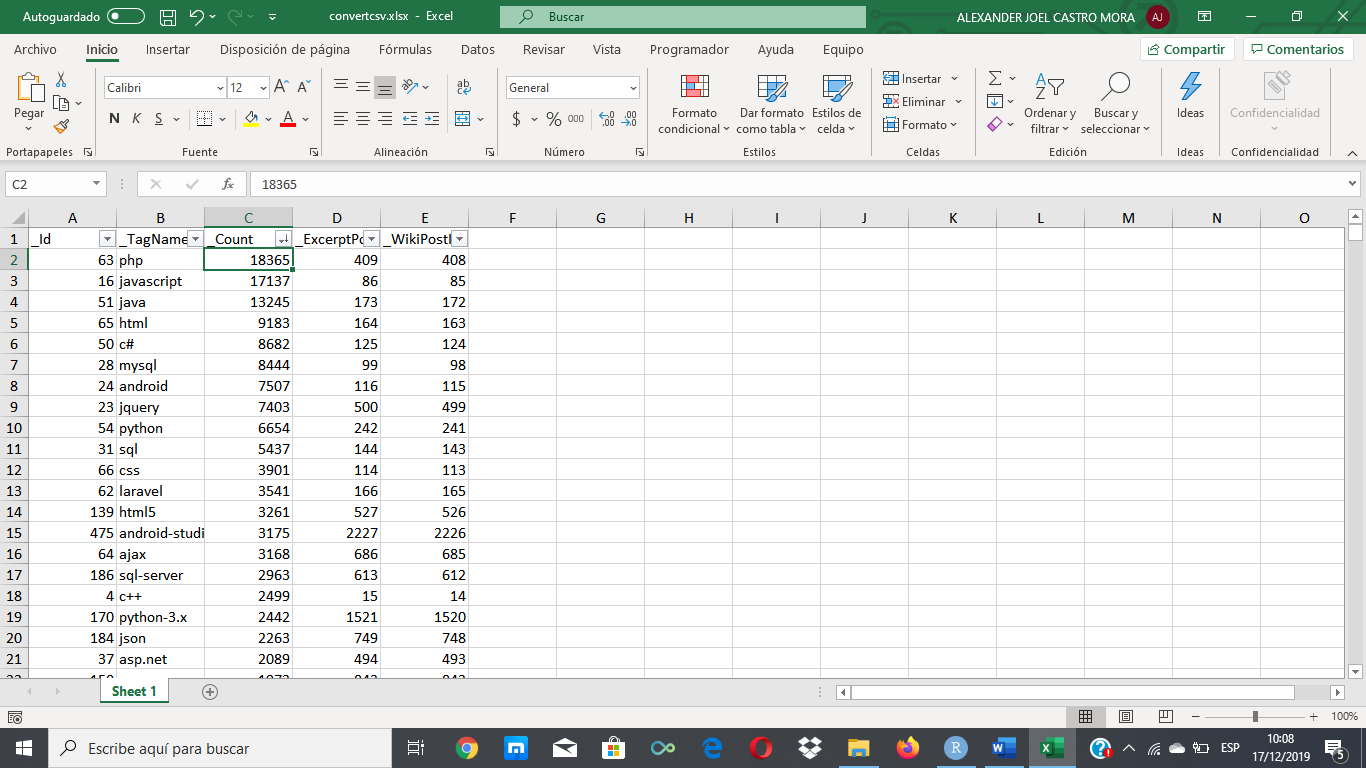
**TEMA: NUBE DE PALABRAS**

**CREACION DE NUBE DE PALABRAS**

Para la realización del siguiente proyecto se necesito el uso de:

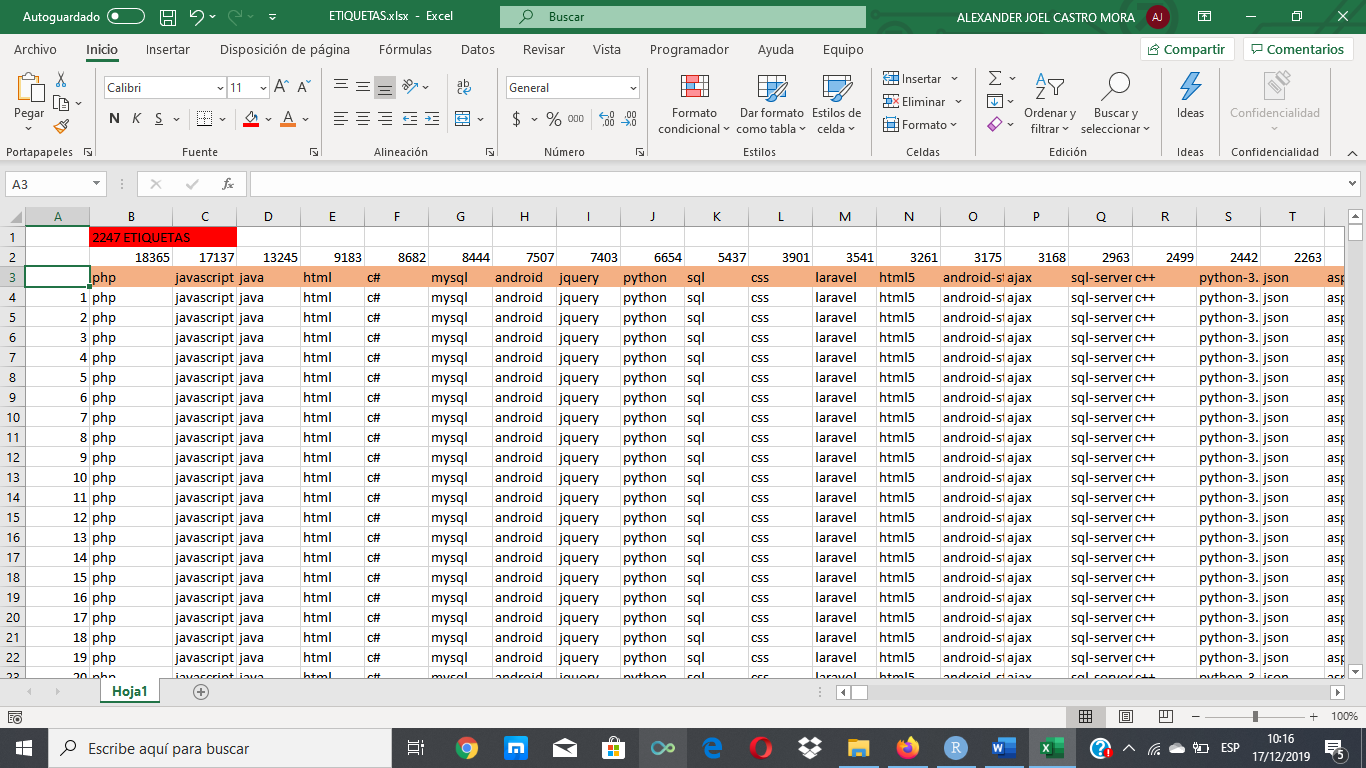
* R Studio (Software para minería de Datos)
* Los datos adquiridos por la plataforma Stack Exchange Data Dump, que contiene los Tags de todos los usuarios de Stack OverFlow.

El documento obtenido era de formato XML por lo cual se prefirió transformarlo a otro tipo de formato como lo es el CSV Y XLSX en nuestro caso usamos el ultimo formato mencionado para una mejor organización y comprensión del documento.



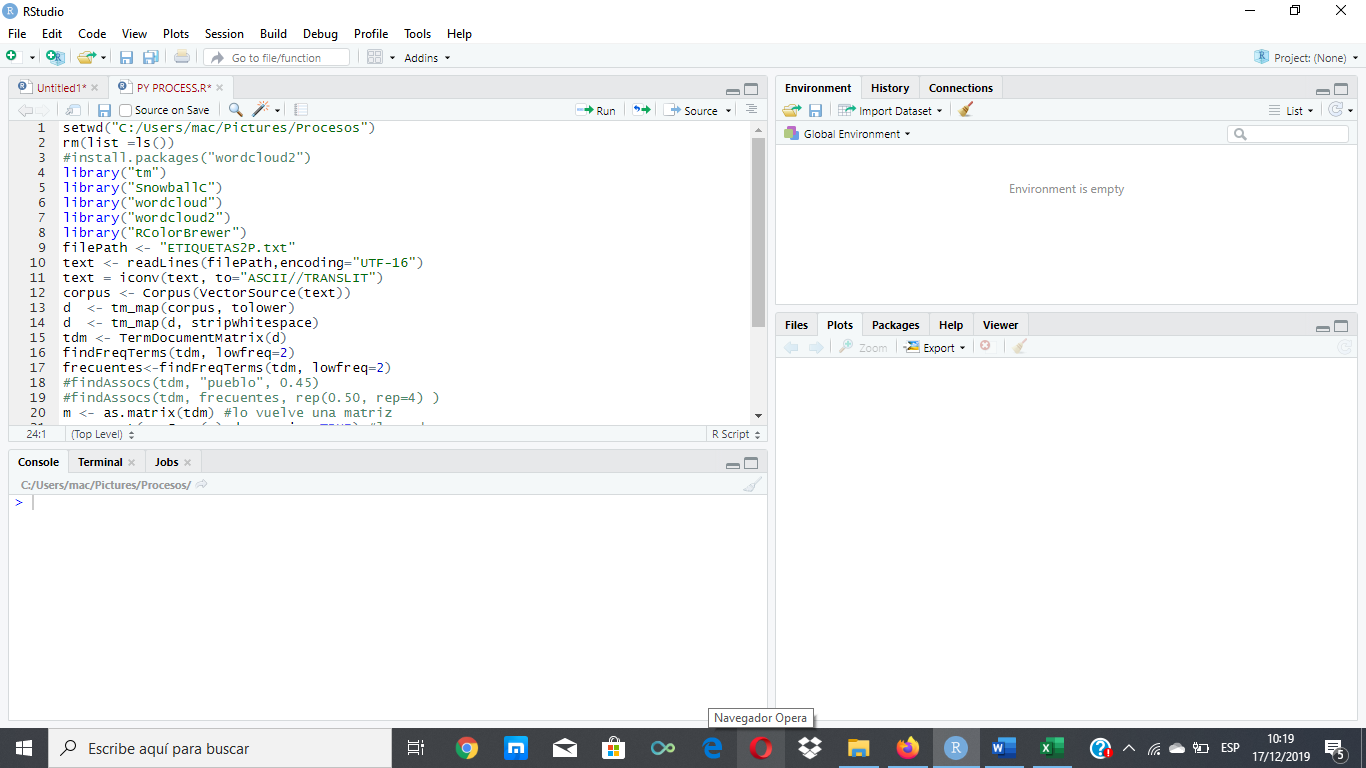
Esta es la vista previa del archivo del cual para la creación de la Nube solo nos va a interesar las columnas de TagName y Count, que son las que nos indican el nombre de la etiqueta y que tan frecuente es en la plataforma.

Debido a los inconvenientes de poder relacionar el TagName con la Concurrencia se opto por crear una nueva tabla con las mismas etiquetas replicadas según su frecuencia, lo que nos dejaría un archivo como el siguiente.



Una vez obtenida esta tabla, podemos empezar a utilizar el entorno de trabajo de R Studio.

Tendremos una interfaz como la siguiente donde creamos un nuevo Script donde ingresaremos el código para generar nuestra nube.

**Nota:** Un punto importante es que la tabla anterior la convertimos a un formato \*.txt para un análisis mucho mas rápido por R aunque también puede leer archivos de Excel, decidimos trabajar en \*.txt.

Empezamos especificando la ruta donde se va a encontrar nuestro archivo, con el cual vamos a trabajar, mediante las siguientes líneas.

*setwd("C:/Users/mac/Pictures/Procesos")*

*rm(list =ls())*

Procedemos con la instalación de las librerías que sean necesarias mediante la siguiente línea y así sucesivamente con todas las librerías.

*install.packages("wordcloud2")*

Continuamos con la carga de estas librerías, mostraremos cual serán las necesarias.

*library("tm")*

*library("SnowballC")*

*library("wordcloud")*

*library("wordcloud2")*

*library("RColorBrewer")*

Como siguiente seleccionamos el archivo y procedemos con la lectura.

*filePath <- "ETIQUETAS2P.txt"*

*text <- readLines(filePath,encoding="UTF-16")*

Convertimos el archivo en un vector de caracteres y lo cargamos en un corpus.

*text = iconv(text, to="ASCII//TRANSLIT")*

*corpus <- Corpus(VectorSource(text))*

Realizamos una denominada Limpieza, la cual va a consistir en que todas las palabras del documento se conviertan a minúsculas y quitamos espacios en blanco de más que se encuentren.

*d <- tm\_map(corpus, tolower)*

*d <- tm\_map(d, stripWhitespace)*

Una vez limpio nuestro Corpus lo pasamos a un formato de base de datos seguido de definir hasta el mínimo de frecuencia que queremos tomar en cuenta.

*tdm <- TermDocumentMatrix(d)*

*findFreqTerms(tdm, lowfreq=2)*

*frecuentes<-findFreqTerms(tdm, lowfreq=2)*

Procedemos con comandos de orden y formatos.

*m <- as.matrix(tdm) #lo vuelve una matriz*

*v <- sort(rowSums(m),decreasing=TRUE) #lo ordena y suma*

*df <- data.frame(word = names(v),freq=v) #lo nombra y le da formato de data.frame*

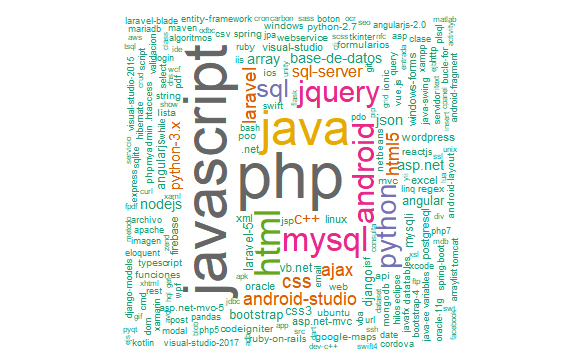
Realizado todo lo anterior ya tenemos un archivo el cual podremos transformarlo a una nube de palabras con la función “**WordCloud**”.

Con la función max.words podemos definir cuantas palabras queremos mostrar en este caso serán 2247 que es la cantidad de etiquetas que tenemos.

*wordcloud(words = df$word, freq = df$freq, min.freq = 2,*

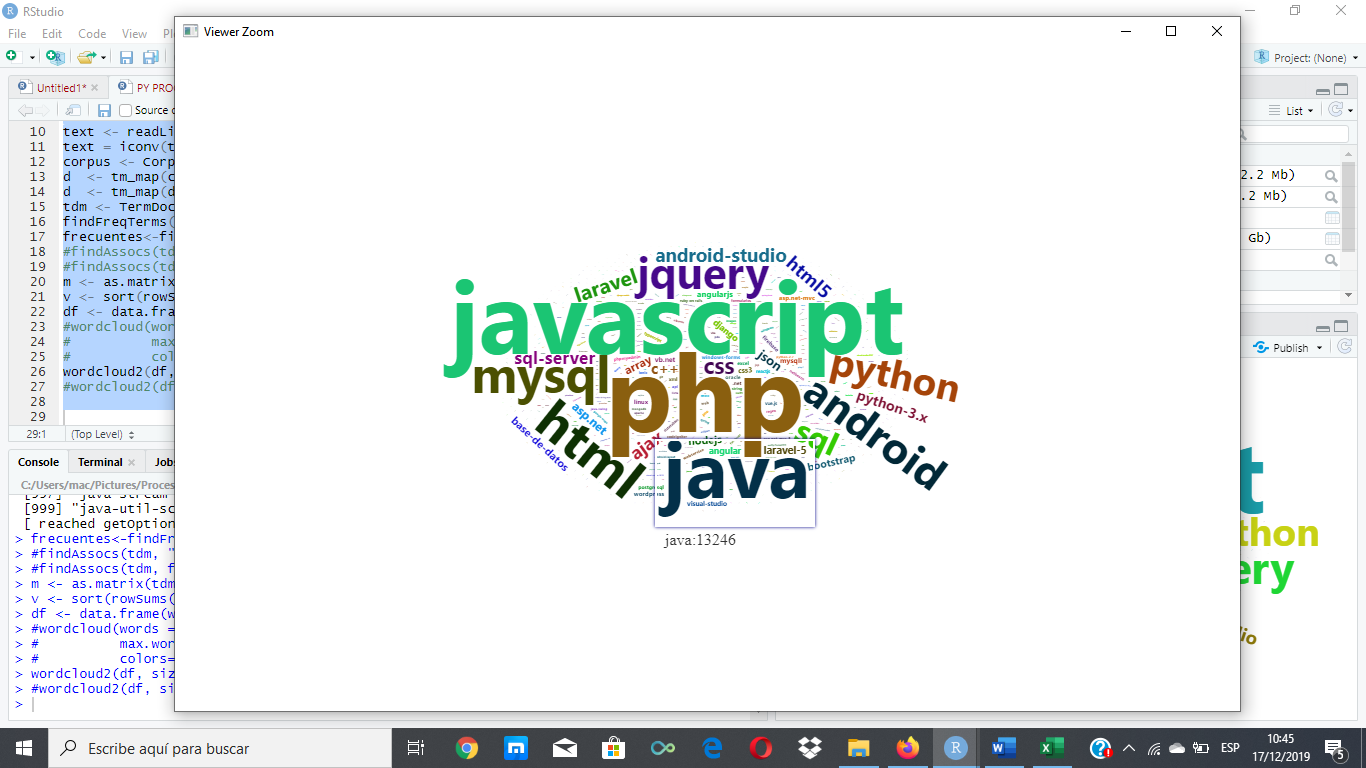
*max.words=2247, random.order=FALSE, rot.per=0.35,*

*colors=brewer.pal(8, "Dark2"))*

Esto se vería de la siguiente forma.

También podemos usar una variante de la WordCloud que es la siguiente.

*wordcloud2(df, size=0.6)*

Siendo un poco más visual e interactiva, ya que podemos seleccionar la etiquetas y nos mostrara cual es la frecuencia de esta.