MINERÍA DE TEXTO Y ANÁLISIS DE PALABRAS

FCEA FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN
IESTA 80 AÑOS INSTITUTO DE ESTADÍSTICA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

GRUPO BEM

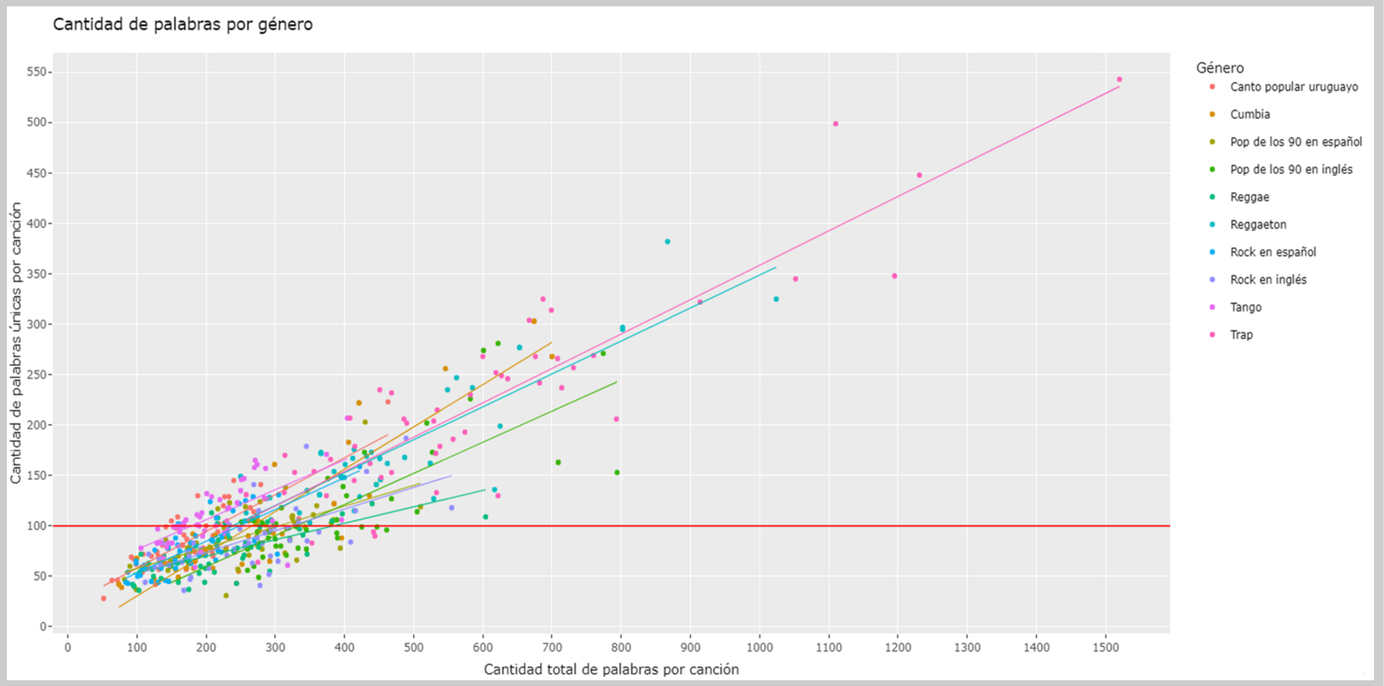
Bruno Lucchini, Emiliano Cabrera y Martín Osorio

ALGO DE CON*TEXTO*

TRABAJAMOS CON ESTE TIPO DE DATOS UTILIZANDO **LETRAS DE CANCIONES DE DISTINTOS GÉNEROS MUSICALES**.

PARA SELECCIONAR LA MUESTRA, ELEGIMOS **10 ESTILOS VARIADOS** Y DESCARGAMOS PARA CADA UNO LAS **LETRAS DE 50 CANCIONES DE MANERA ALEATORIA**, BASÁNDONOS EN RANKINGS DE POPULARIDAD.

CONSIDERANDO QUE EL TEXTO ES UN TIPO DE DATOS MUY DESORDENADO, RECURRIMOS A LIBRERÍAS DE R ESPECÍFICAS PARA SU MANEJO, TALES COMO **TIDYVERSE, TM, TOKENIZERS Y STOPWORDS**, LO QUE NOS PERMITIÓ IMPORTAR LOS DATOS EN RSTUDIO DESDE FICHEROS .TXT Y ESTRUCTURARLOS A GUSTO PARA REALIZAR EL ANÁLISIS CORRESPONDIENTE.



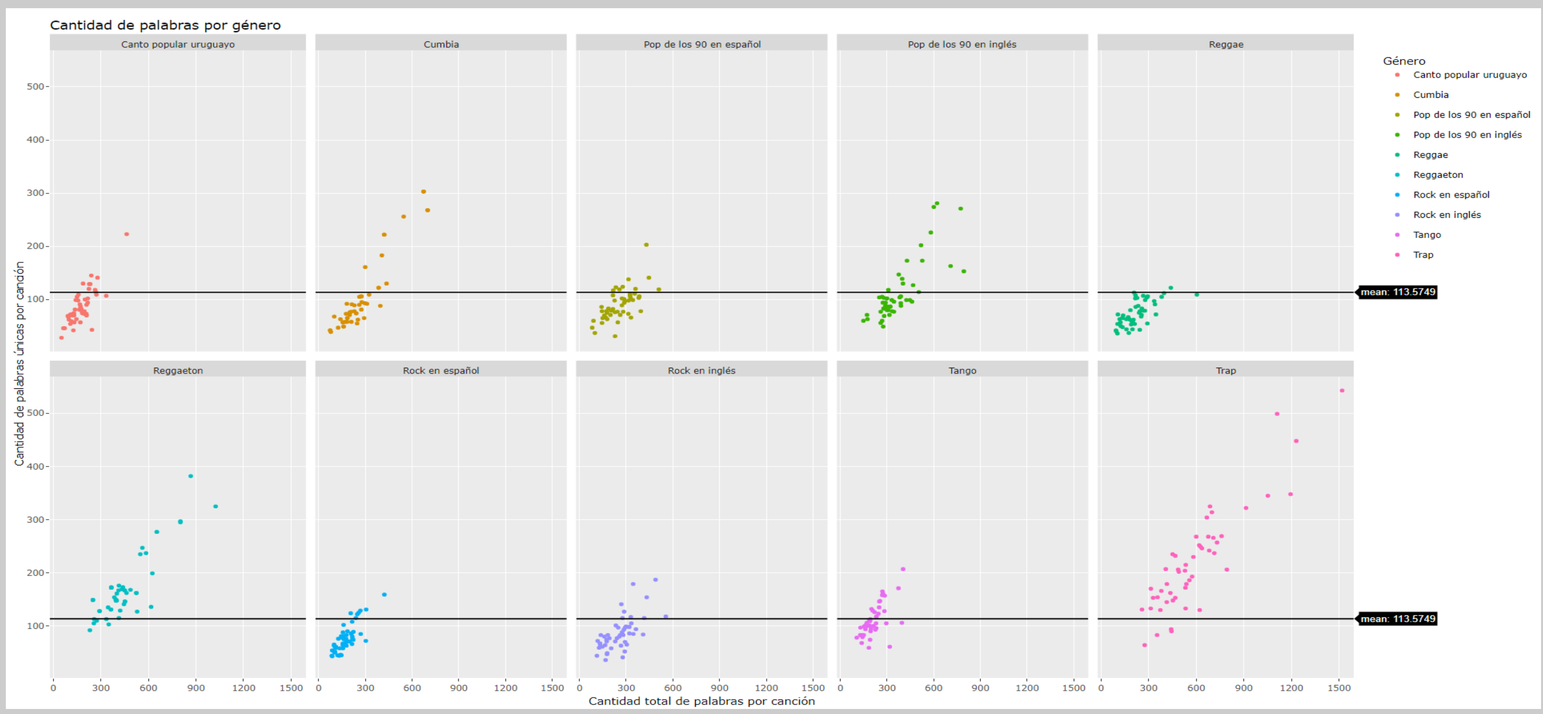
OBJETIVO

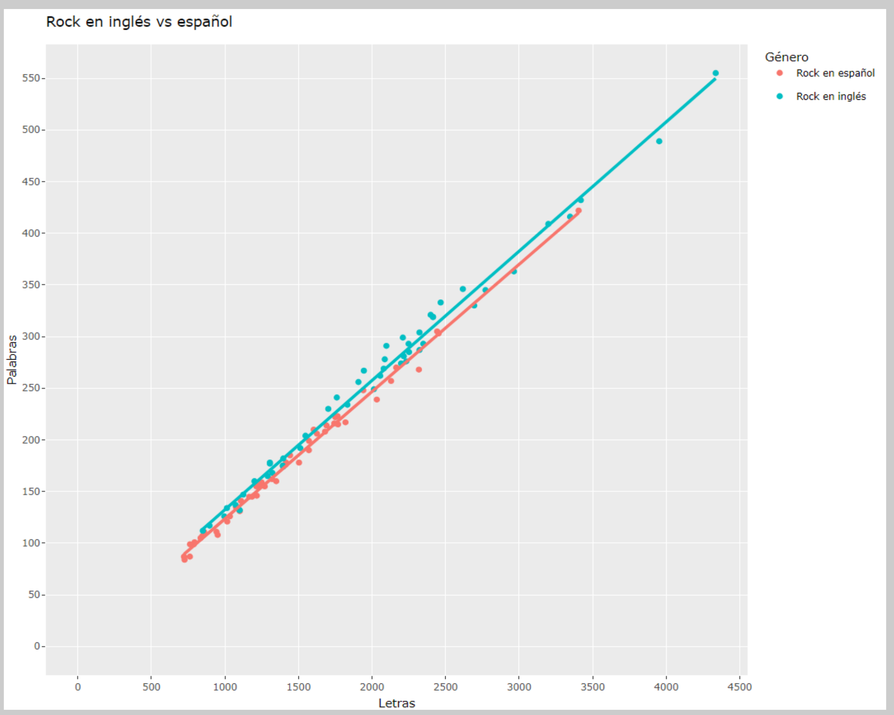
NUESTRA META PRINCIPAL FUE **OBSERVAR DIFERENCIAS EN CANTIDAD DE PALABRAS EMPLEADAS POR GÉNERO**.

TAMBIÉN NOS INTERESABA CONOCER LA **RELACIÓN ENTRE LA CANTIDAD DE PALABRAS TOTALES Y PALABRAS ÚNICAS** PARA ENTENDER LAS CARACTERÍSTICAS DE CADA ESTILO.

DESCUBRIMOS QUE, EN PROMEDIO, **EL TANGO ES EL QUE EMPLEA MÁS VOCABULARIO**: POR CADA 100 PALABRAS UTILIZADAS 54 SON ÚNICAS; MIENTRAS QUE **EL POP EN INGLÉS ES EL QUE MÁS PALABRAS REPITE**, GENERANDO SOLO 30 PALABRAS ÚNICAS POR CADA 100 PALABRAS EMPLEADAS.

TAMBIÉN REALIZAMOS UN ANÁLISIS COMPARATIVO ENTRE EL **INGLÉS Y EL ESPAÑOL**, PUES QUERÍAMOS CONOCER **LA CANTIDAD DE PALABRAS EMPLEADAS EN FUNCIÓN DE LA LONGITUD DE LAS** MISMAS, Y, AUNQUE YA LO ASUMÍAMOS, CONFIRMAMOS QUE **EN INGLÉS SE TIENDE A UTILIZAR MÁS PALABRAS** (UN PROMEDIO DE 259 CONTRA 176), MIENTRAS QUE **EN ESPAÑOL LAS PALABRAS SON MÁS LARGAS**: EN PROMEDIO 810 LETRAS CADA 100 PALABRAS, A DIFERENCIA DEL INGLÉS, CUYO PROMEDIO ES DE 776.





NUBES DE PALABRAS

SON UN TIPO DE GRÁFICO DISEÑADO PARTICULARMENTE PARA **MOSTRAR DATOS NOMINALES, DONDE LA FRECUENCIA ES REPRESENTADA POR EL TAMAÑO DE LA VARIABLE,** EN ESTE CASO LA PALABRA.

CREAMOS UNA NUBE POR GÉNERO PARA PODER MOSTRAR LAS PALABRAS MÁS FRECUENTES EN CADA ESTILO MUSICAL.

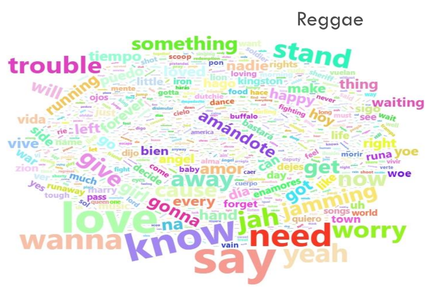
```
#Creamos dataframe con las palabras de todas las canciones del género


palabras <-c()
j=0;
for (i in canciones_Tango){
  j<-j+1
  agregar <-removeWords(palabras_Tango[[j]],stop)
  palabras <-c(palabras,agregar)
}
limpias_Tango <- table(palabras)
limpias_Tango <- data_frame(word = names(limpias_Tango),
                            count = as.numeric(limpias_Tango)
)


#Con el paquete wordcloud2 creamos la nube de palabras:

wordcloud_Tango <- wordcloud2(
  data = limpias_Tango,
  size = 40,
  color = 'random-light')


#Generamos el gráfico:
wordcloud_Tango
```

Tango

Canto popular uruguayo

Pop de los 90 en español

Reggaetón



Reggae

Cumbia

Pop de los 90 en inglés

Rock en inglés

Rock en español

Trap

CONCLUSIÓN

EXISTE UNA **TENDENCIA BASTANTE PAREJA ENTRE LOS DISTINTOS GÉNEROS** A UTILIZAR CIERTA CANTIDAD DE PALABRAS EN RELACIÓN CON EL VOCABULARIO EMPLEADO.

VIMOS QUÉ PALABRAS SON PROPIAS DE CADA GÉNERO Y CUÁLES SE REPITEN A TRAVÉS DE LOS DIFERENTES ESTILOS MUSICALES, **SIENDO EL AMOR UNA TEMÁTICA MODAL.**

ASIMISMO, **PUDIMOS DERRIBAR PREJUICIOS EN TORNO A ALGUNOS GÉNEROS** RESPECTO DE LA RIQUEZA LINGÜÍSTICA CONTENIDA EN SUS CANCIONES.