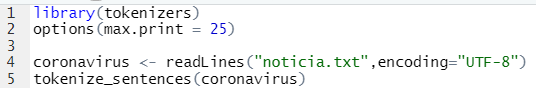
**TEXT MINING NOTICIAS**

***TÉCNICAS DE PRE-PROCESAMIENTO:***

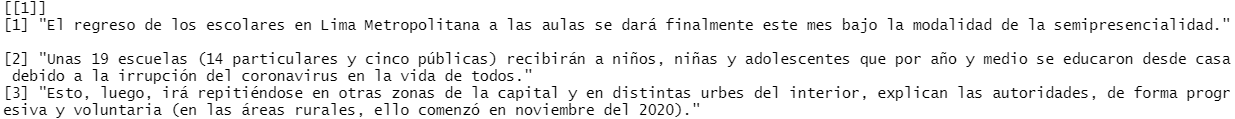
Se escogió una noticia referente a los acontecimientos en Perú sobre el Coronavirus, el cual se procedió a analizar y a realizar los siguientes puntos de pre-procesamiento:

1. **TOKENIZATION**

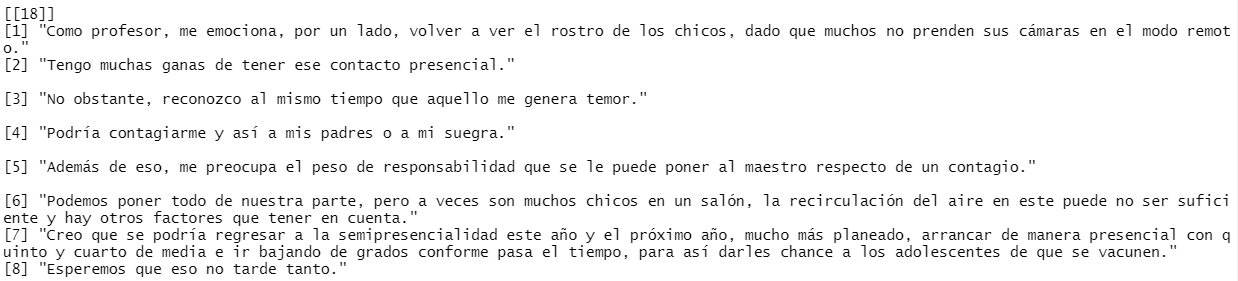
* **Tokenización por oración**



Al realizar la tokenización en la noticia se encontraron los párrafos encontrados y las oraciones encontradas por cada párrafo. Encontrando que nuestra noticia cuenta con 18 párrafos, en el 1er párrafo se identifico 3 oraciones y en el párrafo 18avo se identificó 8 oraciones.

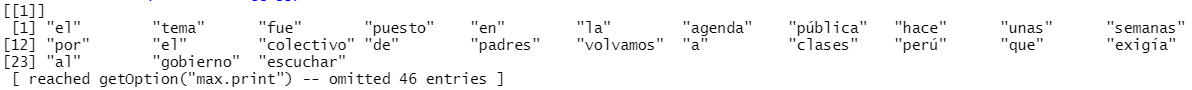


**…**



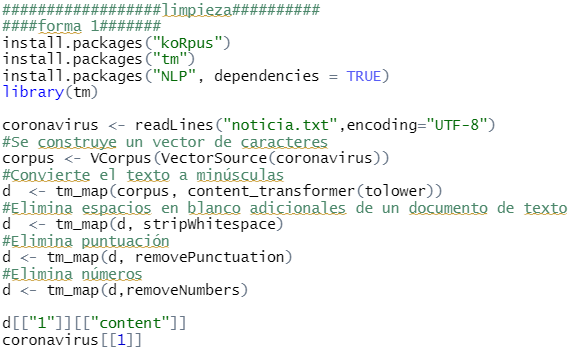
* **Tokenización por palabra**

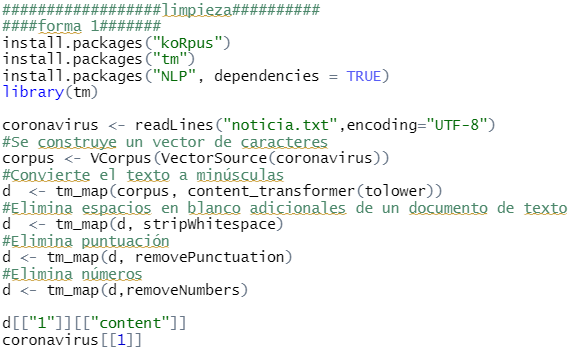




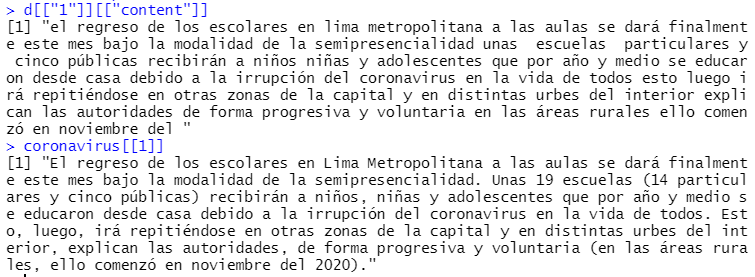
1. **NORMALIZATION**

* **Limpieza con la librería “tm”**



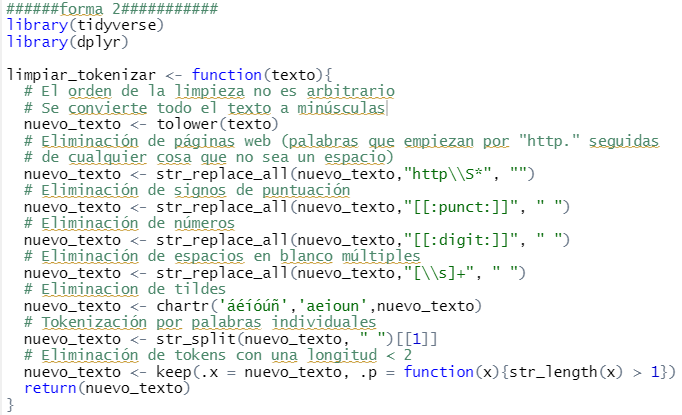


A continuación, se compara un texto limpiado, en está forma no se limpia las tildes ni los caracteres atípicos.

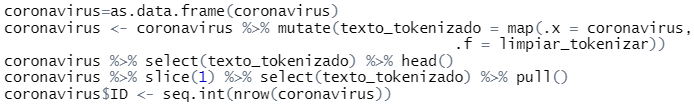


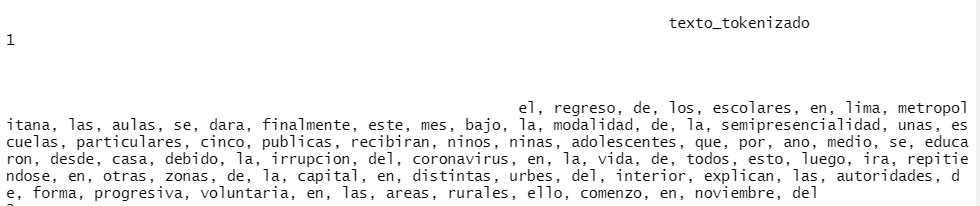
* **Limpieza con la librería “tidyverse” y “dplyr”**

Se crea una función de limpieza

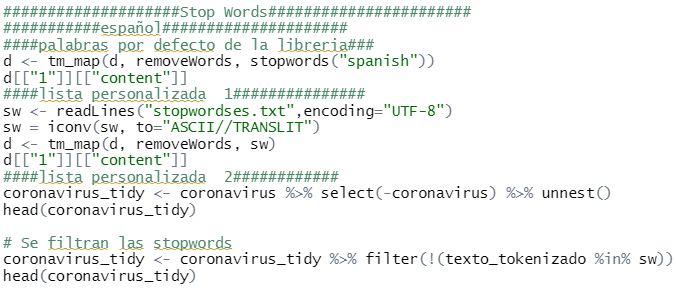


Con esta función a diferencia de la primera forma, aquí si se quita las tildes y los signos especiales como las “ñ”, elimina los conectores (a, y)

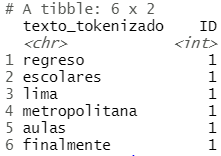
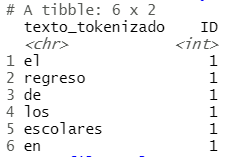




* **Stopwords**



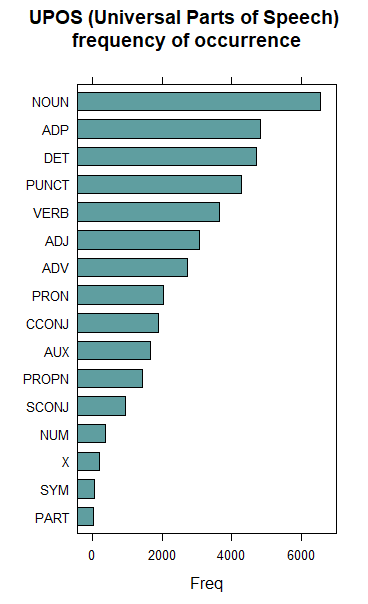
Aquí se realiza la limpieza de los stopwords, de parte de las palabras incluidas en la librería y desde una base personalizada, obteniendo las siguientes palabras:

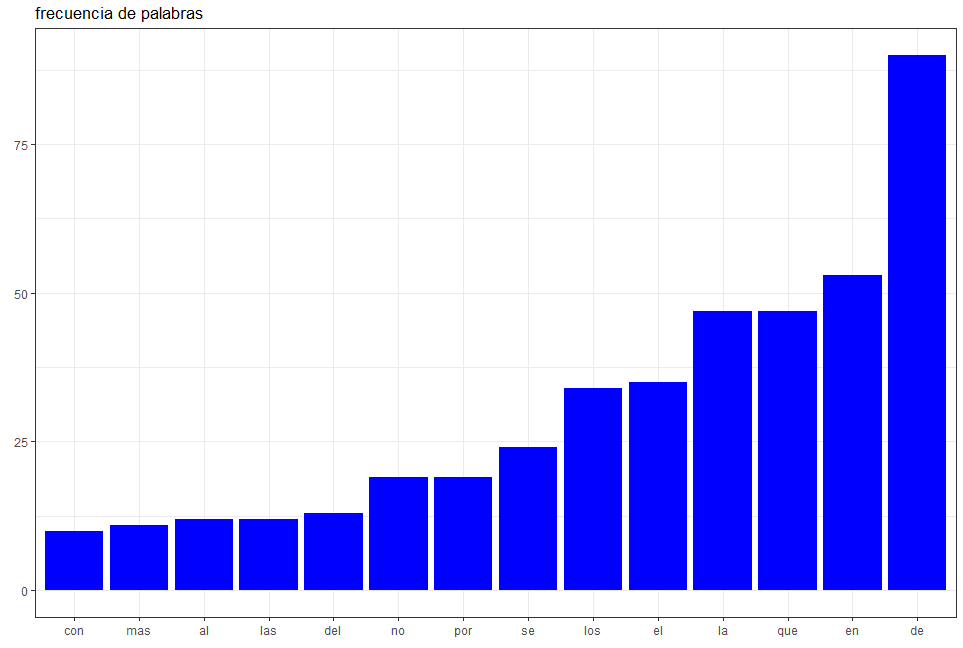
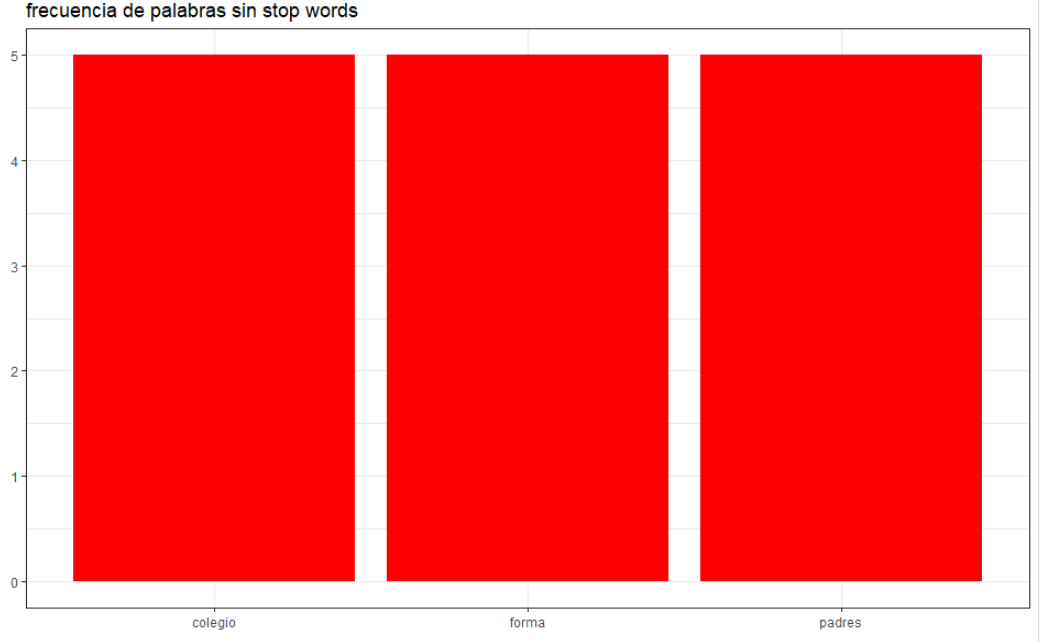


1. **STEMMING Y LEMMATIZATION**



* Utilizando la librería UDpipe yace en el hecho de que nuestro texto se encuentra ahora **anotado**. Se revisará las frecuencias sobre las etiquetas POS:





* Extracción de variables

