**Maestría de Ciencia de Datos**

**Análisis de Estadístico de Texto**

**Alumno: Juan Victor Gonzales Montoya**

1. **Análisis Exploratorio:**

Para iniciar al Análisis Exploratorio se implemento un grafico de nube de palabras para visualizar la frecuencia que tiene los tweets tanto de Daniel Urresti como del Ministerio del Interior de Perú.

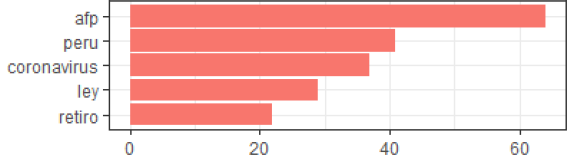
****

**Daniel Urresti**

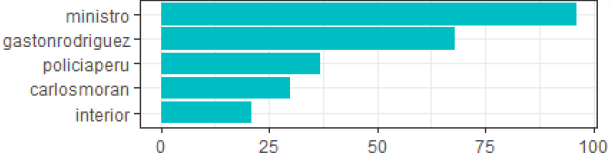
****

**Ministerio del Interior**

Como se puede apreciar en el caso de Daniel Urresti las palabras con mayor frecuencia son ***AFP, Perú, coronavirus*** entre otras, mientras que el caso del Ministerio del Interior las palabras más usadas son ***ministro, gastonrodrigues y policiaperu***, esto se puede corroborar en los siguientes gráficos de frecuencia de palabras:

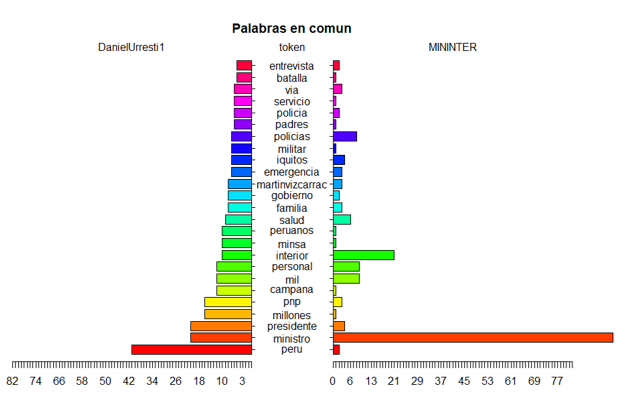


**Daniel Urresti**



**Ministerio del Interior**

Luego se hizo un grafico de las palabras en común entre ambos grupos de tweets el cual se muestra a continuación:



1. **Clasificación:**

A continuación, se normalizó eliminando los stopwords del texto y con el resultado se utilizo el modelamiento de la Maquina de Soporte Vectorial (SVM) y de Naive Bayes, ambos utilizando el indicador de TF\_IDF obteniendo los siguientes resultados:

|  |  |  |  |
| --- | --- | --- | --- |
| **Algoritmo** | **Precision** | **Recall** | **F1** |
| **SVM** | 0.7631 | 0.9892 | 0.8135 |
| **Naive Bayes** | 0.8362 | 0.9465 | 0.9032 |

Al verificar los resultados obtenidos en base al indicador de Precision, se puede constatar que el Modelo con el algoritmo de Naive Bayes fue superior al de SVM