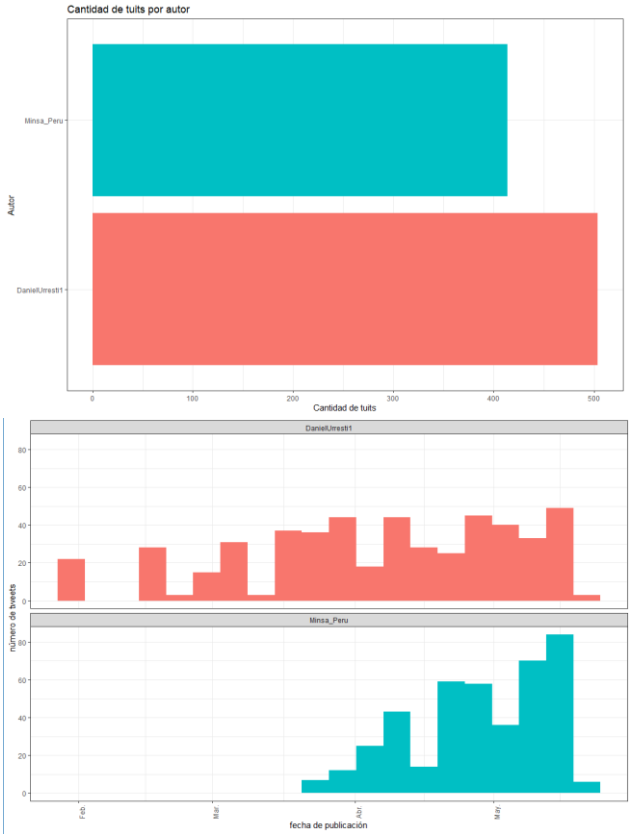
Tarea – Sesión 4

Rayan Figueroa

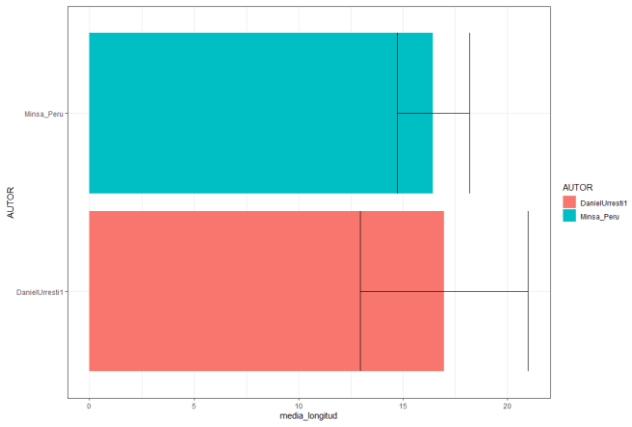
Análisis Exploratorio

1.- Cantidad de Tuits



Se tiene más tuits de DanielUrresti (504) que en el Minsa (414). Los tuits del Minsa inician a mediados de marzo, mientras que los de Urresti desde finales de enero.

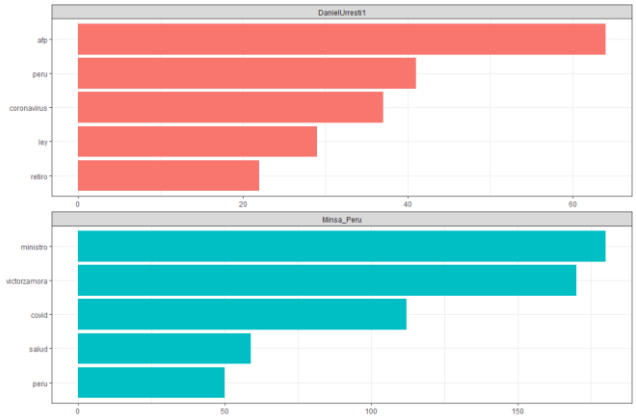
2.- Media de palabras



	AUTOR	media_longitud	sd_longitud
	<chr>	<dbl>	<dbl>
1	DanielUrresti1	17.0	4.02
2	Minsa_Peru	16.4	1.73

DanielUrresti y Minsa_Peru tienen una cantidad de palabras similar. Sin embargo, DanielUrresti tiene una mayor variedad en la longitud de palabras de cada tuit.

3.- Palabras más utilizadas



Nube de palabras:

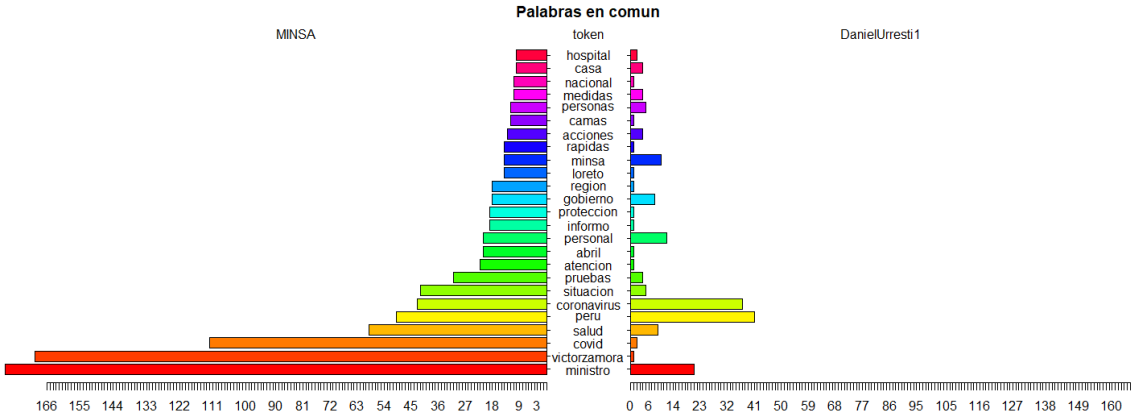
DanielUrresti1

Minsa

Las palabras más usadas por Urresti son afp, peru, coronavirus, ley, retiro. Refiriéndose al retiro de la afp.
Las palabras más usadas por el Minsa son: ministro, victorzamora, covid, salud, Perú. Refiriéndose al estado de emergencia del Covid.

4.- Palabras en común

El numero de palabras comunes entre Minsa y DanielUrresti es de 303.



En las 25 palabras mas comunes se observa que el minsa usa más las palabras ministro, victorzamora y covid y las palabras más usadas de DanielUrresti son Perú, coronavirus y ministro.

5.- Modelado

Se realizó el modelamiento con palabras limpias (normalizadas) es decir con limpieza y sin stopwords.

Se realizó dos modelos: SVM con corpus limpio y itdf y Naive Bayes Normalizado.

```
> cbind.data.frame(precision_SVM,precision_NB)
  precision_SVM precision_NB
1          0.95    0.9741935
> cbind.data.frame(recall_SVM,recall_NB)
  recall_SVM recall_NB
1  0.9382716 0.9320988
> cbind.data.frame(F1_SVM,F1_NB)
     F1_SVM     F1_NB
1 0.9440994 0.9526814
```

El modelo que tiene mejores resultados es el de Naive Bayes Normalizado con un accurancy de 97%. Si observamos el recall observamos que SVM es ligeramente mejor que Naive Bayes.
En conclusión, para este ejemplo seria mejor utilizar Naive Bayes