Intelligent Systems
Natural Language Processing

Simon Iavarone

January 2022

1 Problem

I decided to explore corpus from the Leipzig Corpora Collection. I chose to examine two corpus in Spanish, the news corpus for years 2018 and 2021. I took the smallest size of corpus, which is 10 thousands sentences as it enables fast computation of the results. One of the interesting features of these two corpus is their location in time. The data for one of them was collected before the Covid-19 pandemic hit the planet, while the data for the second one was collected during the pandemic. The datasets used for this project can be found at the following website : `https://wortschatz.uni-leipzig.de/en/download/Spanish#spa_news_2021`.

2 Experiments done

The code presenting all experiments done during this project can be found at `https://github.com/Simiavarone/IS_NLP`. In order to perform a comparative analysis of the two corpus previously mentioned, I analysed an histogram presenting headlines length as well as the dendogram for each corpus. These two plots didn't give very useful information. The computation of the distance between the documents was quite long when taking all the lines in the corpus into account. Therefore, this computation was done with only 20% of the lines present in each corpus. The lines selected for the computation were sampled randomly without replacement from the original corpus.

When examining the top features (with $n = 30$) for each corpus, more interesting information appears. It is also worth mentioning that I did the same experiments on the 100 thousands lines corpus that can be found at the same web address. This path didn't give more exciting results so the idea of using larger corpus was quickly abandoned.

3 Analysis of results

Figures 1 and 2 show the histograms of headline length for years 2018 and 2021 respectively. In these figures we observe a slight difference between these two time periods. There were more short headlines in 2018 than in 2021.

In figures 3 and 4, you can see the dendograms for years 2018 and 2021 respectively. These dendograms present an interesting feature of the data. For year 2018, the maximum distance between the headlines (above 500) is higher than for year 2021 (comprised between 400 and 500). An interpretation of this observation could be that as the pandemic rose in 2021, lots of articles were talking about it, thus reducing the distance between articles.

Finally, you can observe top and bottom features for years 2018 and 2021 respectively in figures 5 and 6. We can clearly notice that some tokens related to the Covid-19 pandemic appear in the top features. These tokens are : `salud`, `pandemia`, `covid-19` and `casos`.

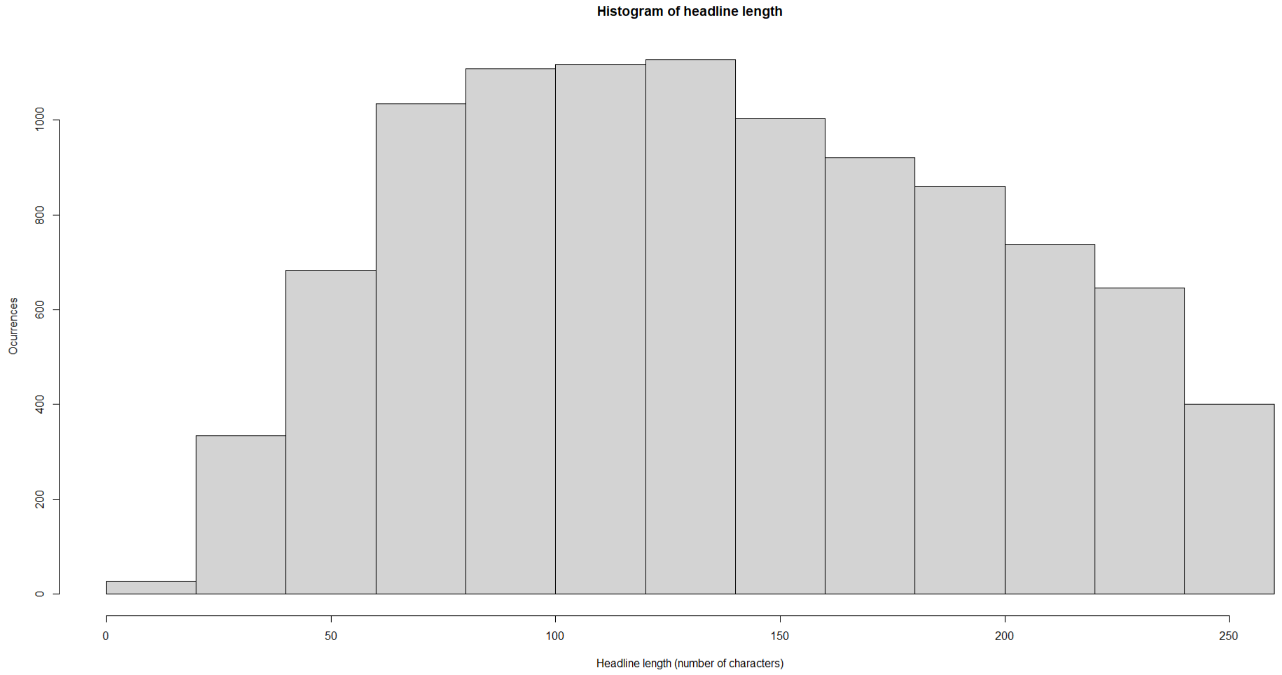


Figure 1 – Histogram of headline length, year 2018

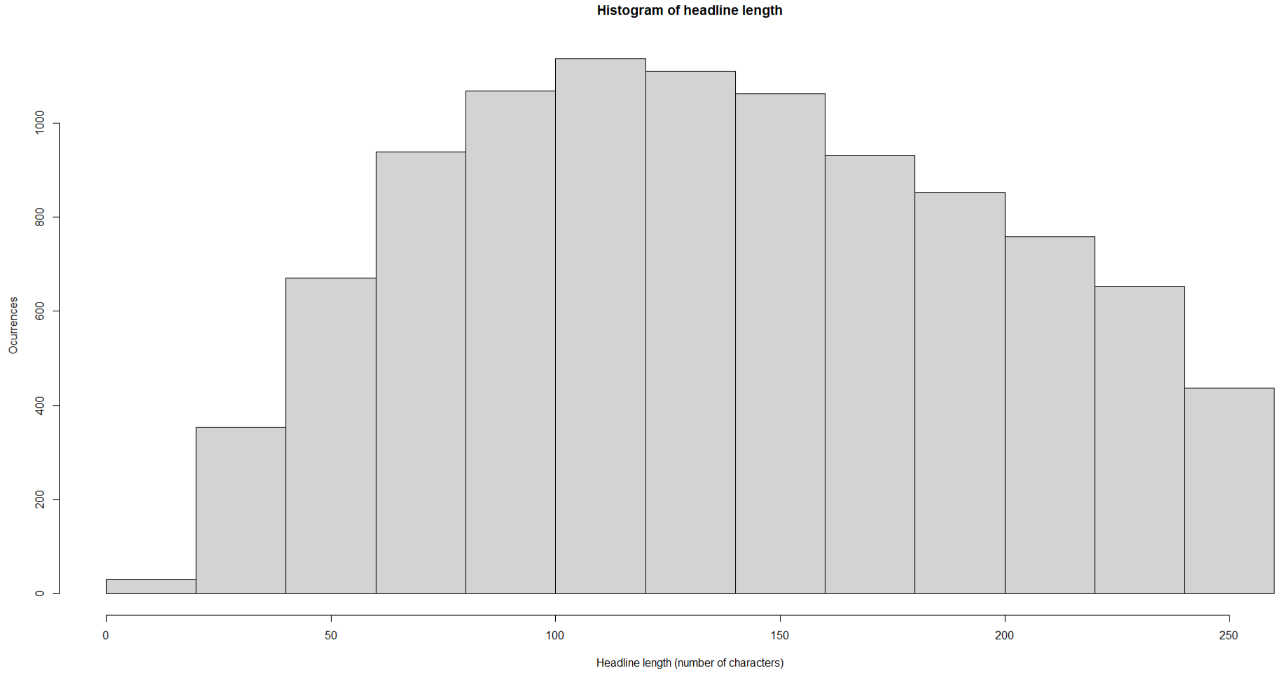


Figure 2 – Histogram of headline length, year 2021

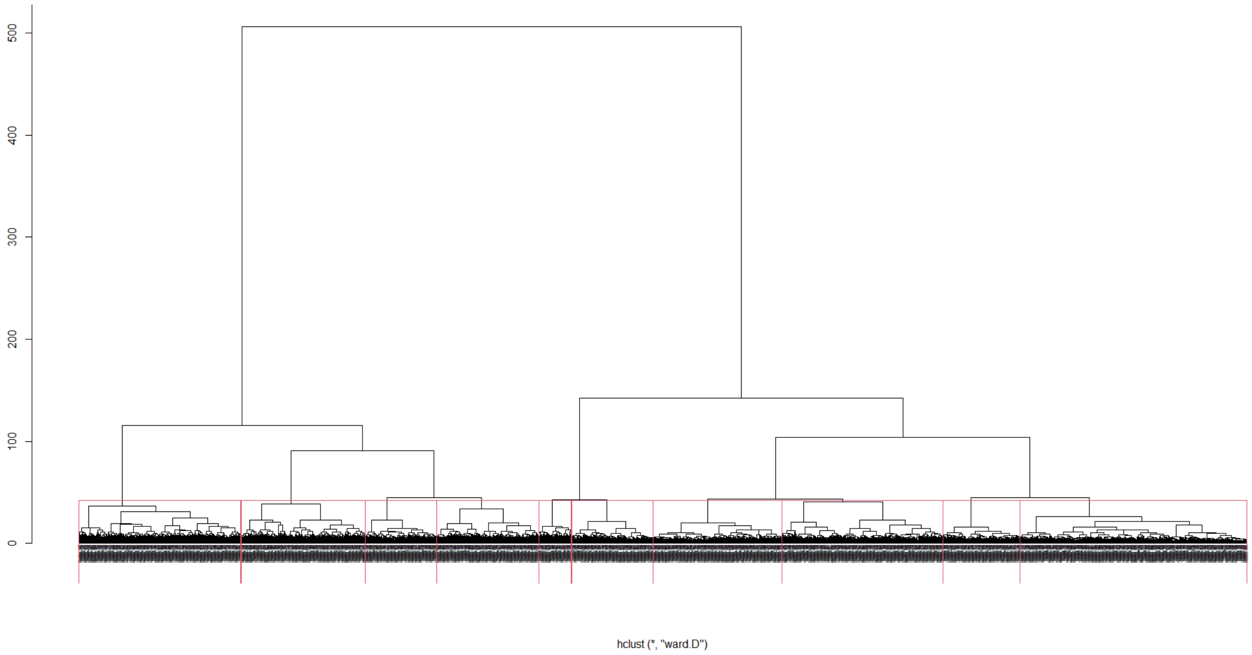


FIGURE 3 – Dendogram for year 2018

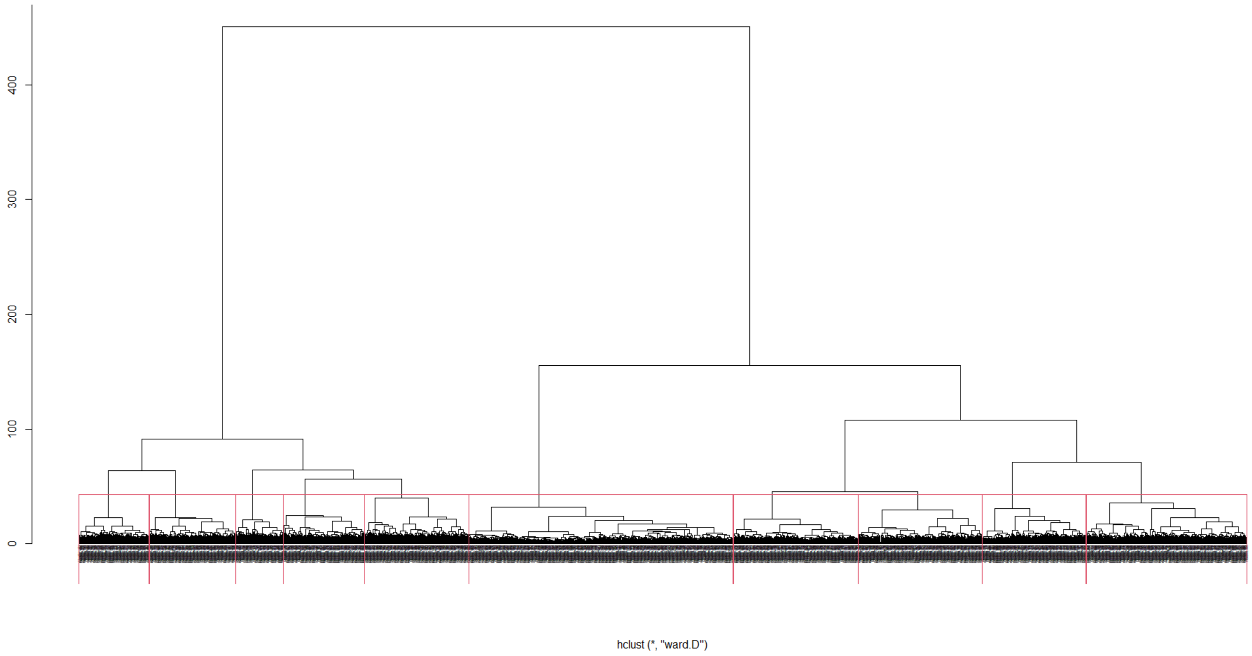
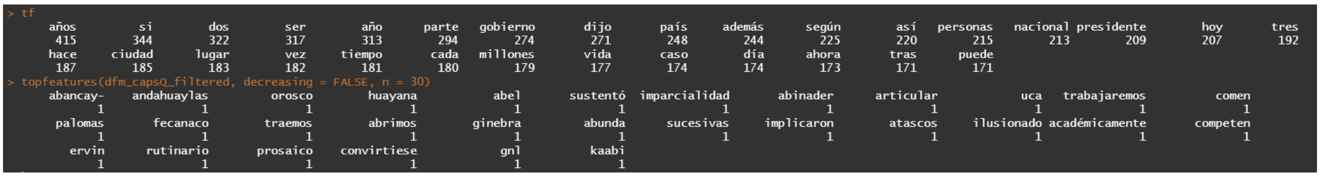
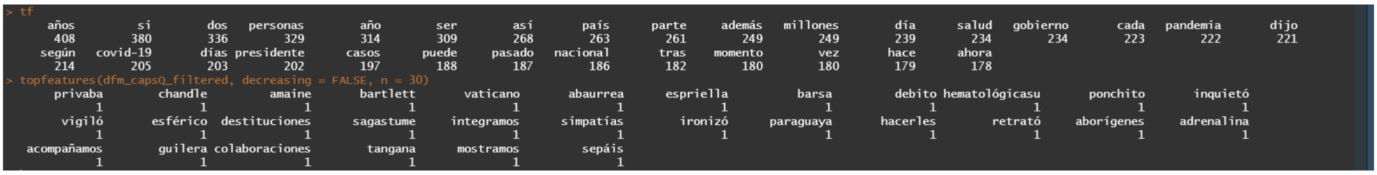


FIGURE 4 – Dendogram for year 2021



```
> tf
      años         si        dos        ser        año      parte   gobierno       dijo       país     además      según        así   personas   nacional presidente        hoy       tres
       415        344        322        317        313        294        274        271        248        244        225        220        215        213        209        207        192
      hace     ciudad      lugar        vez     tiempo       cada   millones       vida       caso        día      ahora       tras      puede
       187        185        183        182        181        180        179        177        174        174        173        171        171
> topfeatures(dfm_capsQ_filtered, decreasing = FALSE, n = 30)
      abancay-    andahuaylas         orosco        huayana           abel       sustentó  imparcialidad       abinader      articular            uca   trabajaremos          comen
             1              1              1              1              1              1              1              1              1              1              1              1
       palomas       fecanaco        traemos        abrimos        ginebra         abunda      sucesivas     implicaron        atascos     ilusionado académicamente       competen
             1              1              1              1              1              1              1              1              1              1              1              1
         ervin      rutinario       prosaico    convirtiese            gnl          kaabi
             1              1              1              1              1              1
```

Figure 5 – Top and bottom features for year 2018



```
> tf
      años         si        dos   personas        año        ser        así       país      parte     además   millones        día      salud   gobierno       cada   pandemia       dijo
       408        380        336        329        314        309        268        263        261        249        249        239        234        234        223        222        221
     según   covid-19       días presidente      casos      puede     pasado   nacional       tras    momento        vez       hace      ahora
       214        205        203        202        197        188        187        186        182        180        180        179        178
> topfeatures(dfm_capsQ_filtered, decreasing = FALSE, n = 30)
       privaba        chandle         amaine       bartlett       vaticano       abaurrea      espriella          barsa         debito hematológicasu       ponchito       inquietó
             1              1              1              1              1              1              1              1              1              1              1              1
        vigiló       esférico  destituciones      sagastume     integramos      simpatías        ironizó      paraguaya       hacerles        retrató     aborígenes     adrenalina
             1              1              1              1              1              1              1              1              1              1              1              1
   acompañamos        guilera colaboraciones        tangana      mostramos         sepáis
             1              1              1              1              1              1
```

Figure 6 – Top and bottom features for year 2021