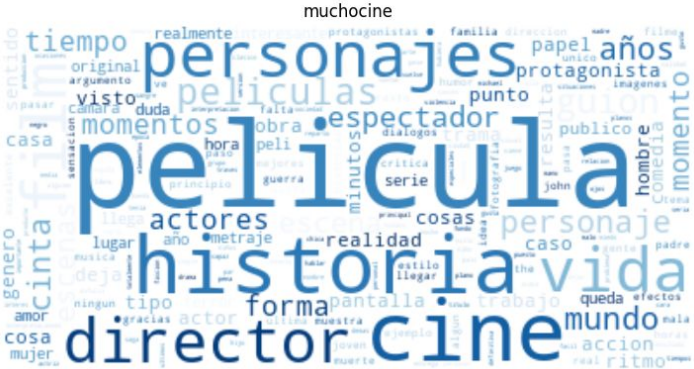
Exploración de texto

Fabián Villena

Introducción

El primer paso en cualquier proyecto de ciencia de datos es la exploración y en el caso de los datos de texto libre no estructurado se deben utilizar técnicas específicas, las cuales son distintas que en los datos estructurados.



El vocabulario

El vocabulario es el conjunto de palabras distintas que están presente en un corpus.

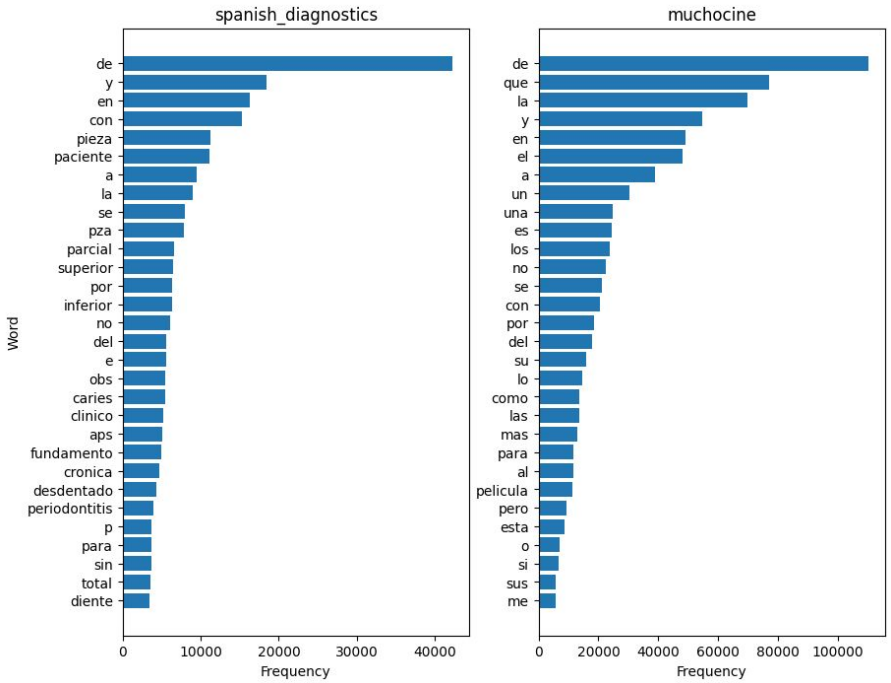
Los vocabularios pueden variar completamente entre corpora y es importante analizar las palabras que están contenidas en el conjunto.

```
de, no, la, y, del,
en, refraccion, los,
a, dientes,
especificada, con,
enfermedades,
especificado, otros,
cronica, encias,
hernia, trastorno,
vicio, o
```

La frecuencia de palabras

Después de tener calculado el vocabulario del corpus que se va a analizar se puede calcular la frecuencia de las palabras.

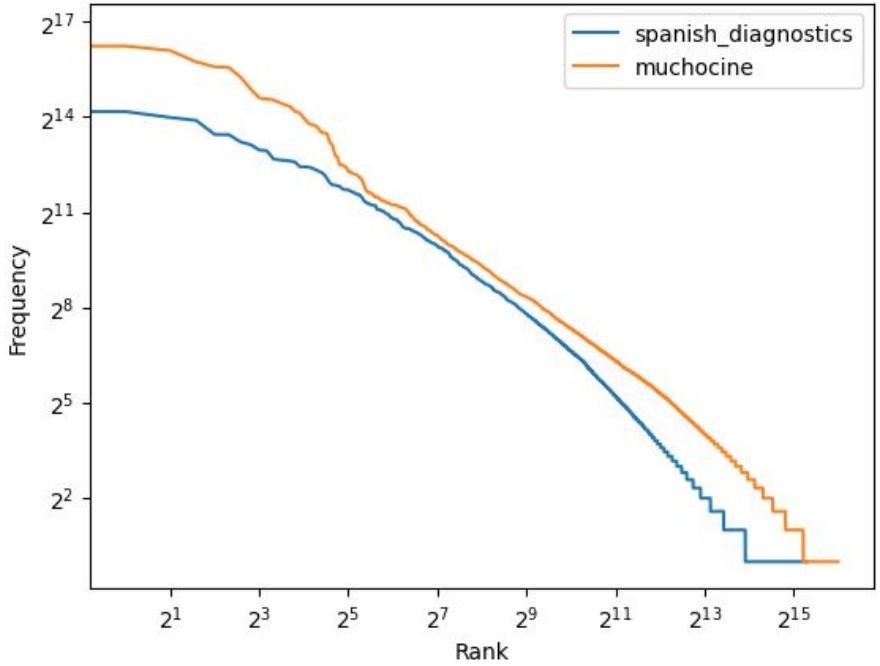
La frecuencia de palabras es la cantidad de veces que está la palabra dentro del corpus o el documento que se está analizando.



Distribución de la frecuencia de palabras

La distribución de la frecuencia de palabras nos puede ayudar a conocer cuál es el contenido y el contexto del corpus o de cada documento.

Existen distribuciones específicas que pueden describir los datos de texto libre no estructurado.



El tamaño del corpus

El largo del corpus es la cantidad de palabras que presenta el conjunto de documentos.

El análisis tomando en cuenta sólo esta métrica no nos aporta mucha información en la exploración del texto.

Table 1. Summary statistics of the corpora

	corpus		
Metric	German	English	Spanish
Articles count	59 539	22 372	12 058
Number of word tokens	20 437 502	12 093 145	51 337 854
Vocabulary size	497 256	144 550	374 877

El tamaño del vocabulario

El tamaño del vocabulario es la cantidad de palabras distintas que presenta un corpus. Esto nos puede comunicar la complejidad que puede presentar el texto, pero se puede ver también confundida por el tamaño del corpus.

Table 1. Summary statistics of the corpora

	corpus		
Metric	German	English	Spanish
Articles count	59 539	22 372	12 058
Number of word tokens	20 437 502	12 093 145	51 337 854
Vocabulary size	497 256	144 550	374 877

Diversidad léxica

La diversidad léxica es un aspecto de la riqueza léxica y se refiere a la razón entre el tamaño del vocabulario y la cantidad total de tokens en el corpus.

Genre	**Tokens**	**Types**	**Lexical diversity**
skill and hobbies	82345	11935	0.145
fiction: science	14470	3233	0.223
press: reportage	100554	14394	0.143
fiction: romance	70022	8452	0.121
religion	39399	6373	0.162

Concordancia

La concordancia nos muestra cada instancia de una palabra específica junto con las palabras de contexto que la acompañan.

```
>>> text1.concordance("monstrous")
Displaying 11 of 11 matches:
ong the former , one was of a most monstrous size . ... This came towards us ,
ON OF THE PSALMS . " Touching that monstrous bulk of the whale or ork we have r
ll over with a heathenish array of monstrous clubs and spears . Some were thick
d as you gazed , and wondered what monstrous cannibal and savage could ever hav
that has survived the flood ; most monstrous and most mountainous ! That Himmal
they might scout at Moby Dick as a monstrous fable , or still worse and more de
th of Radney .'" CHAPTER 55 Of the monstrous Pictures of Whales . I shall ere l
ing Scenes . In connexion with the monstrous pictures of whales , I am strongly
ere to enter upon those still more monstrous stories of them which are to be fo
ght have been rummaged out of this monstrous cabinet there is no telling . But
of Whale - Bones ; for Whales of a monstrous size are oftentimes cast up dead u
>>>
```

Colocación

Una colocación es una secuencia de palabras que frecuentemente ocurren simultáneamente, por ejemplo “manzana roja” es una colocación y por el contrario “manzana azul” no lo es.

Una característica de las colocaciones es que son resistentes a la sustitución por palabras que tienen significados similares. Por ejemplo “manzana colorada” suena extraño.

Bigramas y n-Gramas

Una forma de tomar en cuenta estas colocaciones es mediante la utilización de bigramas, los cuales son pares de palabras.

No sólo podemos añadir al vocabulario bigramas, sino que también cualquier conjunto de n palabras.

```
>>> text4.collocations()
United States; fellow citizens; four years; years ago; Federal
Government; General Government; American people; Vice President; Old
World; Almighty God; Fellow citizens; Chief Magistrate; Chief Justice;
God bless; every citizen; Indian tribes; public debt; one another;
foreign nations; political parties
>>> text8.collocations()
would like; medium build; social drinker; quiet nights; non smoker;
long term; age open; Would like; easy going; financially secure; fun
times; similar interests; Age open; weekends away; poss rship; well
presented; never married; single mum; permanent relationship; slim
build
>>>
```

Nubes de palabras

Las nubes de palabras son un tipo de visualización de resultados de análisis en Procesamiento de Lenguaje Natural. **Esta visualización consiste en mapear el tamaño de la palabra hacia un valor de importancia relativa de la palabra dentro de un texto**, este valor puede ser la frecuencia bruta o también podría ser el TF*IDF.