EDA and NLP basics

Exploring the innards of the Spanish poetry

PyConES 2020

Andrea Morales and Miguel López

andreamorgarz@gmail.com
@ andreamorgar

laskerjpc@gmail.com
@ wizmik12

https://github.com/andreamorgar/poesIA

Natural Language Processing (NLP)

What does it means?

- A field of Artificial Intelligence
- Deals with the interaction between computers and humans using the natural language.
- Makes the human language understandable for computers.
- Also uses machine learning to derive meaning from human languages.

https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32

Natural Language Processing (NLP)

Interesting applications

- Sentiment Analysis
- Chatbots & Virtual Assistants
- Text Classification
- Text Summarization

Steps

Finally, we will be able to extract valuable information from the data. There exists so many different applications such as text classification.

Different topics will lead to different word contexts and some representation will be more suitable.

Preprocess the text using well-known techniques such as lemmatization, tokenization,...

Collect the data, e.g. from scratch, an open repository, scraping a website,...

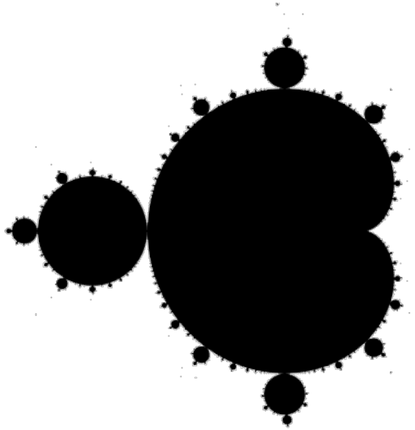
https://blog.insightdatascience.com/how-to-solve-90-of-nlp-problems-a-step-by-step-guide-fda605278e4e

Packages used

matplotlib

Beautiful Soup

GENSIM
topic modelling for humans



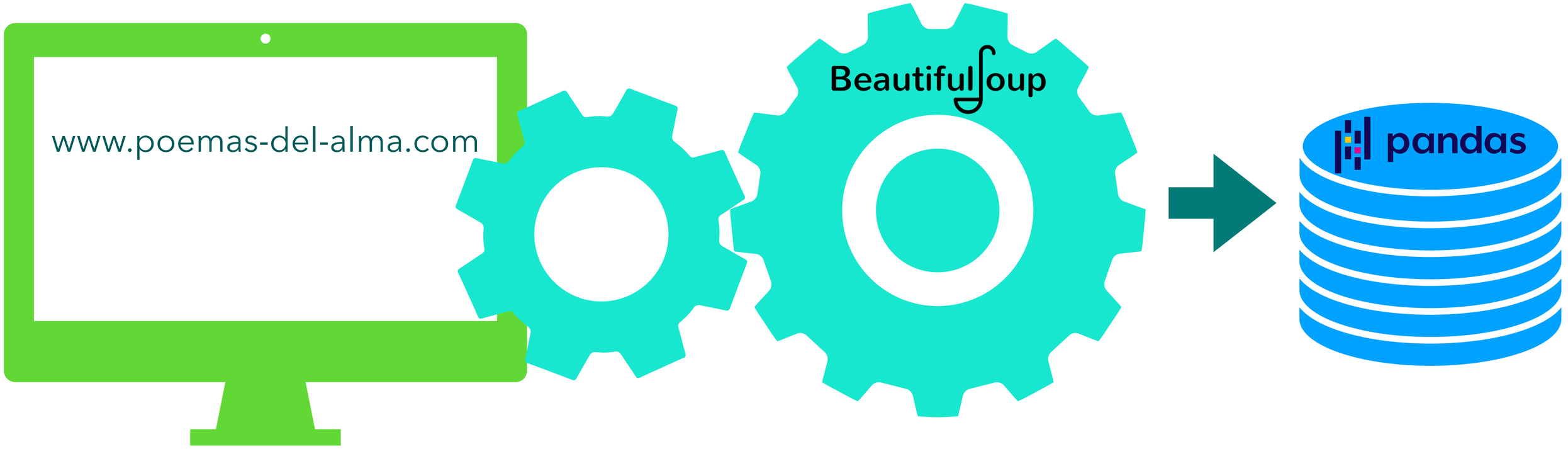
NLTK

pandas

TextBlob

Gather the data

Creating our poetry dataset: let's scrap!



Website

Web Scrapping

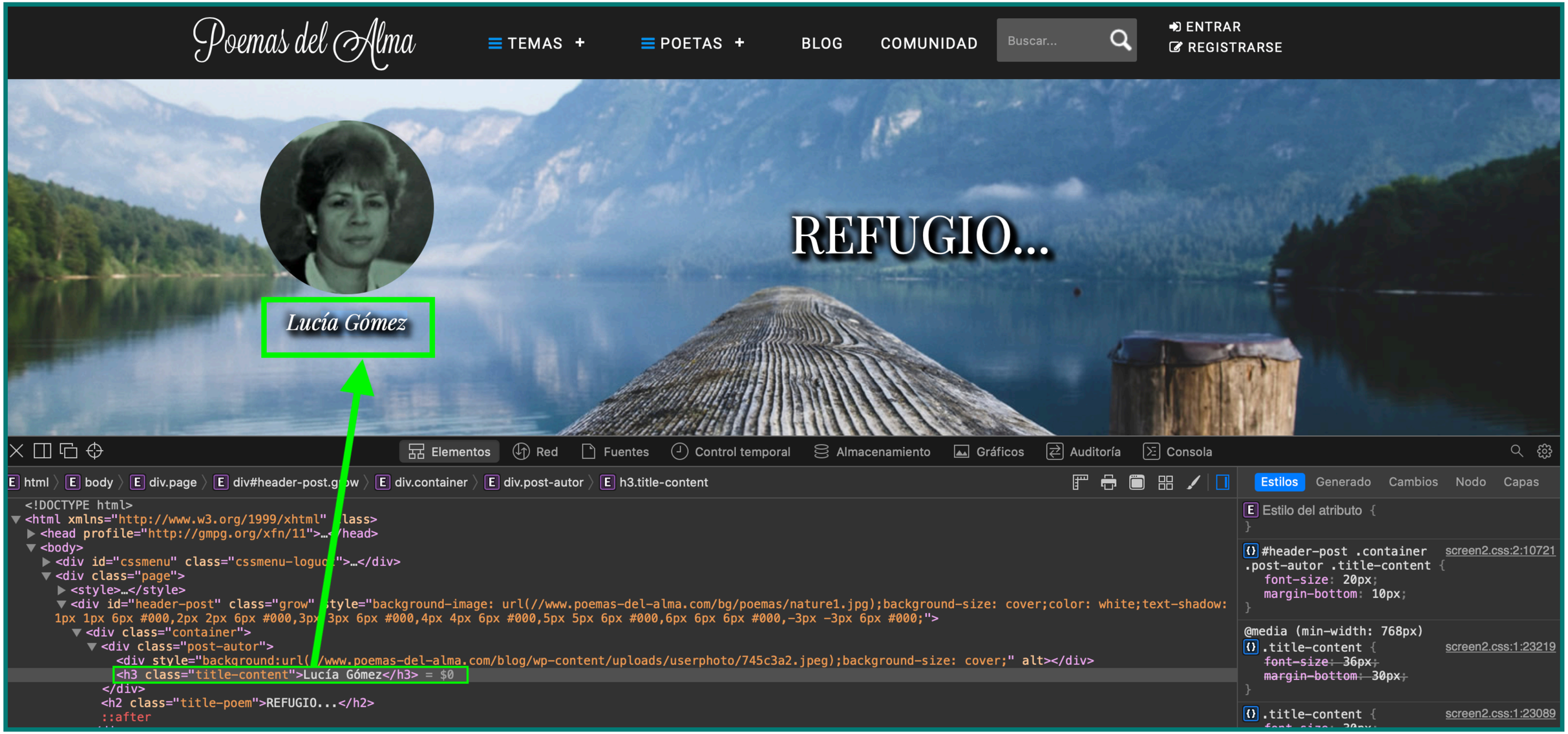
Poetry dataset

Data published on https://github.com/andreamorgar/poesIA/blob/master/data/poems.csv

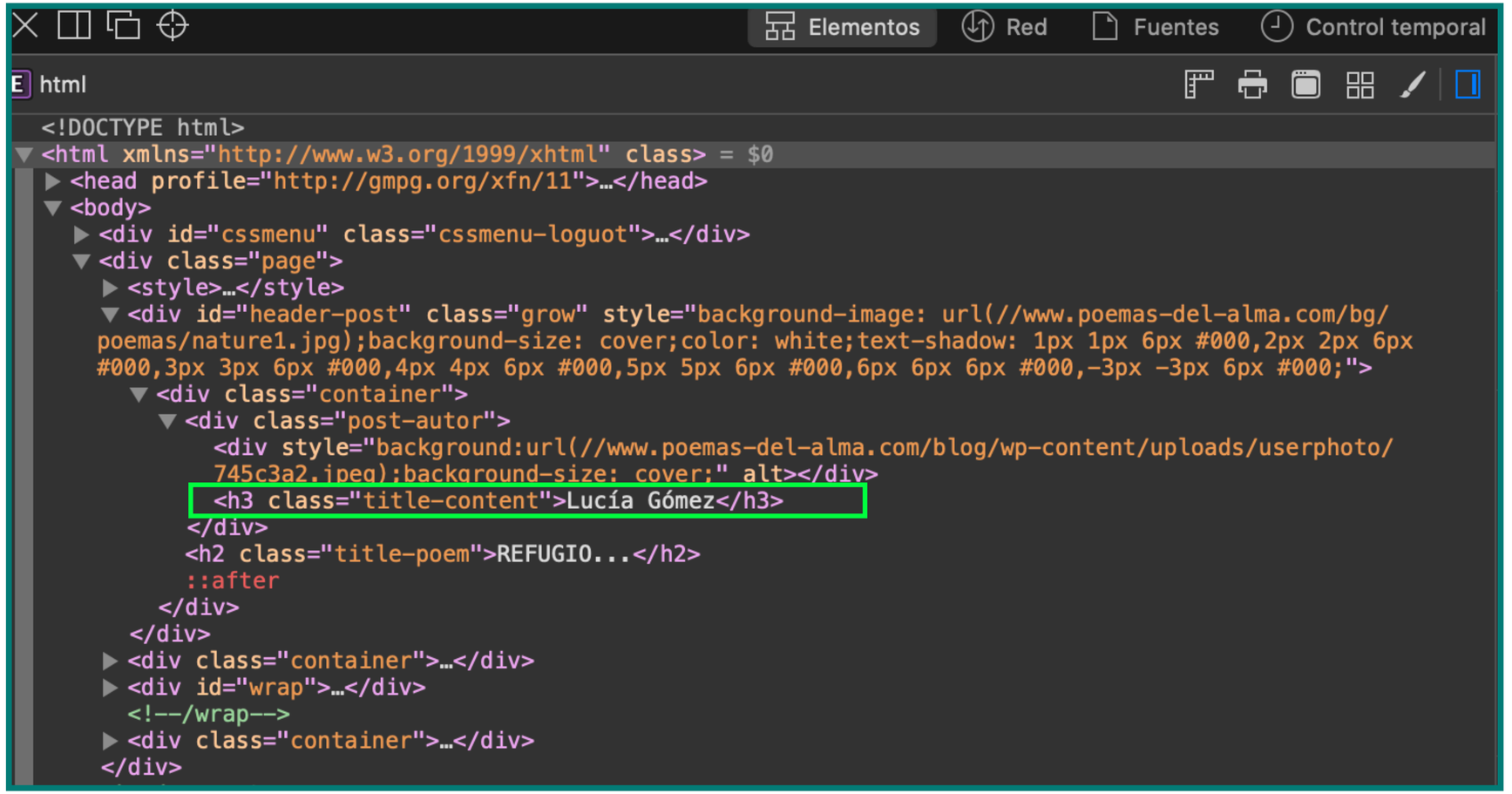
How scrap?


Poemas del Alma
TEMAS +
POETAS +
BLOG
COMUNIDAD
Buscar...
ENTRAR
REGISTRARSE
Lucía Gómez
REFUGIO...
INICIO > LUCÍA GÓMEZ > REFUGIO...
COMPARTIR
TWITTEAR
¿Me vas a abandonar, dulce tristeza mía?
¿No sabes que en ti, puedo encontrar refugio?
Contigo bajé las escaleras y si te vas,
veré el vacío y el umbral de tus lares.
Me quedaré sola entre tantos recuerdos,
en la casa de la noche mía.
En las soledades, hay silencios largos
en los cuales las cosas se abandonan
y el azul del cielo se muestra más ameno.
La soledad puede cubrir el tedio del invierno
justo cuando el hielo del corazón,
empieza a derretirse.
Información del poema
Autor: Lucía Gómez (Offline)
Publicado: 21 de septiembre de 2020 a las 10:20
Categoría: Sin clasificar
Lecturas: 9

How scrap?


Poemas del Alma
TEMAS +
POETAS +
BLOG
COMUNIDAD
Buscar...
ENTRAR
REGISTRARSE
REFUGIO...
Lucía Gómez
Elementos
Red
Fuentes
Control temporal
Almacenamiento
Gráficos
Auditoría
Consola
html body div.page div#header-post.grow div.container div.post-autor h3.title-content
<h3 class="title-content">Lucía Gómez</h3> = $0
Estilos
Generado
Cambios
Nodo
Capas
#header-post .container .post-autor .title-content {
font-size: 20px;
margin-bottom: 10px;
@media (min-width: 768px)
.title-content {
font-size: 36px;
margin-bottom: 30px;

How scrap?



We look for the tags of our interest and extract their text

```
poem_author = poem_soup.find("h3",attrs={"class":"title-content"}).text
```

Exploring the dataset

A first view

- We extracted three columns for every poem.
- These columns consist of author, content and title.

	author	content	title
0	Leopoldo Lugones	\n\nEn el parque confuso\nQue con lánguidas br...	LA MUERTE DE LA LUNA
1	Marilina Rébora	\n\nPorque si tú no velas, vendré como ladrón;...	PORQUE SI TÚ NO VELAS
2	Antonio Colinas	\n\nPequeña de mis sueños, por tu piel las pal...	POEMA DE LA BELLEZA CAUTIVA QUE PERDÍ
3	José María Hinojosa	\n\nLos dedos de la nieve\nrepiquetearon\nen e...	SENCILLEZ
4	Rubén Izaguirre Fiallos	Naciste en Armenia,\npero te fuiste a vivir al...	Breve Carta a Consuelo Suncín

Exploring the dataset

A first view

- Relevant characteristics are revealed with the *describe* method.
- There are 267 different authors and 5128 different poems.
- *Pablo Neruda* is the most prolific author of this dataset and *Cien sonetos de amor* is the most repeated titled poem (maybe there are 100 sonnets).

	author	content	title
count	5131	5131	5131
unique	267	5128	4842
top	Pablo Neruda	\n\n	Cien sonetos de amor
freq	357	2	100

Clean the data

We need to clean the text to proper work with the data

We part from this ugly text…

```
[13] poems_df.content[455]
```

```
'A veces vivo un poco,\ny ostento la evidencia\ncomo un coleccionis
el engaño del verbo flagelado.\n\nMi intemperie\ndescansa un instan
n unitaria de la casa.'
```

```
ún trofeo\nrutila en las escarchas de mi nombre\ny emerge la que era\nen
pedestal de hierba de sus ojos,\nhasta volver,\ncrucificada,\na la oració
```

… to a cleaned text which we can work with:

```
['veces', 'vivo', 'ostento', 'evidencia', 'coleccionista', 'algún', 'trofeo', 'rutila', 'escarchas', 'nombre', 'emerge', 'engaño', 'verbo',
'flagelado', 'intemperie', 'descansa', 'instante', 'pedestal', 'hierba', 'ojos', 'volver', 'crucificada', 'oración', 'unitaria', 'casa']
```

Clean the data

Removing newline characters: ‘\n’ and ‘\r’

- Newline characters are not useful at all for the textual information. First, we removed them all.
- A simple way of doing this is to replace all this kind of character with an empty space.

```
'A veces vivo un poco, y ostento la evidencia como un coleccionista.  Algún trofeo rutila en las escarchas de mi nombre y emerge la que era en el enga
ño del verbo flagelado.  Mi intemperie descansa un instante en el pedestal de hierba de sus ojos, hasta volver, crucificada, a la oración unitaria de
la casa.'
```

Clean the data

Let’s tokenize the text!

- We convert the strings into tokens: we use the *RegexpTokenizer* function.
- The result is a list for every poem consisting of the words that compose it.
- We also convert into lowercase every letter.

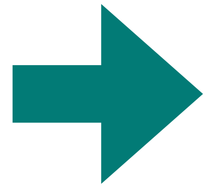
```
['a',
 'veces',
 'vivo',
 'un',
 'poco',
 'y',
 'ostento',
 'la',
 'evidencia',
 'como',
 'un',
 'coleccionista',
 'algún',
 'trofeo',
 'rutila',
 'en',
 'las',
 'escarchas',
 'de',
 'mi',
 'nombre',
 'y',
 'emerge',
 'la',
 'que',
 'era',
 'en',
```

Clean the data

Removing (Spanish) stop words

- In the list on the left, there are some words that not have any substantial meaning (*stop words*) like prepositions or articles. That words are often removed in NLP tasks.
- In this case, we remove the Spanish stop words that appear in the poems.

```
['a',
 'veces',
 'vivo',
 'un',
 'poco',
 'y',
 'ostento',
 'la',
 'evidencia',
 'como',
 'un',
 'coleccionista',
 'algún',
 'trofeo',
 'rutila',
 'en',
 'las',
 'escarchas',
 'de',
 'mi',
 'nombre',
 'y',
 'emerge',
 'la',
 'que',
 'era',
 'en',
```



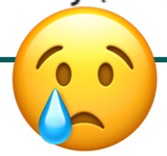
```
['veces',
 'vivo',
 'ostento',
 'evidencia',
 'coleccionista',
 'algún',
 'trofeo',
 'rutila',
 'escarchas',
 'nombre',
 'emerge',
 'engaño',
 'verbo',
 'flagelado',
 'intemperie',
 'descansa',
 'instante',
 'pedestal',
 'hierba',
 'ojos',
 'volver',
 'crucificada',
 'oración',
 'unitaria',
 'casa']
```

Clean the data

Before and after the preprocessing task

Original text

```
0       \n\r\nEn el parque confuso\r\nQue con lánguida...
1       \n\r\nPorque si tú no velas, vendré como ladró...
2       \n\r\nPequeña de mis sueños, por tu piel las p...
3       \n\r\nLos dedos de la nieve\r\nrepiquetearon\r...
4       Naciste en Armenia,\r\npero te fuiste a vivir ...
                                ...
5128    \n¿Vienes? Me llega aquí, pues que suspiras, \...
5129    \n\r\nNada es memoria: todo es invención.\r\nL...
5130    \nFelicidad: Muy dentro de tí.\r\nSerenidad: E...
5131    \nMis manos \r\nabren las cortinas de tu ser \...
5132    \n\r\nY ahora danos\r\nuna muerte honorable,\r...
Name: content, Length: 5133, dtype: object
```



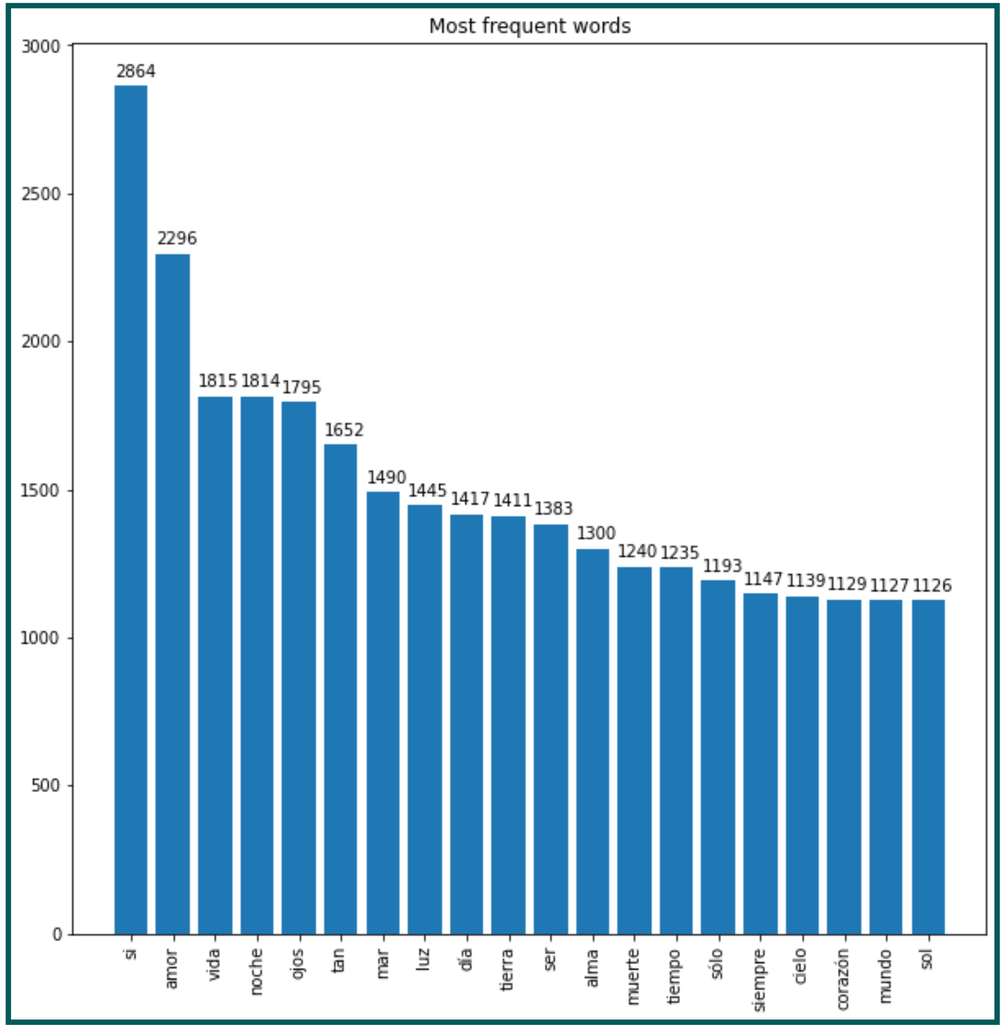
Preprocessed text

```
0       [parque, confuso, lánguidas, brisas, cielo, sa...
1       [si, velas, vendré, ladrón, llegar, sepas, hor...
2       [pequeña, sueños, piel, palomas, pálida, prese...
3       [dedos, nieve, repiquetearon, tamboril, espaci...
4       [naciste, armenia, vivir, mundo, tres, nombres...
                                ...
5128    [vienes, llega, aquí, pues, suspiras, soplo, m...
5129    [memoria, invención, recuerdo, invento, obra, ...
5130    [felicidad, dentro, tí, serenidad, cada, amane...
5131    [manos, abren, cortinas, ser, visten, desnudez...
5132    [ahora, danos, muerte, honorable, vieja, madre...
Name: content, Length: 5133, dtype: object
```

Counting the most frequent words

Bar chart

- We get the 20 most frequent words.
- We ordered them decreasingly
- We also added the number of occurrences of each one.



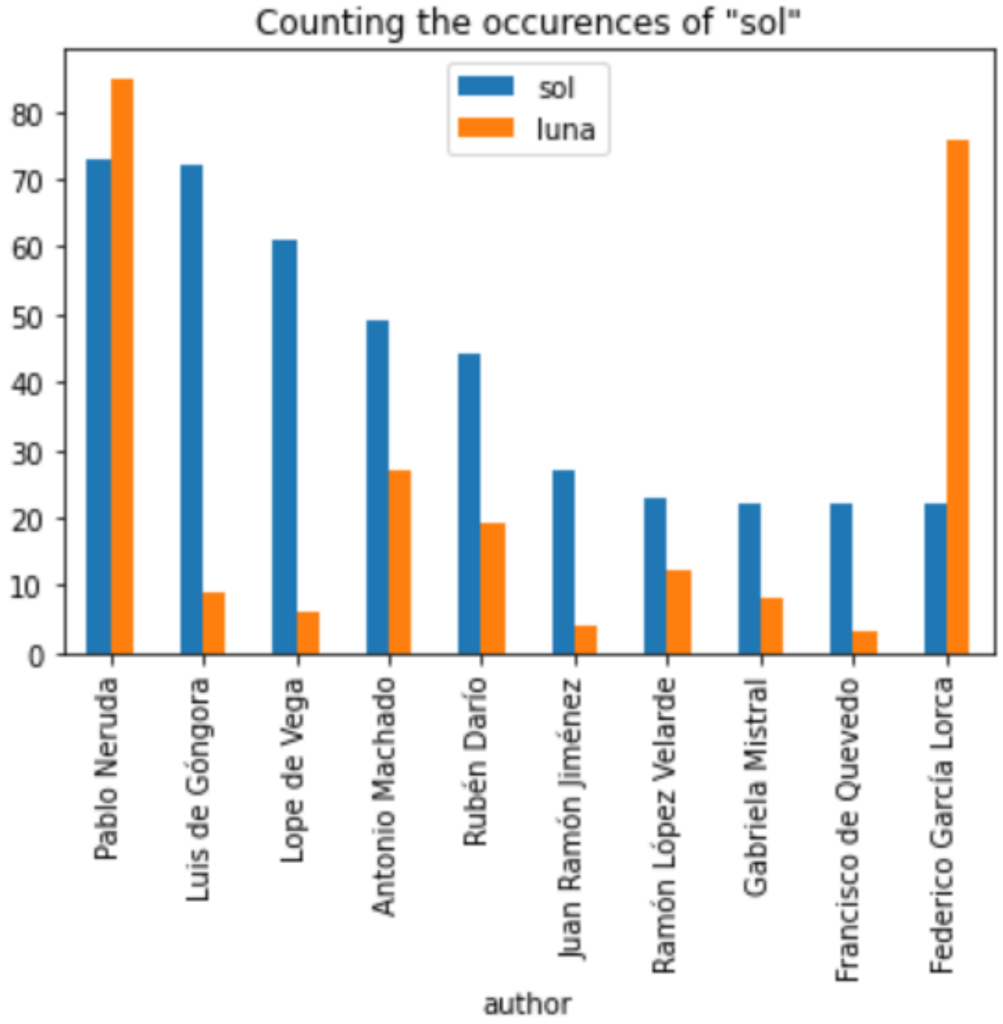
Counting the most frequent words

Word cloud

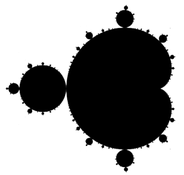
- With this visualization, we will see the most-used words in the whole corpus.
- Notice that the **font size of the words is relevant**: most-used words are the biggest ones!
- In our corpus, ***día***, ***sueño*** and ***noche*** are examples of the most popular words. This is not a surprise: they are such poetic words!

Counting occurrences

- We can calculate the **occurrences of two opposite words** as ***sol*** (sun) and ***luna*** (moon) which are used very frequently in Spanish poetry.
- Pablo Neruda uses more ***luna*** than ***sol***!



Sentiment analysis



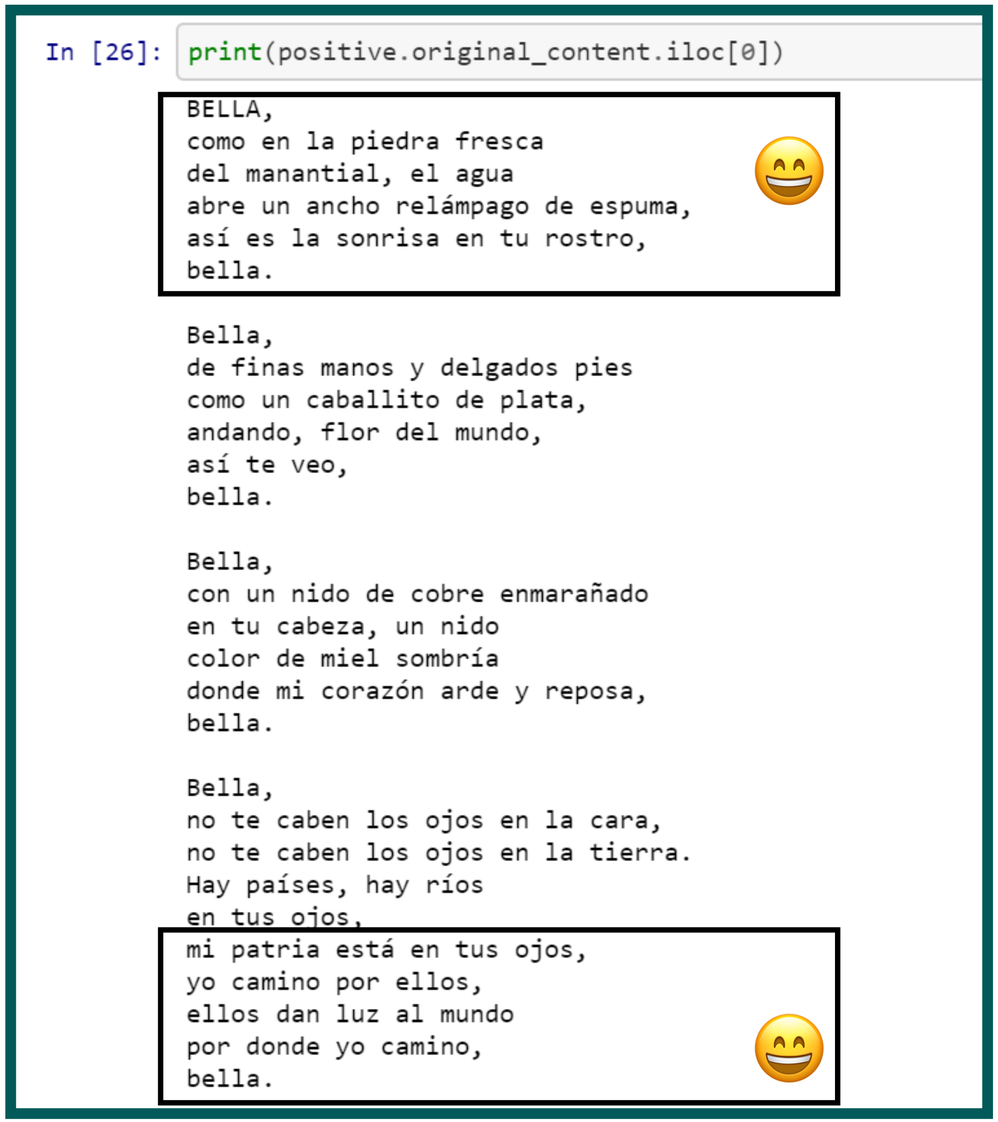
TextBlob

- We collect Pablo Neruda's poems and we analyze the sentiment of them.
- Positive values refer to positive/good feeling poems and negative values refer to negative/sad poems. A zero value is neutral.

We give a grade which represents polarity sentiment

	author	content	title	original_content	content_joined	sentiment
171	Pablo Neruda	['bella', 'piedra', 'fresca', 'manantial', 'ag...	Bella	BELLA,\r\ncomo en la piedra fresca\r\ndel mana...	bella piedra fresca manantial agua abre ancho ...	0.565025
345	Pablo Neruda	['feliz', 'opiné', 'delante', 'sabio', 'examin...	Sin embargo me muevo	De cuando en cuando soy feliz!,\r\nopiné delan...	feliz opiné delante sabio examinó pasión demos...	0.507692
6	Pablo Neruda	['cien', 'sonetos', 'amor', 'volar', 'tiempo',...	Cien sonetos de amor	Cien sonetos de amor\r\n\nHay que volar en est...	cien sonetos amor volar tiempo dónde alas avió...	0.506061
31	Pablo Neruda	['vez', 'dejadme', 'ser', 'feliz', 'pasado', '...	Oda al día feliz	ESTA vez dejadme\r\nser feliz,\r\nnada ha pasa...	vez dejadme ser feliz pasado nadie parte algun...	0.502614
312	Pablo Neruda	['cien', 'sonetos', 'amor', 'olvidé', 'manos',...	Cien sonetos de amor	Cien sonetos de amor\r\n\nPero olvidé que tus ...	cien sonetos amor olvidé manos satisfacían raí...	0.500000

The most positive poem



```
In [26]: print(positive.original_content.iloc[0])
```

```
BELLA,
como en la piedra fresca
del manantial, el agua
abre un ancho relámpago de espuma,
así es la sonrisa en tu rostro,
bella.

Bella,
de finas manos y delgados pies
como un caballito de plata,
andando, flor del mundo,
así te veo,
bella.

Bella,
con un nido de cobre enmarañado
en tu cabeza, un nido
color de miel sombría
donde mi corazón arde y reposa,
bella.

Bella,
no te caben los ojos en la cara,
no te caben los ojos en la tierra.
Hay países, hay ríos
en tus ojos,
mi patria está en tus ojos,
yo camino por ellos,
ellos dan luz al mundo
por donde yo camino,
bella.
```

- We can order the poems by their sentiment value to get the most cheerful one.
- As we expected, these poems can be so delightful.

"Bella, no te caben los ojos en la cara, no te caben los ojos en la tierra."

"Bella, your eyes don't fit on your face, your eyes don't fit on earth."

The most negative poem

Also are they really touching…

"Y entre la noche negra -desesperadas- corren y sollozan las almas de los obreros muertos"

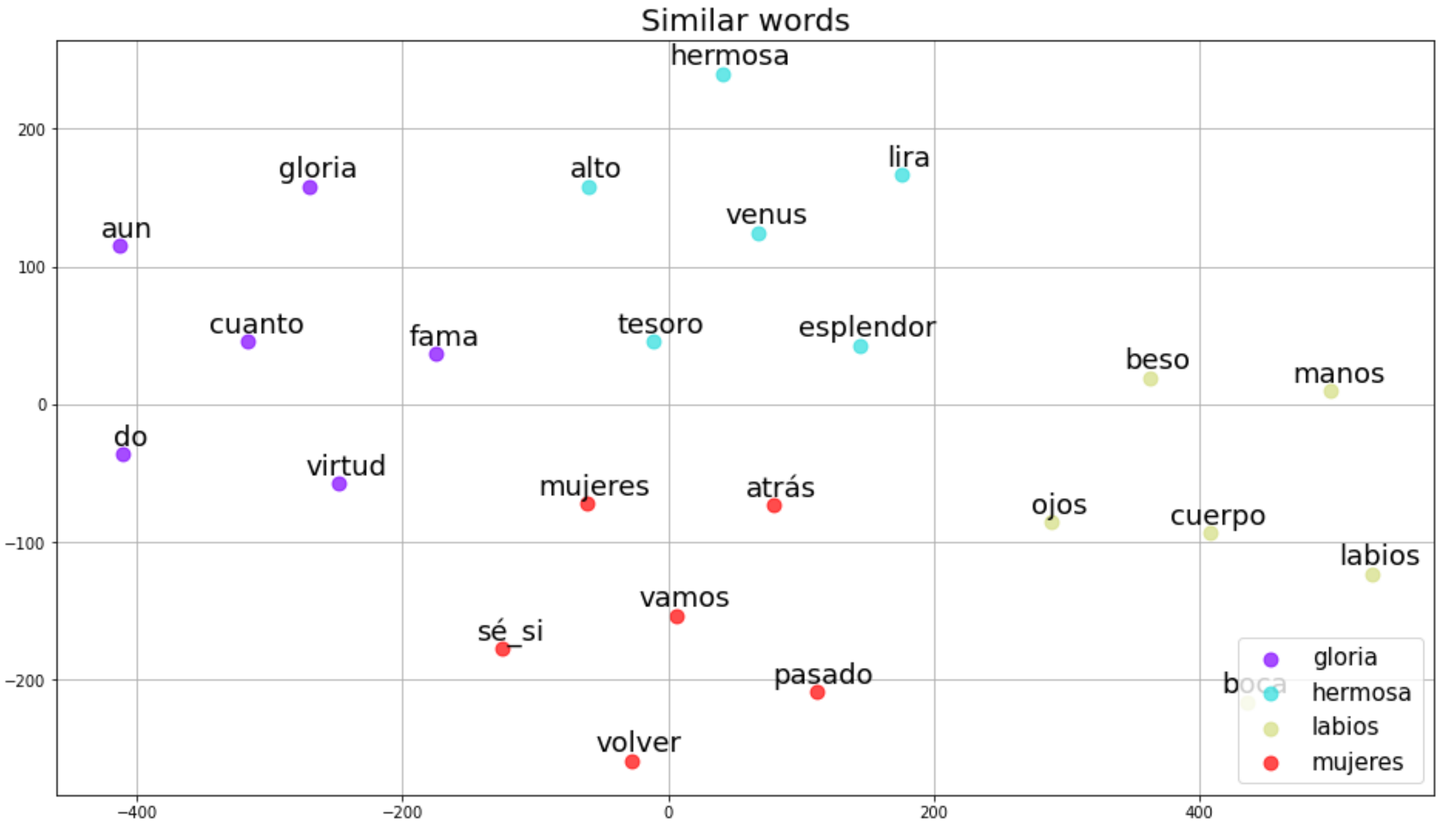
"And through the black night -desperate- the souls of the dead workers run and sob"

```
In [24]: print(negative.original_content.iloc[0])
```

```
Fierro negro que duerme, fierro negro que gime
por cada poro un grito de desconsolación.

Las cenizas ardidas sobre la tierra triste,
los caldos en que el bronce derritió su dolor.

Aves de qué lejano país desventurado
graznaron en la noche dolorosa y sin fin?

Y el grito se me crispa como un nervio enroscado
o como la cuerda rota de un violín.

Cada máquina tiene una pupila abierta
para mirarme a mí.

En las paredes cuelgan las interrogaciones,
florece en las bigornias el alma de los bronces
y hay un temblor de pasos en los cuartos desiertos.

Y entre la noche negra -desesperadas- corren
y sollozan las almas de los obreros muertos.
```

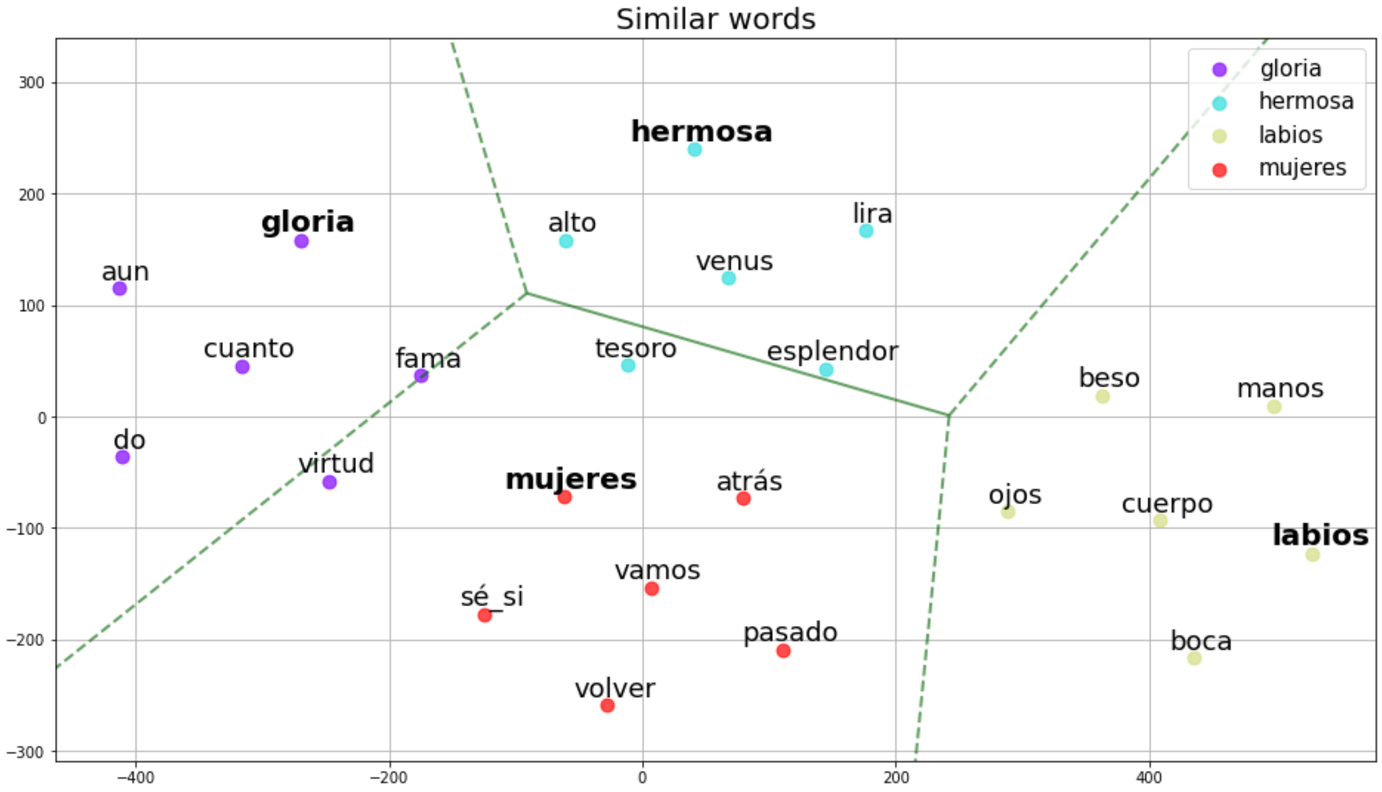
Similarity between words

- In the poetry context, we found similarities between words using the word embedding.
- We plot the most similar words to 4 words: ***gloria***, ***hermosa***, ***labios*** and ***mujeres***.
- To get the 2D-representation we used a TSNE algorithm.



Similarity between words: Voronoi sets

- We show the same picture. This time we plot Voronoi sets to delimiter the 2D space using the 4 words as centroids.
- For each centroid (in bold) word there is a corresponding region consisting of all points of the plane closer to that centroid word than to any other.



Applications

Once we have a valid representation we can solve many problems

- **Poem recommendation**: we can use the semantic vectors to recommend other similar ones
- **Poem/author identification**
- **Poem generation:** using existing poems to create synthetic texts https://github.com/andreamorgar/poesIA

… and many others!

Final ideas

- We have seen how much information we can extract from a website
- Now we have an overall idea of how to go with a basic NLP problem
- We were able to capture semantics and resolve similarity problem
- We studied the nature of the data in the dataset… with really interesting conclusions!

This opens a door to more deep tasks involving this kind of vocabulary, predictive tasks and so more!

Thanks for your attention!

Any questions?