Semántica vectorial

Fabián Villena

Introducción

Para poder ingresar textos libres no estructurados a algoritmos para entrenar modelos de aprendizaje automático necesitamos una técnica para convertir el lenguaje natural en un vector de números.

Estas representaciones en un espacio pueden ser construidas con varias técnicas, con simples conteos de palabras hasta la utilización de redes neuronales artificiales.

Los modelos se ajustan con características numéricas

Debemos transformar los datos no estructurados hacia una representación numérica que pueda ser ingresada a modelos de aprendizaje automático.

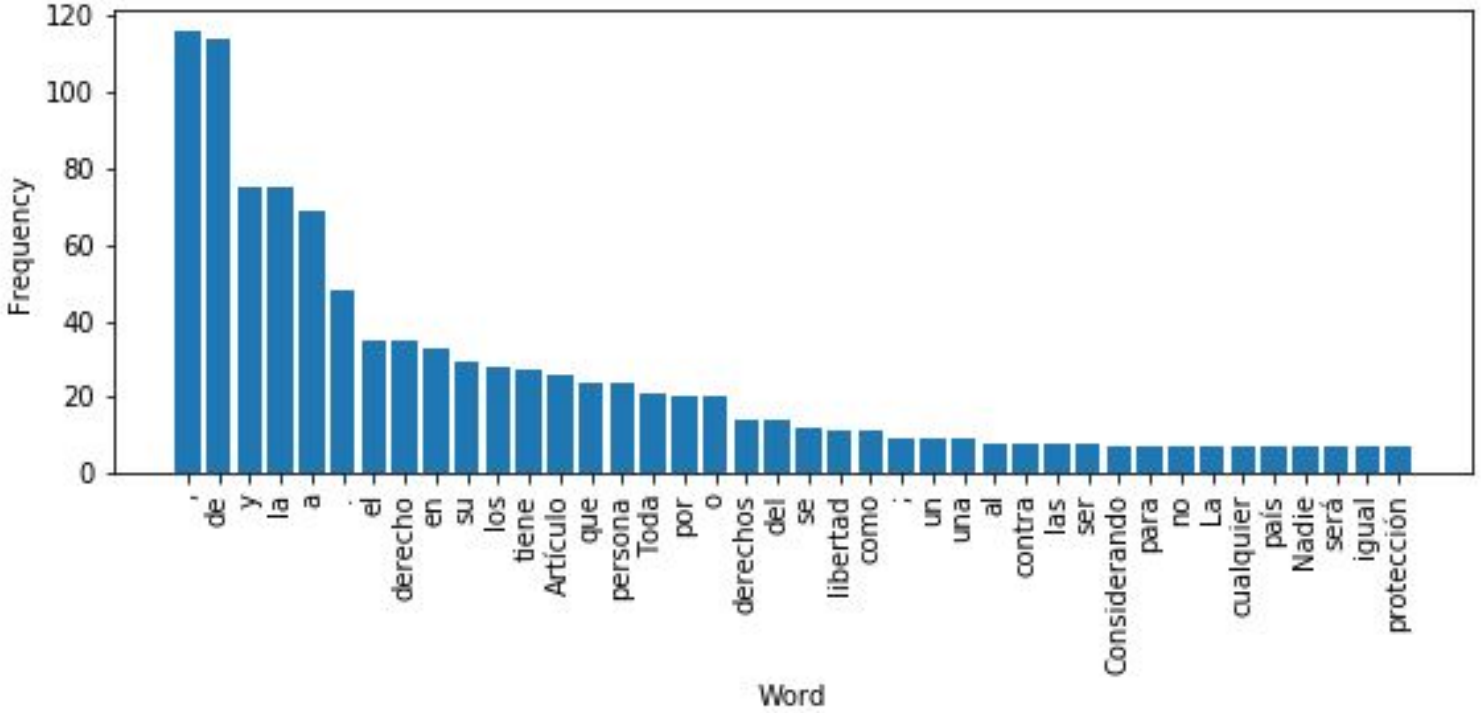
Estas representaciones deben **trasladar los significados de los documentos hacia un espacio vectorial**.

Doc #	diente	dolor	…	caries
1	0.03	1.21	…	0.83
2	0.10	0.36	…	1.13
3	0.00	1.54	…	1.58
…	…	…	…	…
n	0.56	1.11	…	0.67

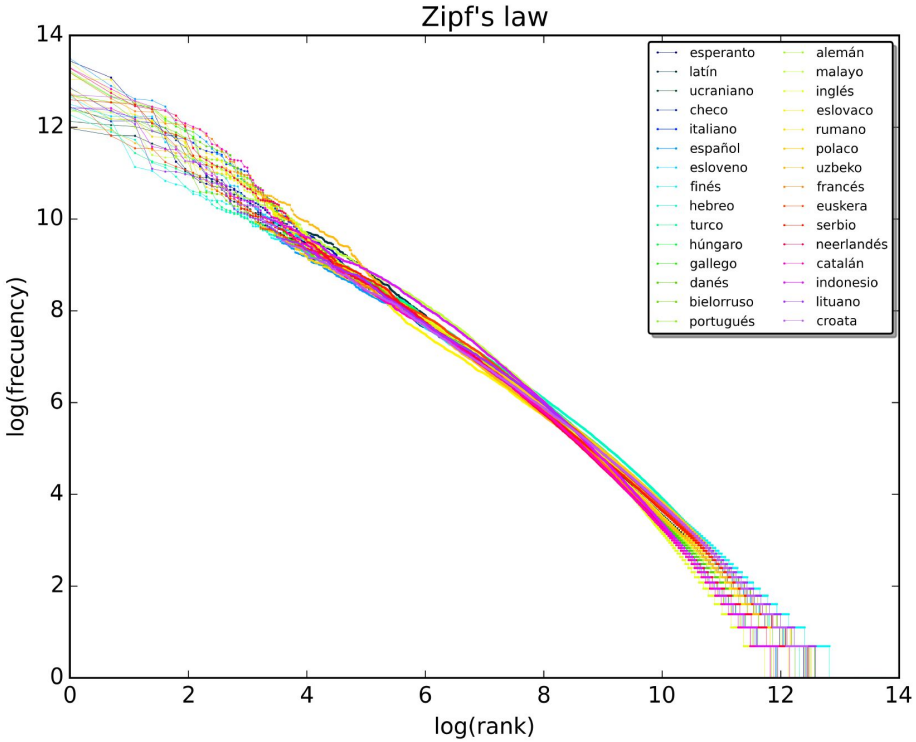
Ley de Zipf

- La ley de Zipf es una ley empírica sobre la frecuencia de palabras en un corpus.
- Esta ley establece que la frecuencia de una palabra en un corpus es inversamente proporcional a su posición en una tabla ordenada de frecuencias.
- Esta ley se relaciona con el principio de mínimo esfuerzo. Normalmente utilizamos la menor cantidad de palabras para comunicar una idea.

Ley de Zipf en la Declaración de Derechos Humanos



Ley de Zipf en múltiples *corpora*



Stop words

Las stop words o palabras vacías son palabras que no tienen significado y normalmente son las palabras más frecuentes en un corpus. **Las stop words incluyen artículos, pronombres, preposiciones, etc. pero también incluyen palabras específicas del corpus** que no aportan significado a los documentos.

Se pueden usar listas predefinidas de stop words por lenguaje como también se puede construir la lista utilizando las palabras más frecuentes del corpus.

• de • la • que • el • en • y • a • los • del • se	• las • por • un • para • con • no • una • su • al • lo

Stop words específicos del corpus

Se puede establecer un umbral en el cual las palabras más frecuentes sean consideradas stop words. En un trabajo del dominio médico se estableció el umbral en el percentil 95 y se encontraron **stopwords como *paciente*, *diagnóstico*, *consulta*, las cuales no están en ninguna lista de stop words**, pero pueden calcularse desde la distribución del corpus.

https://scielo.conicyt.cl/pdf/rmc/v147n10/0717-6163-rmc-147-10-1229.pdf

Hipótesis distribucional

Elementos lingüísticos con distribuciones similares tienen significados similares.

Tomando en cuenta esta hipótesis podemos cuantificar y categorizar similaridades semánticas entre elementos lingüísticos basado en sus propiedades distributivas

Semántica vectorial

La semántica vectorial aprende representaciones del significado de las palabras o documentos directamente desde su distribución en el texto.

Estas representaciones son utilizadas en todas las aplicaciones de procesamiento de lenguaje natural que utilizan significado.

La semántica vectorial es la manera estándar de representar significados de palabras o documentos y con ella podemos modelar varios aspectos del significado.

Aprendizaje de representaciones

Las representaciones de palabras o documentos son un ejemplo de aprendizaje de representaciones, en donde automáticamente se aprenden representaciones de un texto de entrada.

Encontrar maneras para aprender representaciones en vez de crearlas a mano es un tema importante en la investigación en NLP.

Matriz término-documento (*Bag of words*)

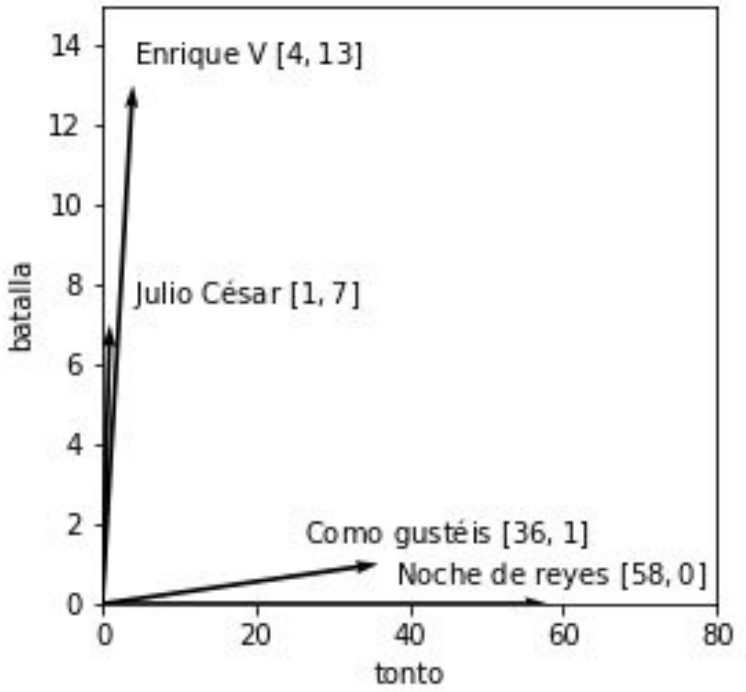
Para representar el significado de un documento a través de su contexto podemos extraer características desde el texto al **contar cuántas veces está presente cada palabra del vocabulario dentro del documento**.

	batalla	bueno	tonto	ingenio
Como gustéis	1	114	36	20
Noche de reyes	0	80	58	15
Julio César	7	62	1	2
Enrique V	13	89	4	3

Los documentos son puntos (vectores) en el espacio

Un vector es una colección de números por lo que podemos interpretar desde la matriz término-documento que **cada documento es un vector** con tantas dimensiones como palabras tenga el vocabulario.

Un espacio vectorial es una colección de vectores caracterizada por sus dimensiones.



El problema de utilizar sólo la frecuencia

La utilización de sólo la frecuencia de las palabras lleva a que existan palabras sobrerrepresentadas que no aportan información al análisis.

Es necesario realizar una ponderación más compleja.

TF*IDF

La utilización de la frecuencia de las palabras en los documentos para poder representarlos no es muy útil cuando las palabras más frecuentes son típicamente stop words y que estas palabras no nos permiten representar diferencias en los documentos.

	batalla	**bueno**	tonto	ingenio
Como gustéis	1	**114**	36	20
Noche de reyes	0	**80**	58	15
Julio César	7	**62**	1	2
Enrique V	13	**89**	4	3

TF

Palabras que ocurren frecuentemente son más importantes que las palabras infrecuentes, pero palabras que son demasiado frecuentes son poco importantes.

Primero debemos ponderar positivamente las palabras que son más frecuentes dentro de un documento mediante una de estas funciones:

$$tf_{t,d} = count(t, d) \qquad tf_{t,d} = \log_{10}(count(t, d) + 1)$$

Donde $tf_{t,d}$ es la frecuencia del término `t` en el documento `d` y `count(t,d)` es la cantidad de veces que el término `t` está en el documento `d`.

IDF

El segundo factor que utilizamos para ponderar las palabras en un documento es el Inverse Document Frequency, el cual le da más peso a palabras que no se repiten en muchos documentos; debido a que palabras que están presentes en muchos documentos, probablemente no generan diferencias entre los mismos, normalmente las palabras con un menor IDF son stop words.

El IDF se calcula tomando en cuenta la cantidad de documentos `N` y $\texttt{df}_t$, la cantidad de documentos donde aparece el término `t`.

$$idf_t = \log_{10}\left(\frac{N}{df_t}\right)$$

Cálculo de la representación

El valor de TF*IDF combina los parámetros calculados utilizando las funciones de TF e IDF, entonces el peso $w_{t,d}$ del término t en el documento d se determina de la siguiente manera:

$$w_{t,d} = tf_{t,d} * idf_t$$

Palabra	**df**	**idf**
Romeo	1	1.57
ensalada	2	1.27
Falstaff	4	0.967
bosque	12	0.489
batalla	21	0.246
ingenio	34	0.037
tonto	36	0.012
bueno	37	0
dulce	37	0

Representación con TF*IDF

Al representar nuestros documentos con TF*IDF se hacen más claras las palabras que generan diferencias entre documentos. Hay palabras que están presentes en todos los documentos y por ende no aportan información para establecer diferencias entre documentos.

	batalla	bueno	tonto	ingenio
Como gustéis	0.074	0	0.019	0.049
Noche de reyes	0	0	0.021	0.044
Julio César	0.22	0	0.0036	0018
Enrique V	0.28	0	0.0083	0.022

Aplicación de TF*IDF

La utilización de este tipo de ponderación genera una representación más válida de las palabras en un documento, sin sobrerrepresentar stop words.

Recuperación de información

Una de las aplicaciones que más utiliza semántica vectorial es la tarea de encontrar un documento dentro de una colección de documentos que más relevancia tenga para una consulta.

También se va a representar la consulta con un vector en el mismo espacio vectorial y se necesitará una manera de comparar estos vectores para obtener los más relevantes.

Similitud de coseno

Para medir la similaridad entre vectores (que están representando documentos) necesitamos una métrica que tome dos vectores de las mismas dimensiones y retorne una medida de similaridad.

La similaridad coseno es el producto punto entre dos vectores normalizado por el largo de los vectores, para que el largo del vector no sesgue la similaridad

$$\text{cosine}(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

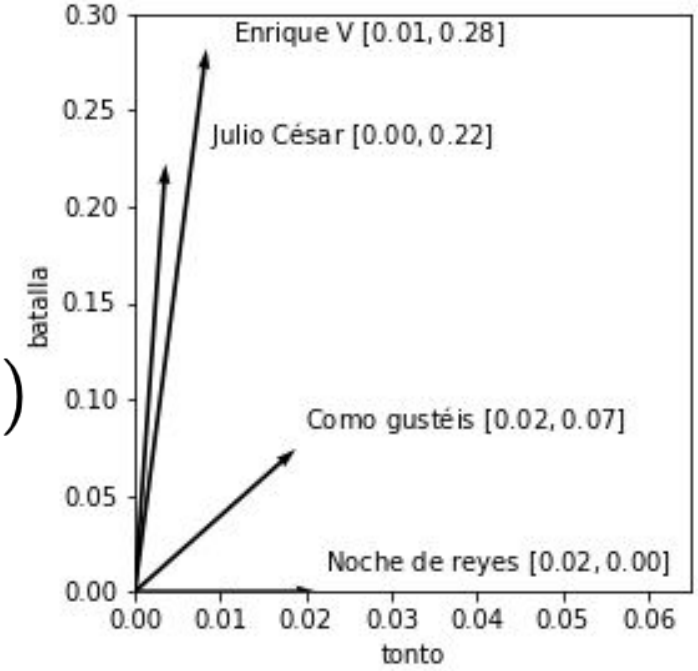
Similaridad entre vectores

$$\text{cosine}(\text{Julio César}, \text{Enrique V})$$

$$= \frac{0*0.01+0.22*0.28}{\sqrt{0^2+0.22^2}*\sqrt{0.01^2+0.28^2}} \approx 0.999$$

$$\text{cosine}(\text{Julio César}, \text{Noche de Reyes})$$

$$= \frac{0*0.02+0.22*0}{\sqrt{0^2+0.22^2}*\sqrt{0.01^2+0^2}} = 0$$



Coseno como métrica de distancia

Esta métrica es la que mejor se comporta con datos de texto y que representa de mejor manera la similaridad entre documentos.