Embeddings

Matemáticas para las ciencias de la computación

Miguel Angel Soto Hernandez

Contenido

- Problemática
- Acercamientos
- ¿Qué es un embedding?
- Matemáticas detrás de los embeddings
  - Algoritmo Skip-gram
- Tipos de embeddings
  - Texto
  - Imágenes
  - Embeddings y filtrado colaborativo

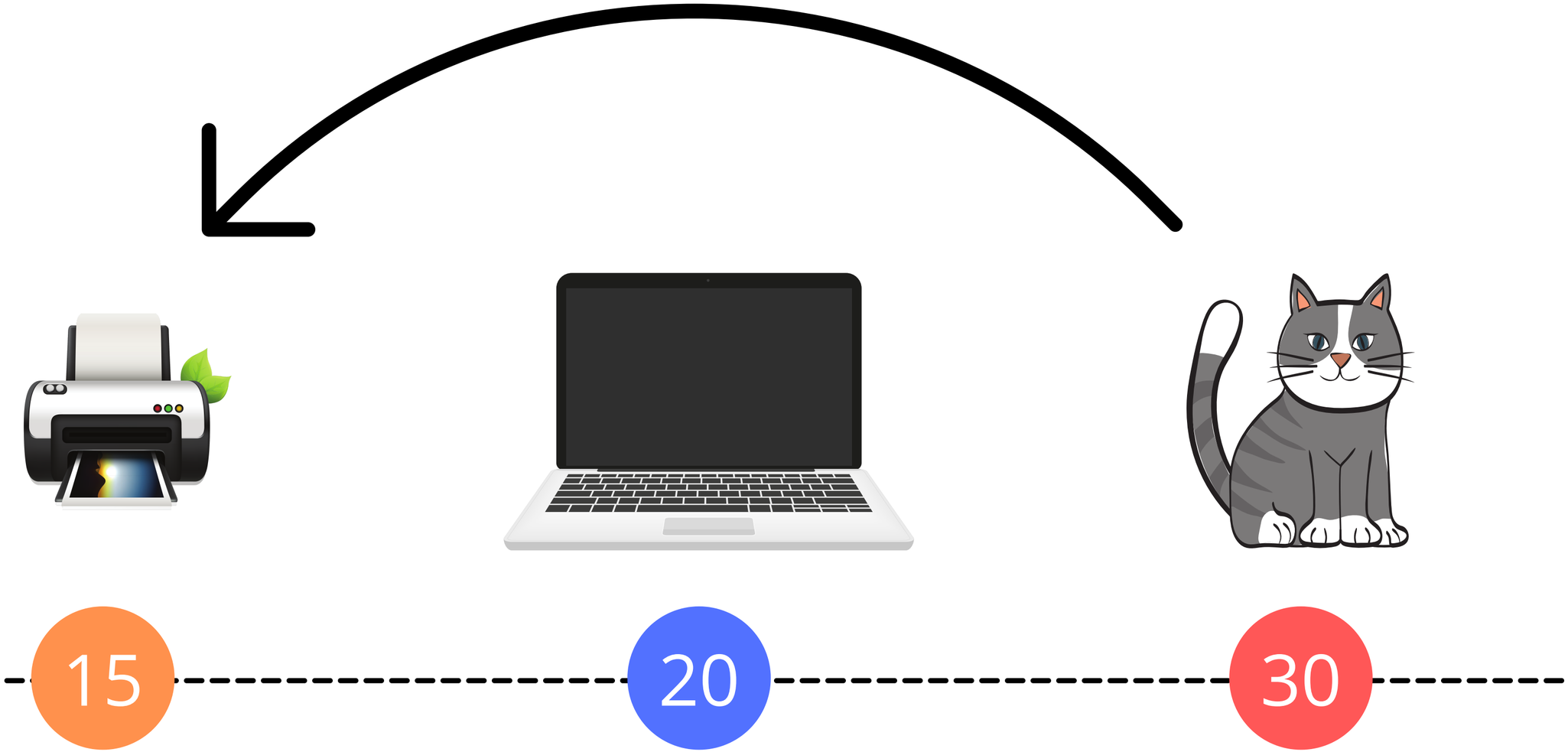
Problemática

¿Cómo hacer comprender a una máquina algo que está escrito de manera natural?


COMMUNICATION
LANGUAGE
RULES
SYNTAX
SPEECH
BOOK
GRAMMAR
IDIOMS
EDUCATION
DICTIONARY
PREPOSITIONS
UNDERSTAND

Yo solo sé que no sé nada

20 30 15 40 20 15 50


15
20
30

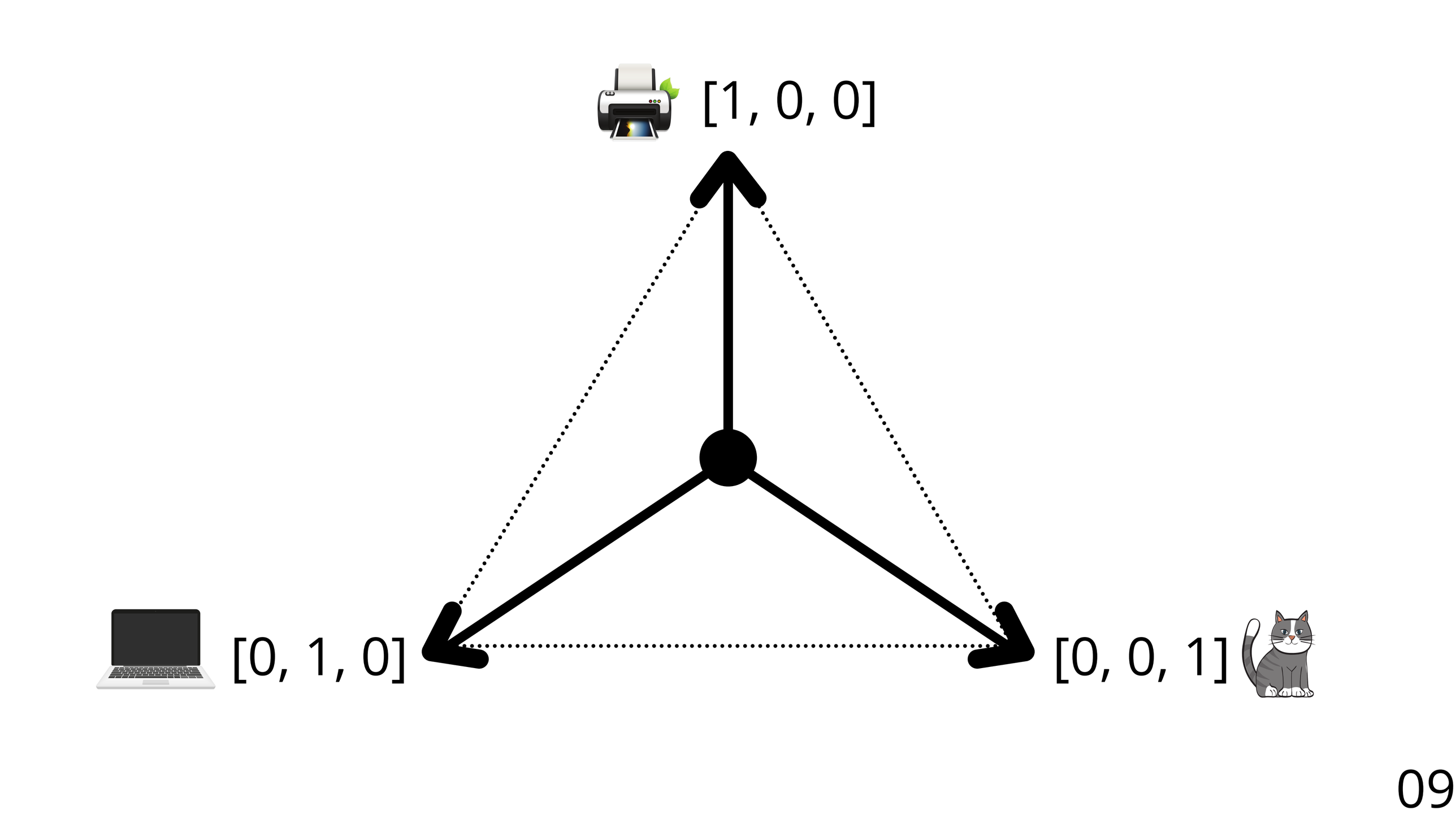
One hot-encoding

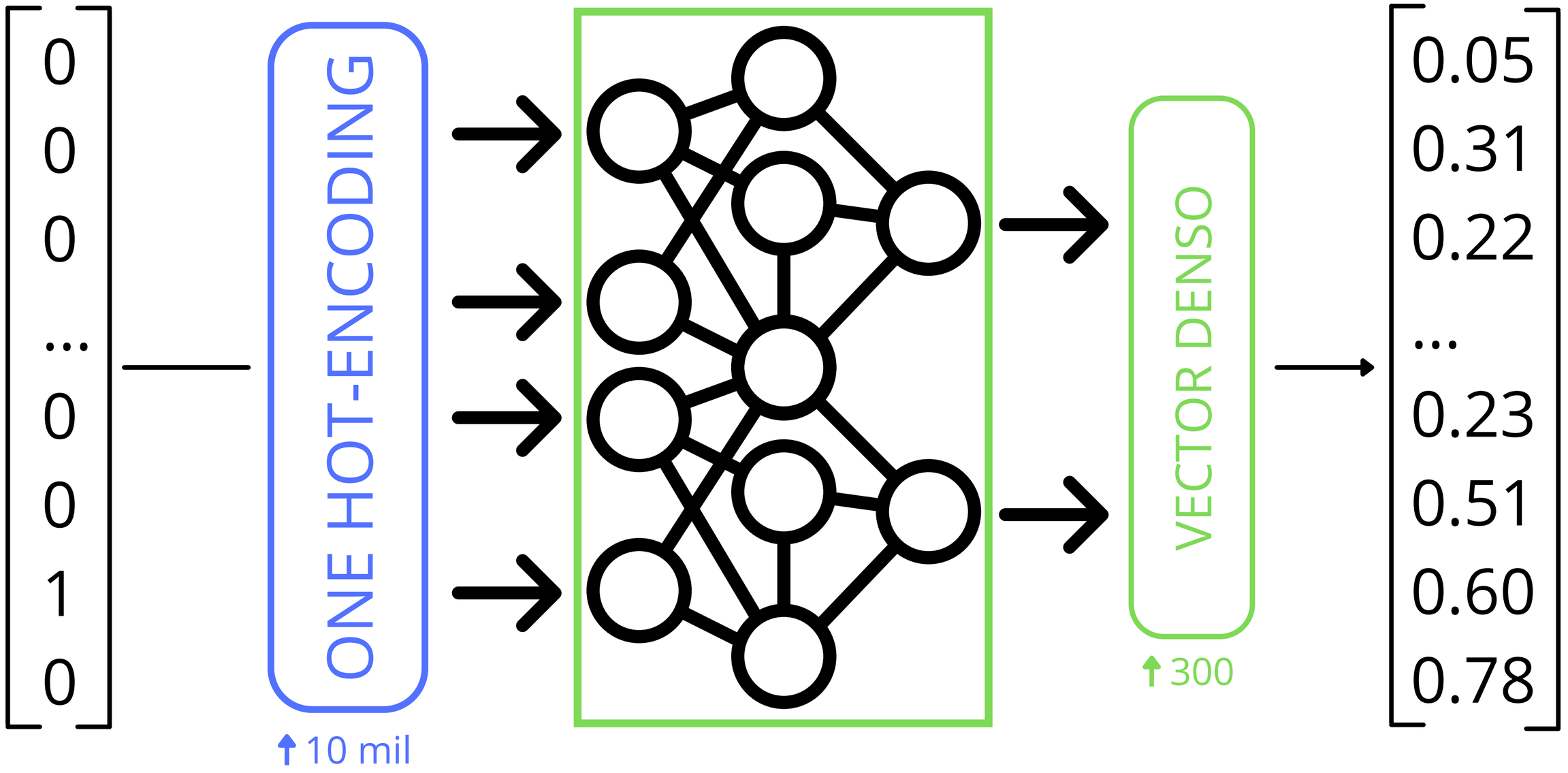
[[0] Matemáticas
[0] Para
[0] Las
[0] Ciencias
[0] De
[0] La
[1] Computación]

Impresora = 0 → [1, 0, 0]

Laptop = 1 → [0, 1, 0]

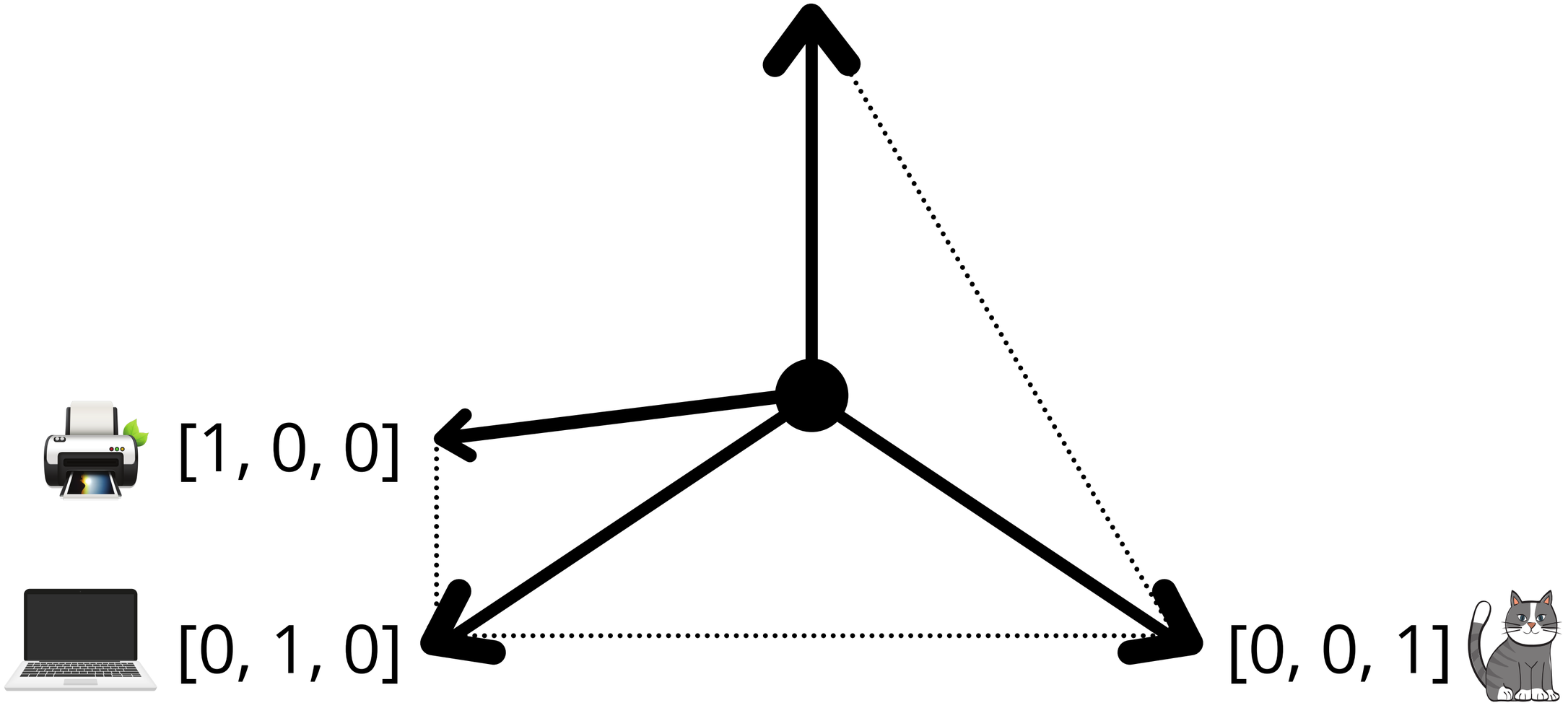
Gato = 2 → [0, 0, 1]

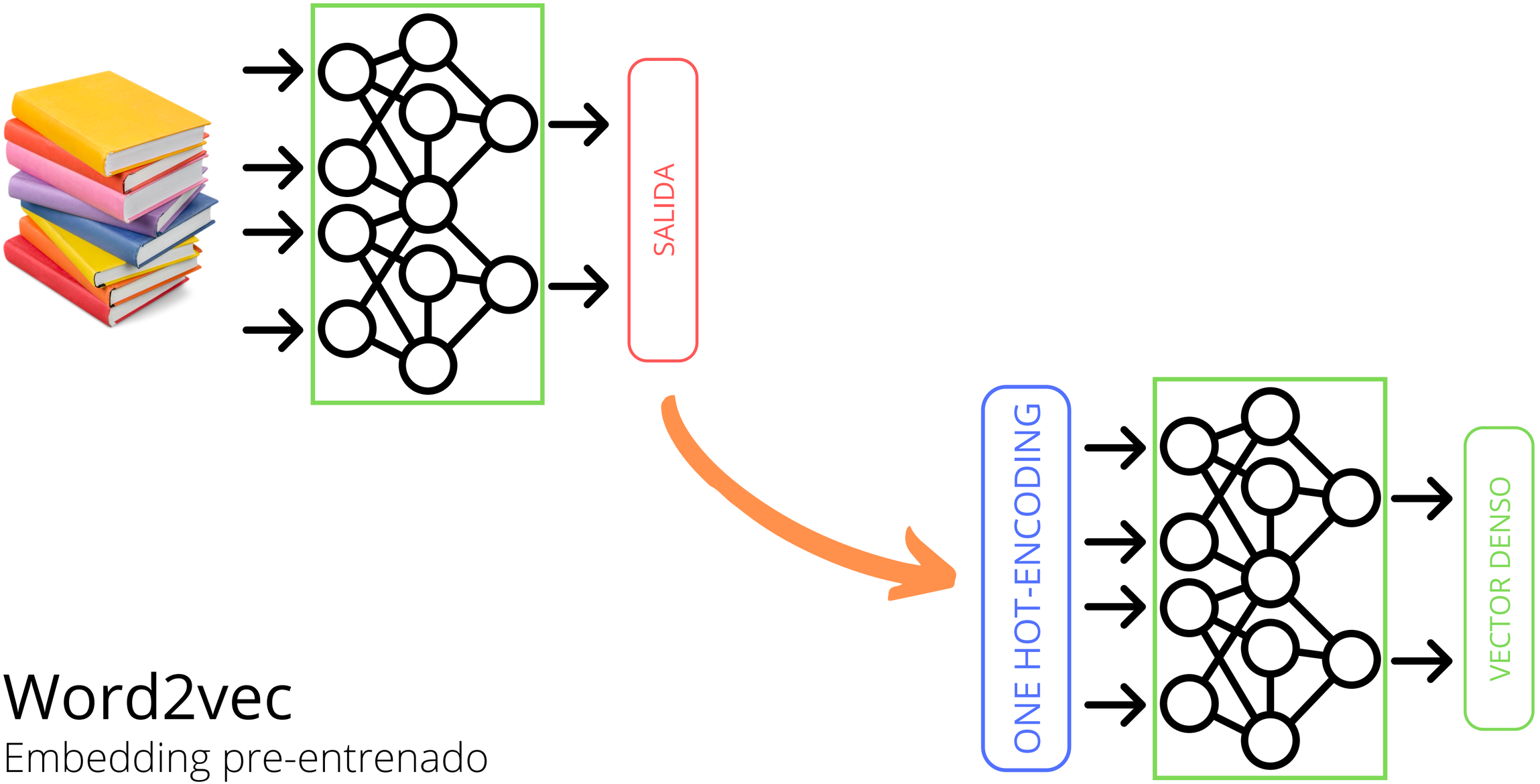

[1, 0, 0]
[0, 1, 0]
[0, 0, 1]


0
0
0
...
0
0
1
0
ONE HOT-ENCODING
10 mil
VECTOR DENSO
300
0.05
0.31
0.22
...
0.23
0.51
0.60
0.78

Embedding

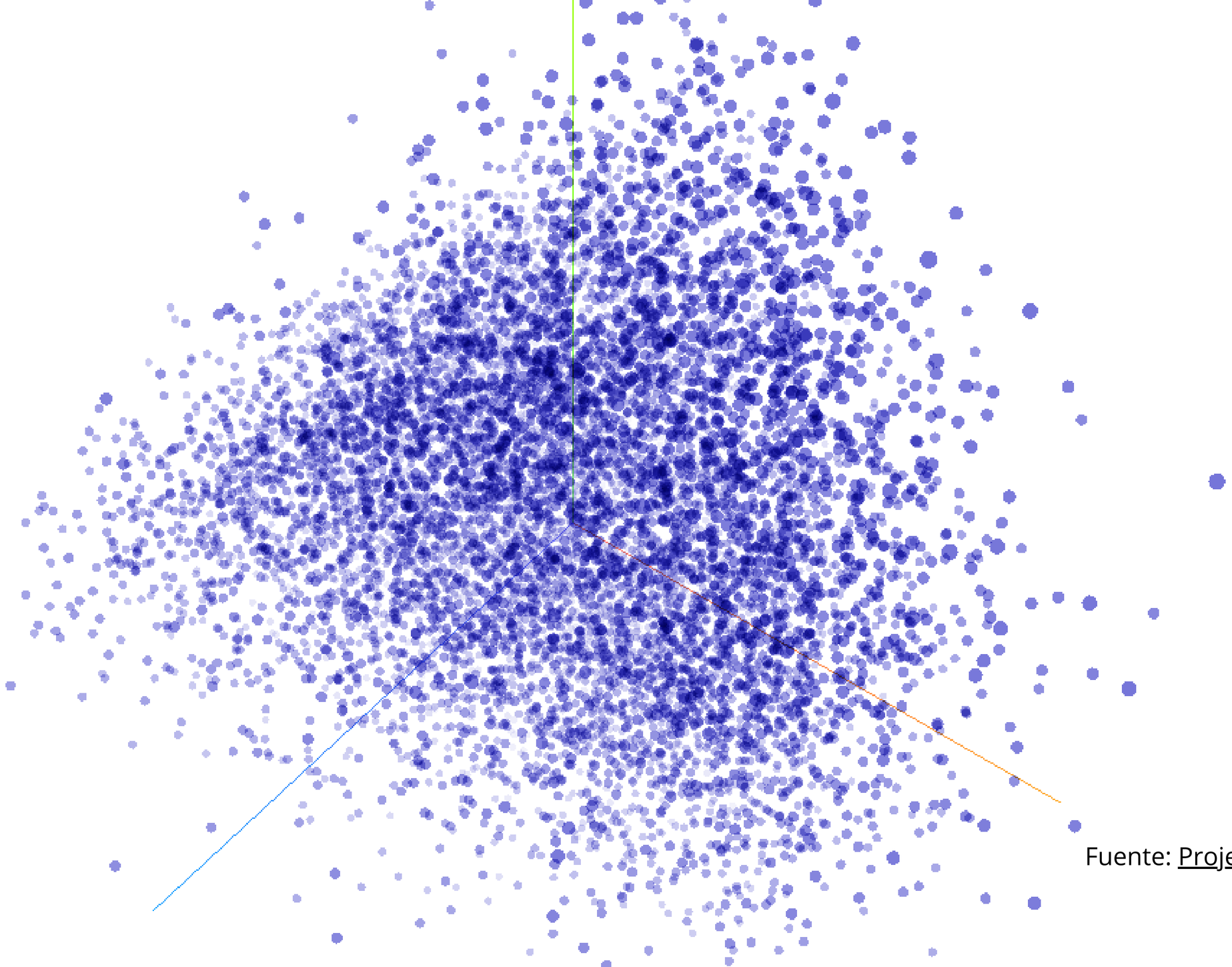
Es una traducción de un vector de alta dimensión en un espacio de baja dimensión.


[1, 0, 0]
[0, 1, 0]
[0, 0, 1]



Word2vec

Embedding pre-entrenado



Fuente: Project Tensorflow

Embeddings de texto

Se pueden representar palabras de una oración en diferentes formas:

- Vector disperso muy grande (alta dimensión)
- Vector pequeño en comparación, pero denso

Word2vec	GloVe	ELMo	FastText
Google	Stanford	AllenNLP	Facebook

Embeddings de imagen

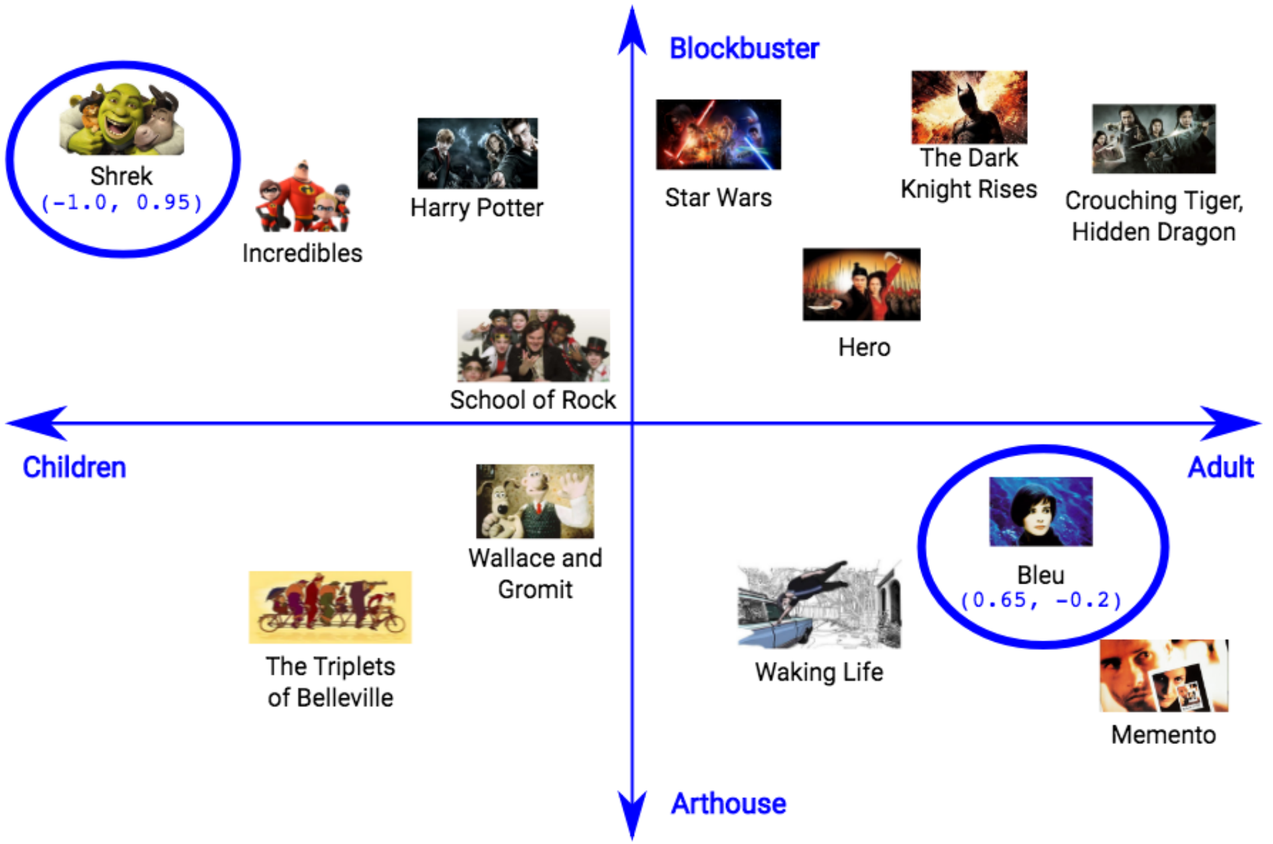
Los sistemas de procesamiento de imágenes funcionan con conjuntos de datos ricos y de alta dimensión que tienen intensidades individuales de píxel sin formato. Sin embargo, una imagen en su forma densa sin formato no es útil para algunas tareas.

Inception
Google

ResNet
Microsotf

NASNet
Google

Embeddings y el filtrado colaborativo



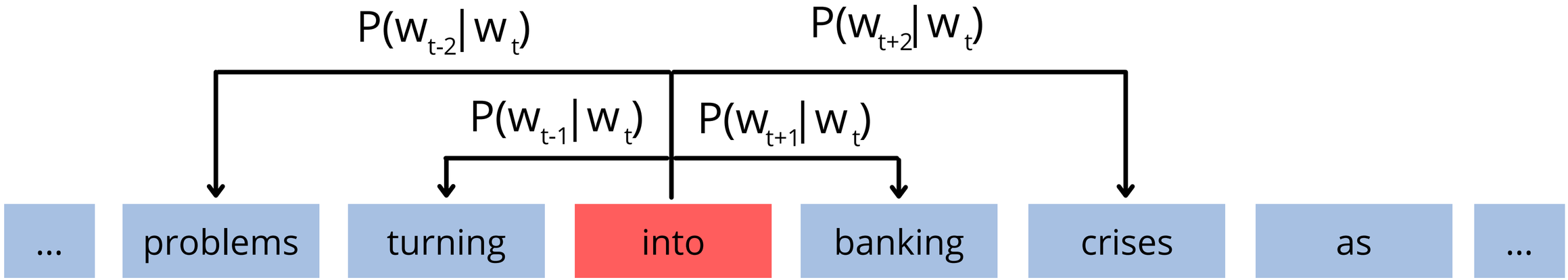
Semántica

Parte de la lingüística que estudia el significado de las expresiones lingüísticas.

Semántica Distribucional

El significado de una palabra viene dado por las palabras que aparecen con frecuencia cerca de ella.

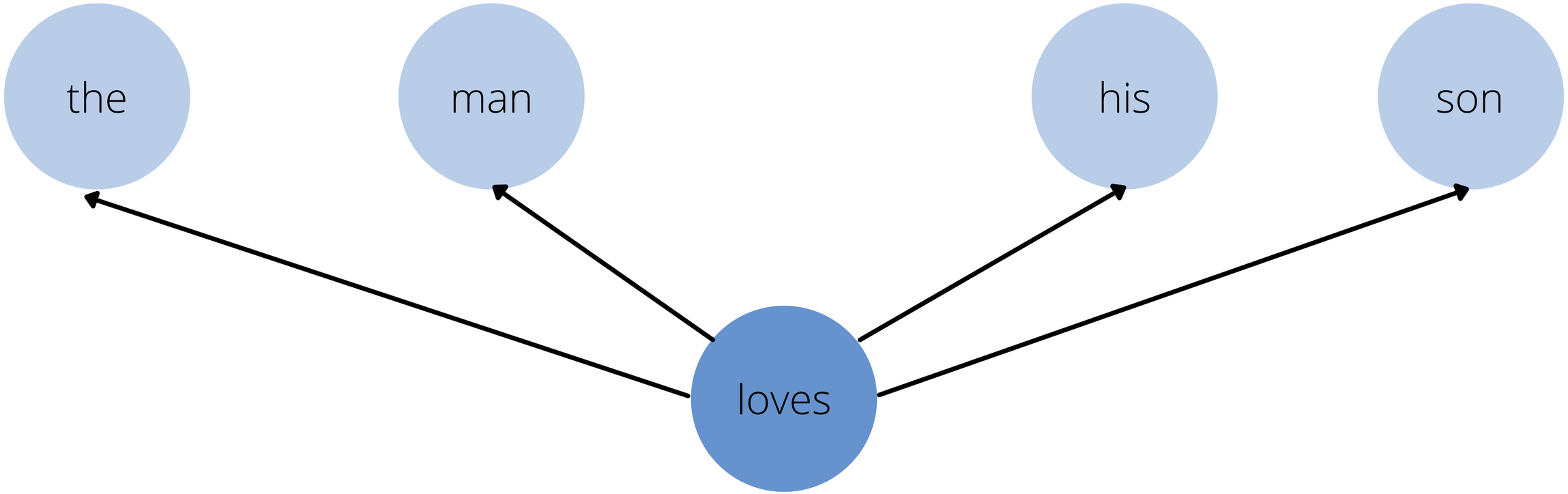
Algoritmo Skip-gram



fuera de las palabras de contexto en ventanas de tamaño 2

centro de la palabra en la posición t

fuera de las palabras de contexto en ventanas de tamaño 2



$$L(\theta) = \prod_{t=1}^{T} \prod_{\substack{-m \le j \le m \\ j \neq 0}} P(w_{t+j} \mid w_t; \theta)$$

P('the', 'man', 'his', 'son' | 'loves')

P('the'| 'loves') * P('man'| 'loves') *
P('his'| 'loves') * P('son'| 'loves')

$$J(\theta) = -\frac{1}{T}\log L(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-m\leq j\leq m \\ j\neq 0}} \log P\left(w_{t+j} \mid w_t; \theta\right)$$

¿Pero cómo calculamos $P(w_{t+1} | w_t; \theta)$?

Vamos a usar dos vectores densos por cada palabra w:

$\mathbf{v}_w$ cuando w es la palabra central

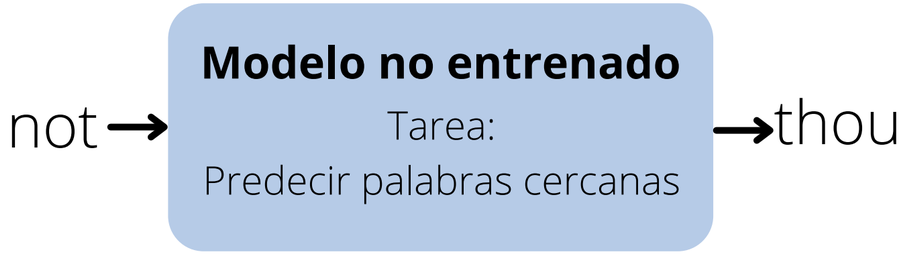
$\mathbf{u}_w$ cuando w es una palabra de contexto

softmax

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \longrightarrow P(O = o \mid C = c) = \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

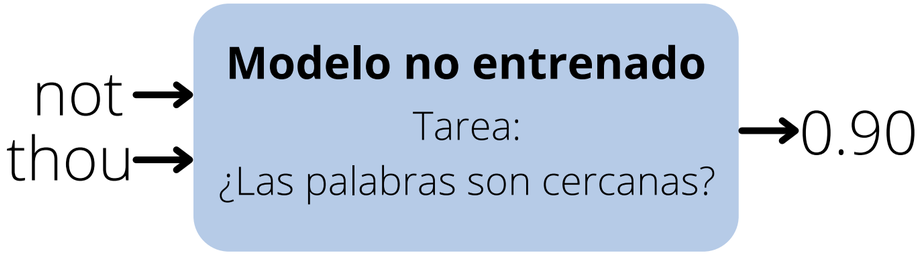
Muestreo negativo

Cambiar de la tarea



$$P(O = o \mid C = c) = \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

A



$$P(D = 1 \mid w_c, w_o) = \sigma(\mathbf{u}_o^\top \mathbf{v}_c),$$

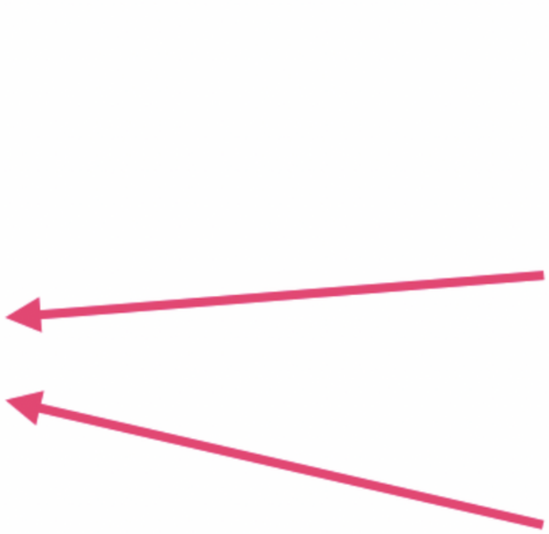
$$\sigma(x) = \frac{1}{1 + \exp(-x)}.$$

input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine

input word	output word	target
not	thou	**1**
not	shalt	**1**
not	make	**1**
not	a	**1**
make	shalt	**1**
make	not	**1**
make	a	**1**
make	machine	**1**

Pick randomly from vocabulary
(random sampling)

input word	output word	target
not	thou	1
not	aaron	0
not	taco	0
not	shalt	1
not	make	1



Word	Count	Probability
aardvark		
aarhus		
aaron		
taco		
thou		
zyzzyva		

$$P(w^{(t+j)} \mid w^{(t)}) =$$

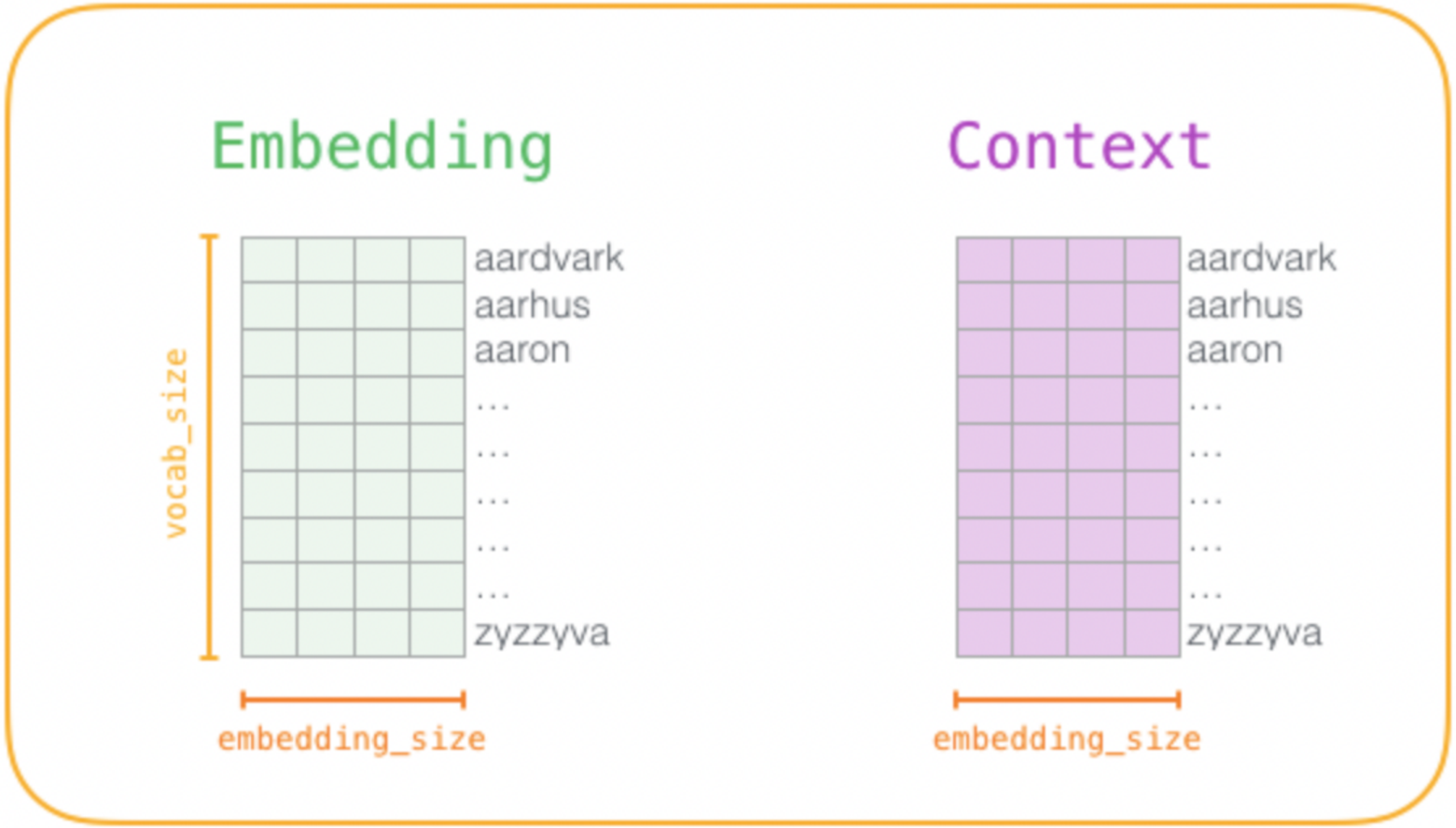
$$P(D = 1 \mid w^{(t)}, w^{(t+j)}) \prod_{k=1,\, w_k \sim P(w)}^{K} P(D = 0 \mid w^{(t)}, w_k).$$

Aquí el cálculo del gradiente en cada paso del entrenamiento ya no está relacionado con el tamaño del diccionario sino esta linealmente relacionado con k

$$
\begin{aligned}
-\log P(w^{(t+j)} \mid w^{(t)}) &= -\log P(D=1 \mid w^{(t)}, w^{(t+j)}) - \sum_{k=1,\, w_k \sim P(w)}^{K} \log P(D=0 \mid w^{(t)}, w_k) \\
&= -\log \sigma\left(\mathbf{u}_{i_{t+j}}^\top \mathbf{v}_{i_t}\right) - \sum_{k=1,\, w_k \sim P(w)}^{K} \log\left(1-\sigma\left(\mathbf{u}_{h_k}^\top \mathbf{v}_{i_t}\right)\right) \\
&= -\log \sigma\left(\mathbf{u}_{i_{t+j}}^\top \mathbf{v}_{i_t}\right) - \sum_{k=1,\, w_k \sim P(w)}^{K} \log \sigma\left(-\mathbf{u}_{h_k}^\top \mathbf{v}_{i_t}\right).
\end{aligned}
$$

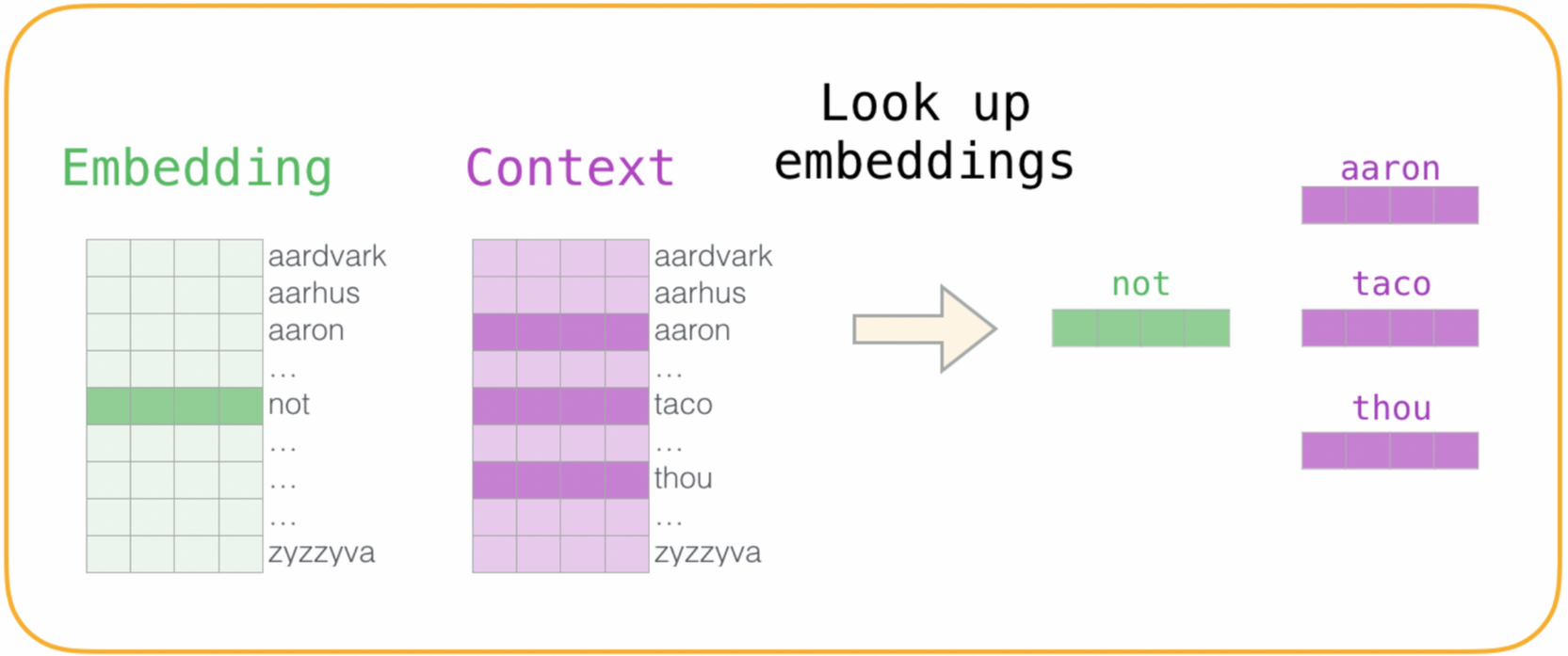
K < V

Proceso de entrenamiento



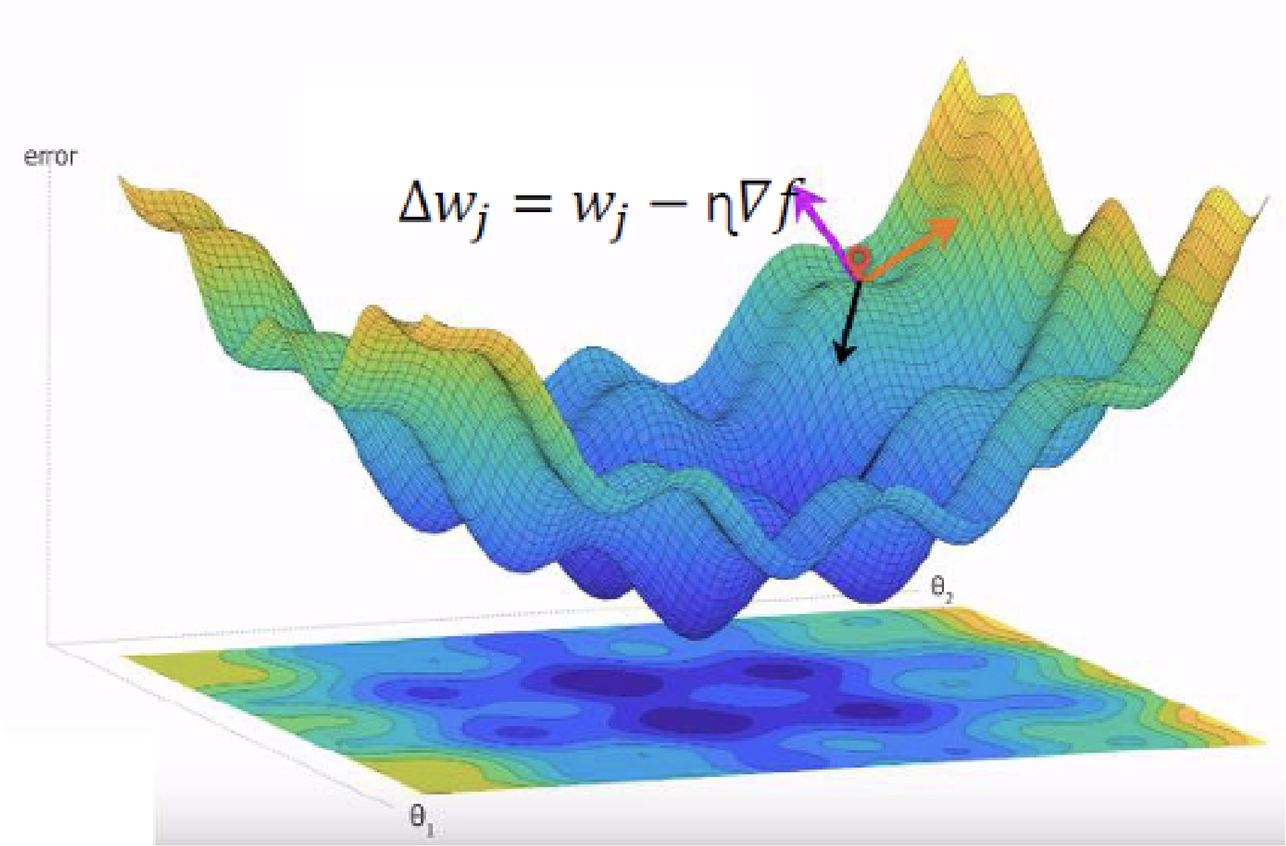
dataset

input word	output word	target
not	thou	1
not	aaron	0
not	taco	0
not	shalt	1
not	mango	0
not	finglonger	0
not	make	1
not	plumbus	0
...	...	...

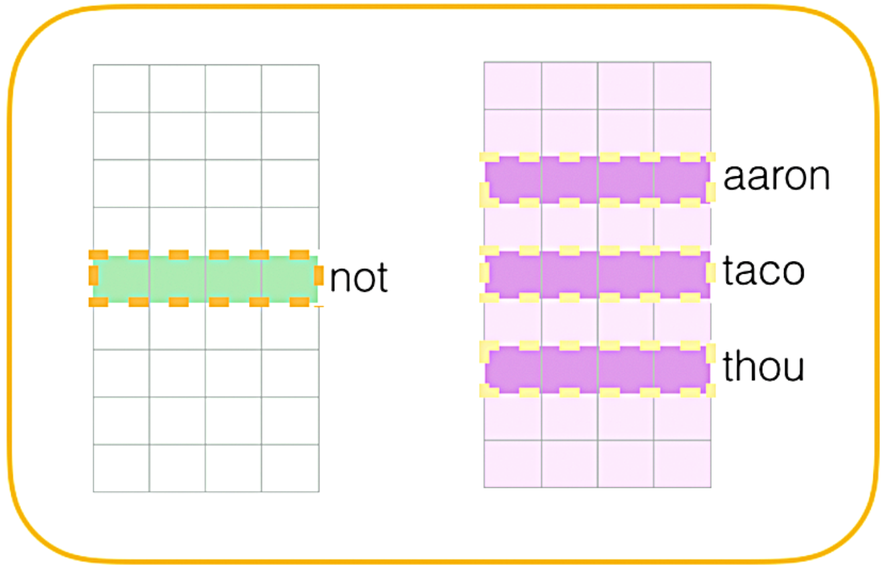


Gradiente descendiente

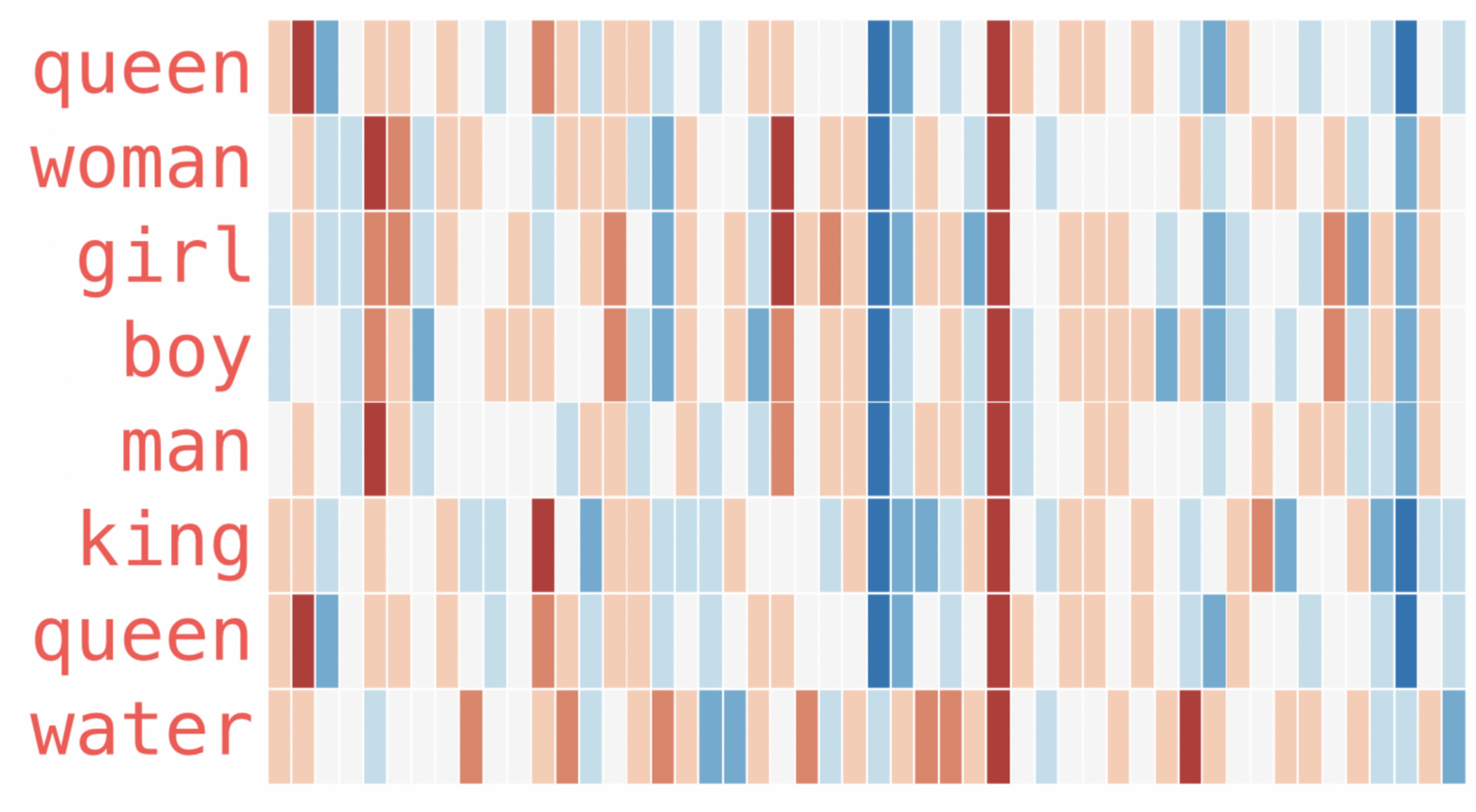
$$\begin{bmatrix} \frac{\partial error}{\partial \theta_1} \\ \frac{\partial error}{\partial \theta_2} \\ \vdots \\ \frac{\partial error}{\partial \theta_k} \end{bmatrix} = \nabla f$$



input word	output word	target	input • output	sigmoid()	Error
not	thou	1	0.2	0.55	0.45
not	aaron	0	-1.11	0.25	-0.25
not	taco	0	0.74	0.68	-0.68



Update Model Parameters


queen
woman
girl
boy
man
king
queen
water

Similitud coseno

$$\text{soft\_cosine}_1(a, b) = \frac{\sum_{i,j}^{N} s_{ij} a_i b_j}{\sqrt{\sum_{i,j}^{N} s_{ij} a_i a_j} \sqrt{\sum_{i,j}^{N} s_{ij} b_i b_j}},$$

Referencias

Le, Q. V., y Mikolov, T. (2014). Distributed representations of sentences and documents. CoRR, abs/1405.4053 . Descargado de http://arxiv.org/abs/1405.4053 Messenger. (2011). Descargado de www.messenger.com/ Mikolov, T. (2013).

Learning representations of text using neural networks. Mikolov, T., Chen, K., Corrado, G., y Dean, J. (2013). Efficient estimation of word representations in vector space. CoRR, abs/1301.3781 . Descargado de http://arxiv.org/abs/1301.3781

Mikolov, T., Joulin, A., y Baroni, M. (2015). A roadmap towards machine intelligence. CoRR, abs/1511.08130 . Descargado de http://arxiv.org/abs/1511.08130 Mikolov, T., Sutskever, I., Chen, K., Corrado, G., y Dean, J. (2013).

Distributed representations of words and phrases and their compositionality. CoRR, abs/1310.4546 . Descargado de http://arxiv.org/abs/1310.4546

Mikolov, T., Yih, W., y Zweig, G. (2013). Linguistic regularities in continuous space word representations. HLT-NAACL, 746–751. Mitkov, R. (2003). The oxford handbook of computational linguistics. Descargado de https://books.google.cl/books?id=yl6AnaKtVAkC&pg=PA754&redir_esc=y#v= onepage&q&f=false

Gracias