Word embeddings

Thursday, August 10, 2023 7:54 PM

- Tf-idf → frecuencias
  - → contador, lleva la cuenta en el corpus
  - → Matriz resultante depende de $|V|$ donde $|V|$: Tamaño del vocabulario
  - → "Maldición de la dimensionalidad"

poca data
muchas features (e.g. tfidf - 16000 columnas)
"features"

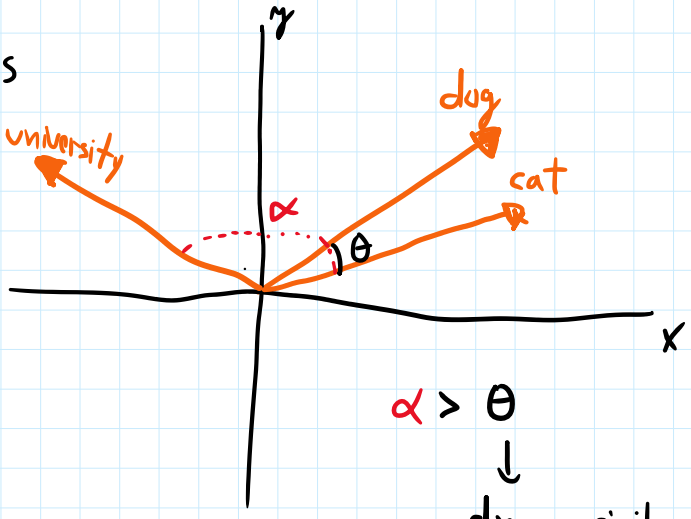
Word Embeddings: → vector

"dog" → | 0.5 | 0.98 | -0.1 | 0.25 | (dense) no ceros

N=4 N es definida por nosotros

"cat" → | 0.1 | -0.9 | 0.981 | 0.31 |

"university" → | 0.01 | 0.008 | 0.5 | -0.21 |



$\alpha > \theta$

↓

dog es similar a cat

Word embeddings → embedding estáticos

→ diccionario con repre. vectoriales de cada palabra en V.

modelo["cat"] = [| | | |]

modelo["dog"] = [| | | |]

<=> "cat" y "dog" ∈ V

→ dinámicos (Transformers (e.g. BERT)

→ dinámicos (Transformers (e.g BERT)
↳ SOTA en NLP

Word2Vec → word → vector (estático)

↳ NO es un algoritmo/modelo
↳ Es una librería/paquete
↳ Skipgram, CBOW

Skipgram:

D1: [Hola, mundo, como, están, en, el, planeta, tierra]
window
w_1 w_2 w_3 w_4 w_5 w_6 w_7 w_8

$$p(w_t, w_3) = p(+|w, c)$$
$$= p(-|w, c) = 1 - p(+|w, c)$$

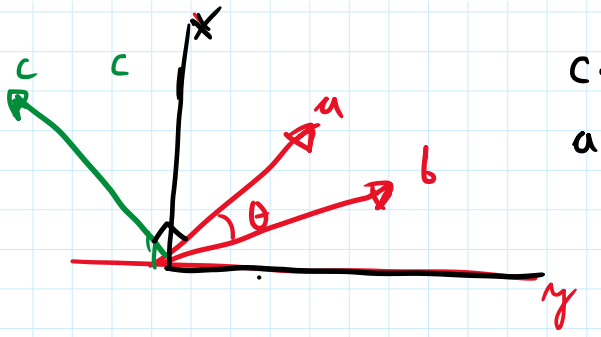
$p(+|w, c)$ = similaridad (w, c)
↳ producto punto (dot product) $\in \mathbb{R}$

$$\langle a, b \rangle \cdot \langle c, d \rangle = a \cdot c + b \cdot d = \mathbb{R}$$

$$\vec{a} \cdot \vec{b} = \cos\theta \rightarrow \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} = \cos\theta \Rightarrow \vec{a} \cdot \vec{b} = |\vec{a}||\vec{b}|\cos\theta$$

θ = ángulo entre $\vec{a}, \vec{b}$

$$|\vec{v}| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} \in \mathbb{R}$$



$c \cdot a = 0$

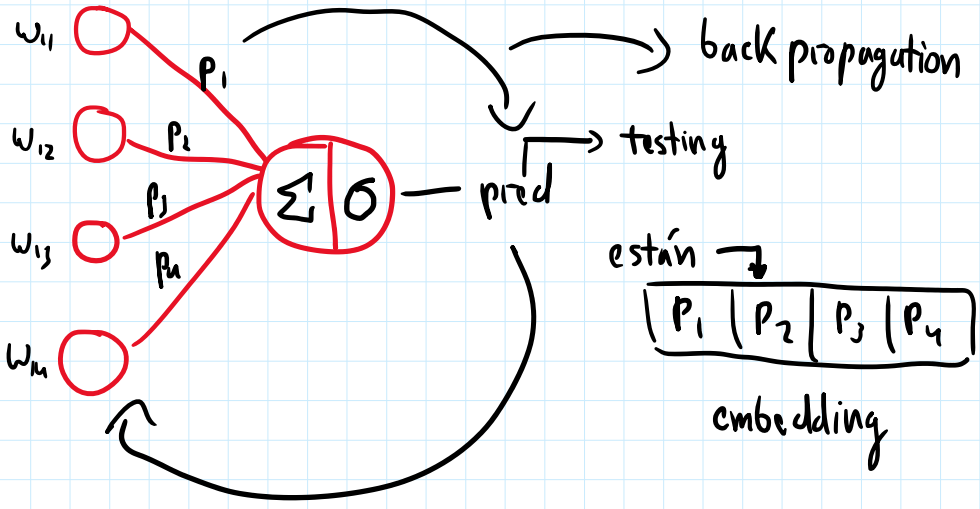
$a \cdot a = 1$

$p(+|w,c) = \sigma(w \cdot c)$ $\quad \sigma(x) = \frac{1}{1+e^{-x}}$ => sigmoide

$p(+|w,c) = \prod_{i=1}^{L} \sigma(w \cdot c_i)$ donde $L = 4$ window = 4

$\sigma(w \cdot c_1) \cdot \sigma(w \cdot c_2) \cdot \ldots$

están → | 0.1 | 0.3 | 0.8 | 0.4 |



extraer muestras negativas → Negative sampling

Skipgram with Negative sampling → regresión logística (sigmoide)

OOV (Out-of-Vocabulary) → Word2Vec (problema)

- random vector
- random (seed) → fijas
- otro modelo → Fasttext (n-grams)

"Hola mundo ESAN"

n=2 (bi-grams)

[Hola_mundo], [mundo_ESAN]

n=2

Hola

[<H, ol, a>] → ola

ol + a = ola → [vector]

$D_1 = [w_1, w_2, w_3]$ →

w_1 →

w_2 →

w_3 →

=

→

+

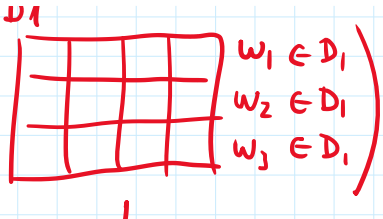
+

3

→ SVM(

SVM(D1, D2, D3) → vector ✓

SVM(D1 $w_1 \in D_1$

SVM (

) X

↳ a vector