IIC3670 Procesamiento de Lenguaje Natural

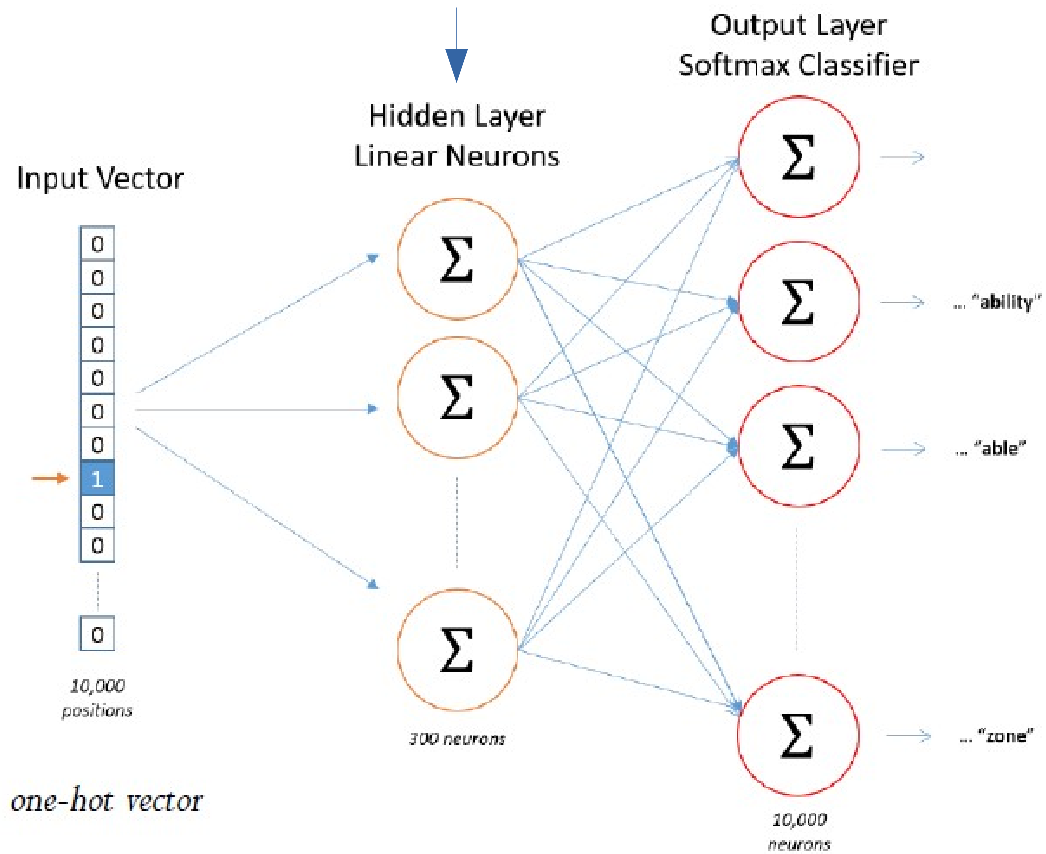
https://github.com/marcelomendoza/IIC3670

- WORD2VEC -

Word vectorization

Capa con n neuronas, n << V

palabra objetivo



Palabra de contexto

Vocabulario

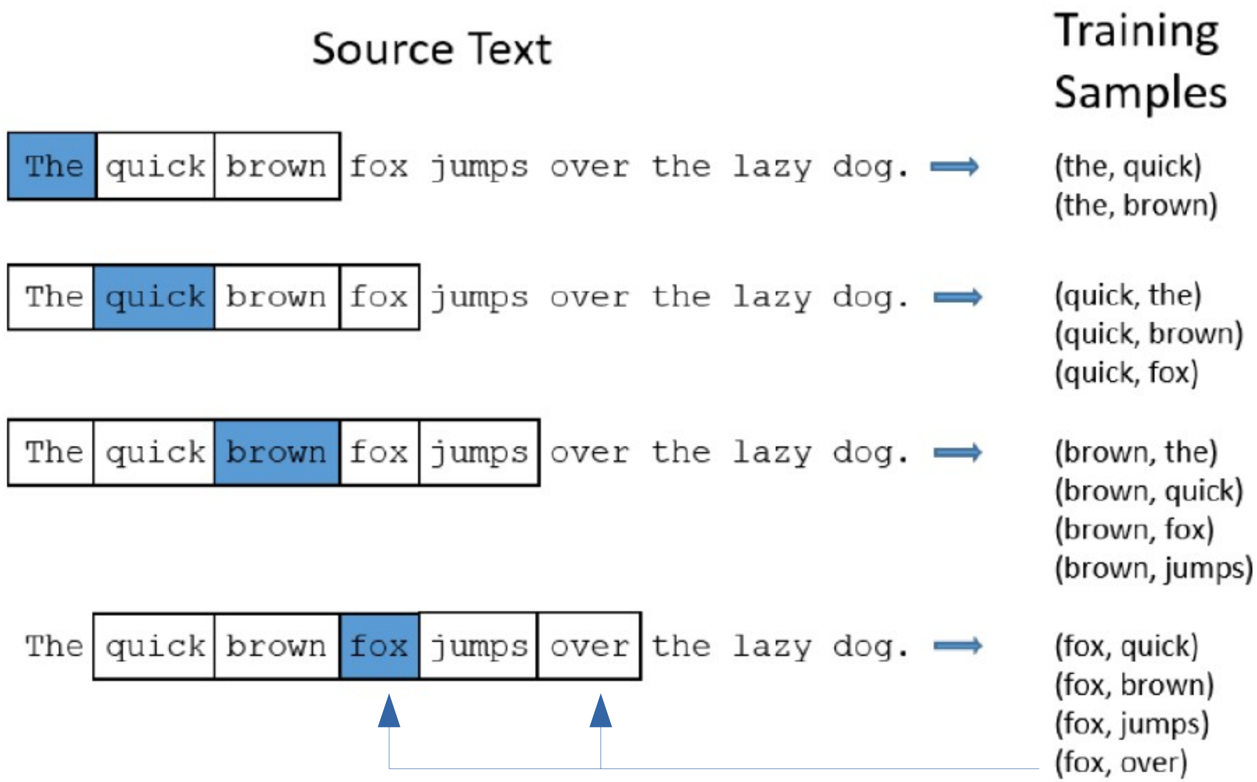
Vocabulario

Word vectorization

La red se entrena
mostrando *bi-gramas (objetivo,contexto)*

- Skip-grams: entrenamiento de la red en base a pares de palabras.

Una ventana deslizante
recorre el texto



Los *bi-gramas* no necesariamente
son consecutivos: ***skip-grams***

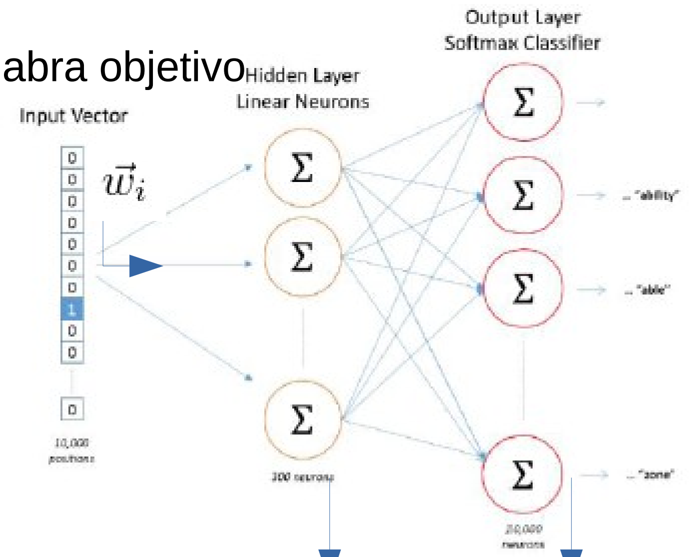
entrada

salida

Word vectorization

- Skip-grams: entrenamiento de la red en base a pares de palabras.

Palabra objetivo



$\vec{w}_i$

W_{in}

Palabra de contexto

$\vec{w}_{i+j}$

W_{out}

Función de pérdida:

$$\mathcal{L}_{SG} = -\frac{1}{|S|} \cdot \sum_{i=1}^{|S|} \sum_{-c \leq j \leq c,\ j \neq 0} \log(p(w_{i+j}|w_i))$$

$|S|$: # training skip-grams (pares)

$-c \leq j \leq c$: tamaño de la ventana de contexto (2c)

W_{in} o W_{out} pueden ser usados como *word embeddings*

- Skip-grams: Como generar el training set de pares de palabras.

Tratando el desbalance entre skip-grams y pares no observados

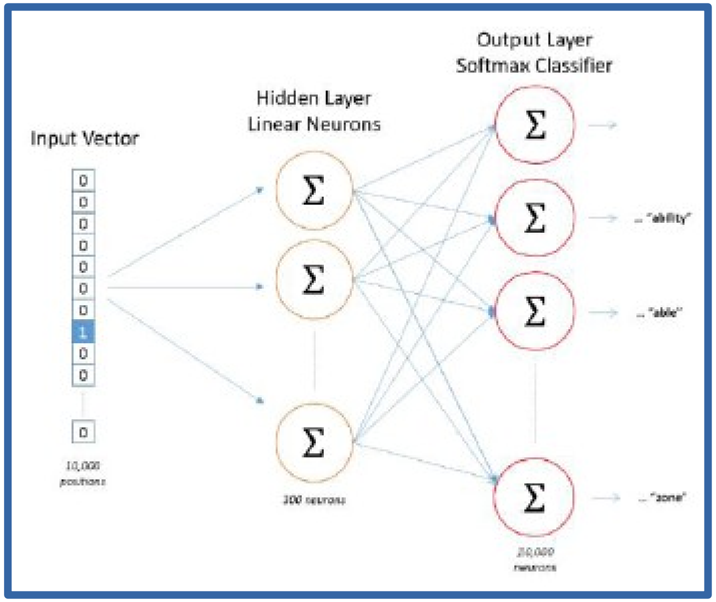
Negative sampling:

- Seleccionamos aleatoriamente k ejemplos negativos (palabras que no están en C). Si no hiciéramos esto, **todas** las palabras que no están en C serían ejemplos negativos ($k = 5$).
- La probabilidad de seleccionar una palabra como ejemplo negativo es:

$$P(w_i) = \frac{f(w_i)^{\beta}}{\sum_{j=0}^{n} f(w_i)^{\beta}}$$

donde $0 < \beta < 1$ ($\beta \approx \frac{3}{4}$).

Word vectorization: skip-grams



¿Por qué funciona?

$$P(+|w,c) = \sigma(\mathbf{c}\cdot\mathbf{w}) = \frac{1}{1+\exp(-\mathbf{c}\cdot\mathbf{w})}$$

$$P(-|w,c) = 1 - P(+|w,c)$$

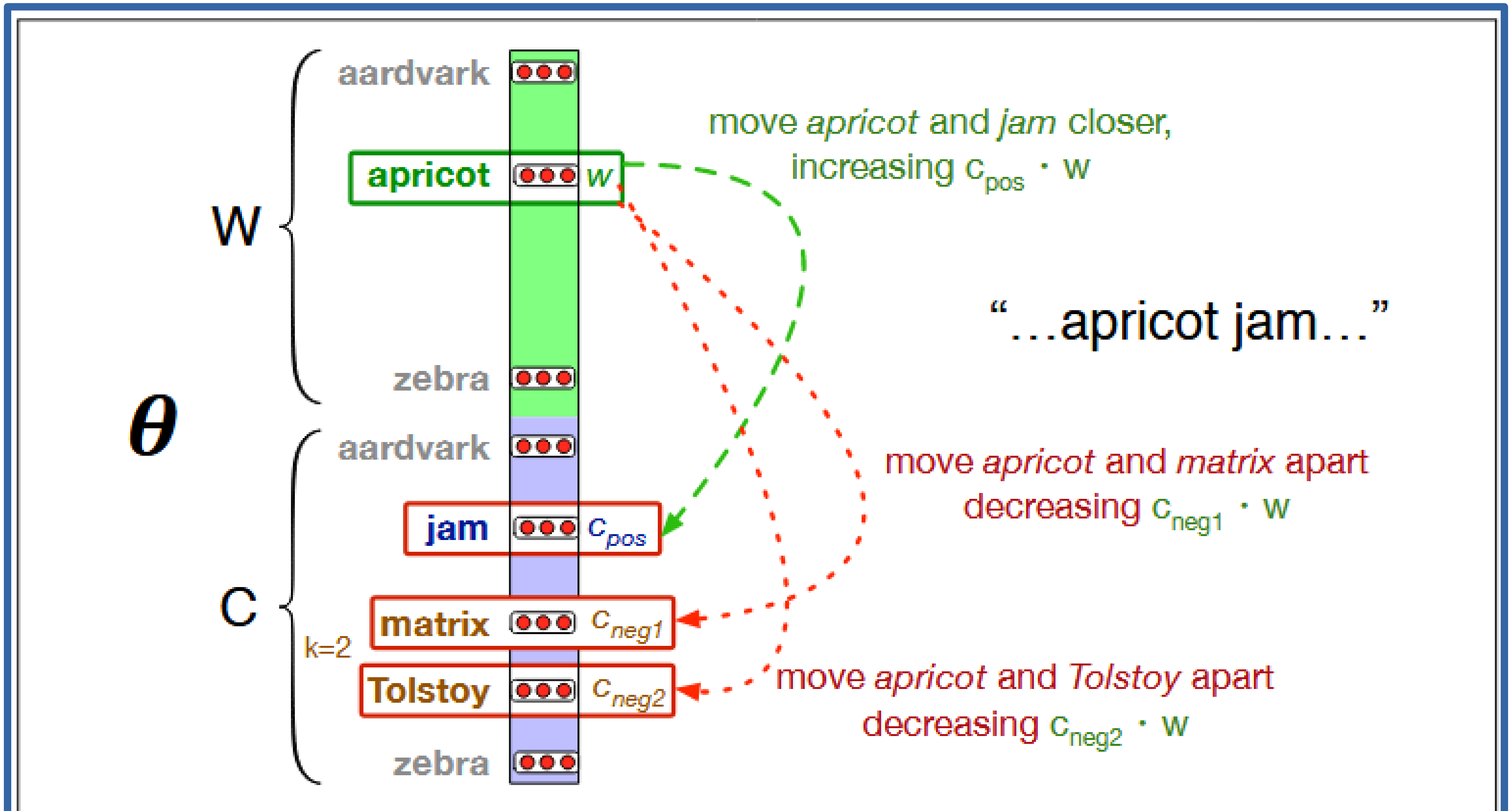
$$= \sigma(-\mathbf{c}\cdot\mathbf{w}) = \frac{1}{1+\exp(\mathbf{c}\cdot\mathbf{w})}$$

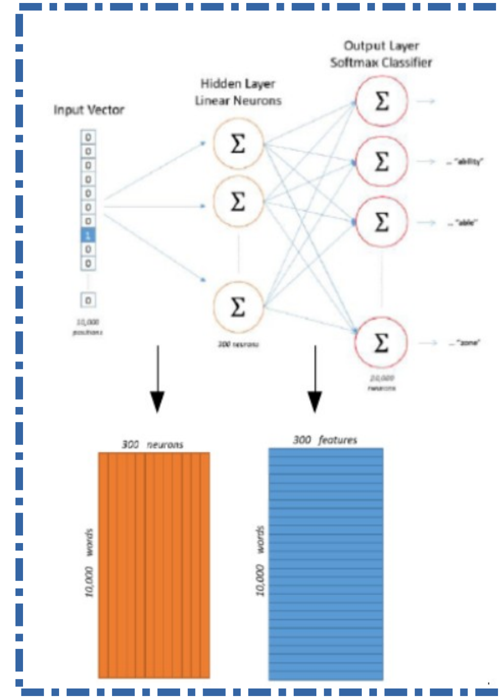
$$Similarity(w,c) \approx \mathbf{c}\cdot\mathbf{w}$$

Palabras relacionadas se alinean (disminuyen su distancia angular) mientras que palabras no relacionadas aumentan su distancia angular (se distancian).

Word vectorization: skip-grams

Palabras relacionadas se alinean (disminuyen su distancia angular) mientras que palabras no relacionadas aumentan su distancia angular (se distancian).





W_{in} W_{out}

Puede tomar cualquiera como embedding

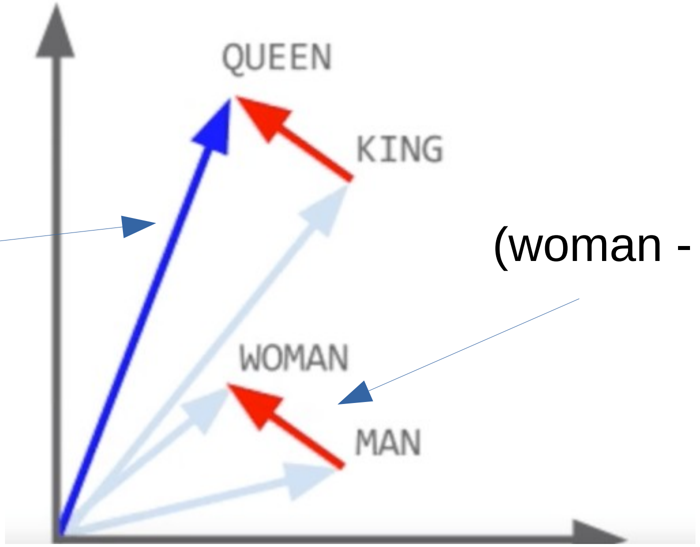
¿Por qué?

Word vectorization

- Operadores en word2vec: word analogies

king + (woman - man)

(woman - man)



$$\underset{b^* \in V}{\arg\max} \left(sim \left(b^*, b - a + a^* \right) \right)$$

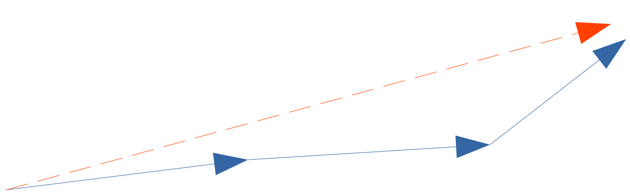
Levy & Goldberg, Linguistic Regularities in Sparse and Explicit Word Representations, ACL’14.

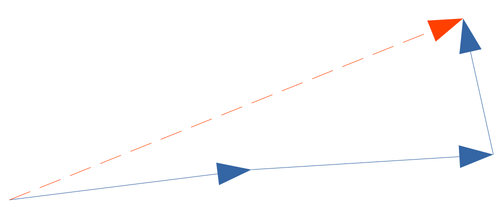
Word vectorization

- Operadores en word2vec: doesnt_match(['king', 'george', 'stephen', 'truck'])

$$\arg\max_{w} f(w) = \left\| \sum_{v \in L \setminus w} \vec{v} \right\|, \quad \forall w \in L$$

Cuarto excluído





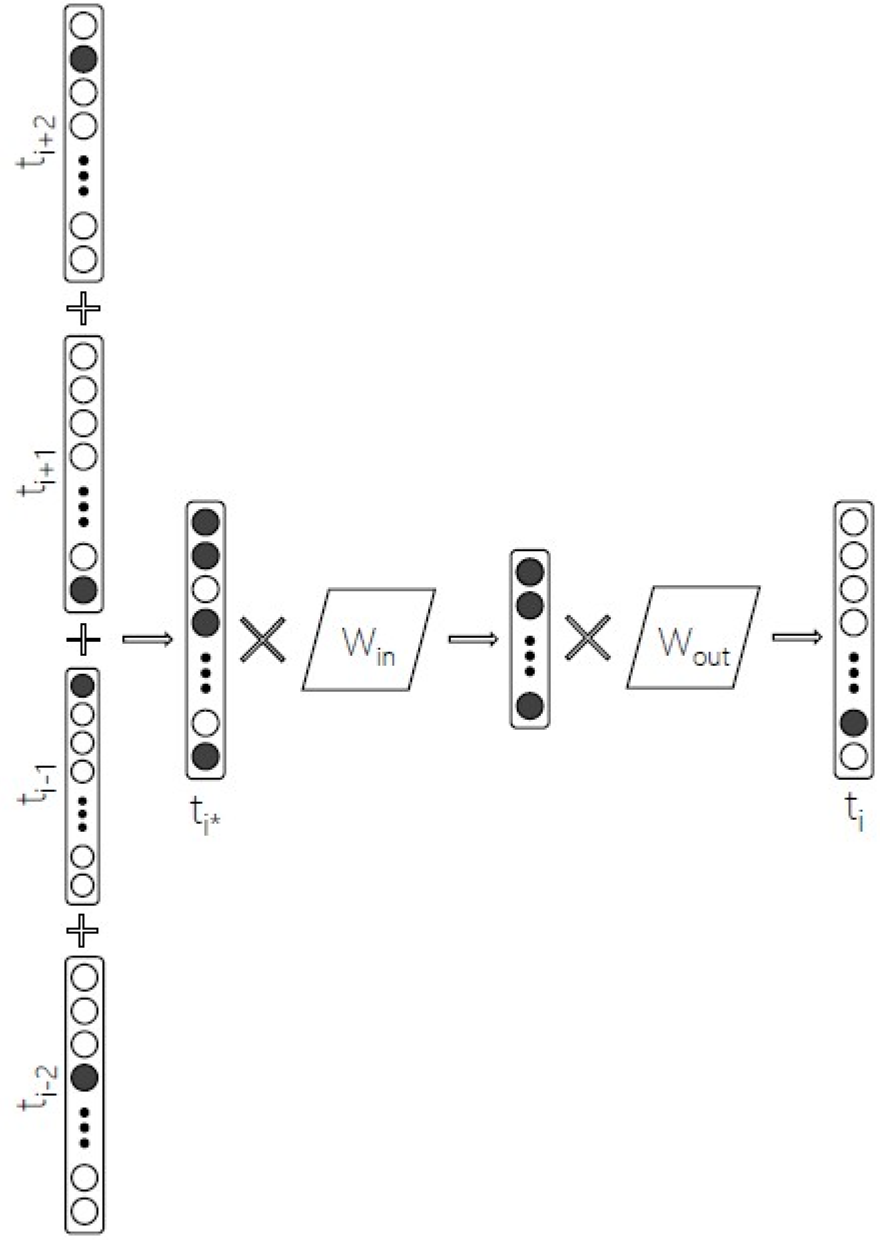
Word vectorization

- Continuous Bag-of-Words (a.k.a. CBOW)

$$\mathcal{L}_{CBOW} = -\frac{1}{|S|} \cdot \sum_{i=1}^{|S|} \log(p(w_i | w_{i-c}, \ldots, w_{i+c}))$$

regularización

$$\mathcal{L} = \mathcal{L}_{CBOW} - \lambda \cdot \sum_{V} ||\vec{w_i}||$$



- GLOVE -

Word vectorization

- **GloVe** (**Glo**bal **Ve**ctors):

context

$$X = \begin{array}{c|c} & \begin{array}{cccccccc} I & like & enjoy & deep & learning & NLP & flying & . \end{array} \\ \begin{array}{c} I \\ like \\ enjoy \\ deep \\ learning \\ NLP \\ flying \\ . \end{array} & \begin{bmatrix} 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & \boxed{0} & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

target

X_{ij}

$X_i = \sum_k X_{ik}$

target

$P_{ij} = P(j|i) = X_{ij}/X_i$ prob. de *j* en el contexto de *i*

context

Word vectorization

- GloVe (Global Vectors):

una baja correlación da un ratio ~ 1

Probability and Ratio	$k = solid$	$k = gas$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	0.96

solid correlaciona con *ice*

gas correlaciona con *steam*

modelo

Corpus

$$F(w_i, w_j, \tilde{w}_k) = \boxed{\frac{P_{ik}}{P_{jk}}}$$

Word vectorization

- Fs que dependen de - : $$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}} .$$

- Lo expresamos vectorialmente: $$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{\boxed{P_{ik}}}{P_{jk}} ,$$

=

- Lo expresamos según F: $$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{\boxed{F(w_i^T \tilde{w}_k)}}{F(w_j^T \tilde{w}_k)} ,$$

F = exp

- Consideramos que: $$w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)$$

bias b_i

- *i* podría intercambiarse por *k*: $$w_k^T \tilde{w}_i = \log(P_{ki}) = \log(X_{ki}) - \log(X_k)$$

bias $\tilde{b}_k$

$$X_{ik} = X_{ki}$$

$$w_i^T w_k = w_k^T w_i$$

$$\longrightarrow \quad w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})$$

Word vectorization

datos

- GloVe (Global Vectors): $$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})$$

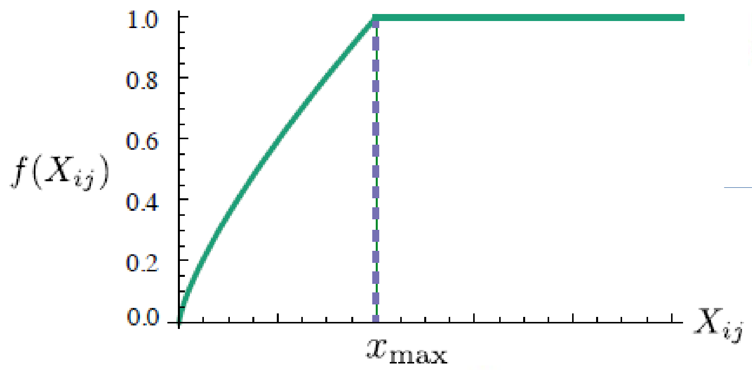
Modelo: Factorización de la matriz de co-ocurrencia

- Tarea: mínimos cuadrados. $$\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

empírico

$$x_{\max} = 100$$
$$\alpha = 3/4$$

- X es sparse. Debemos compensar ese fenómeno:



$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases} .$$

- La función objetivo es: $$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$