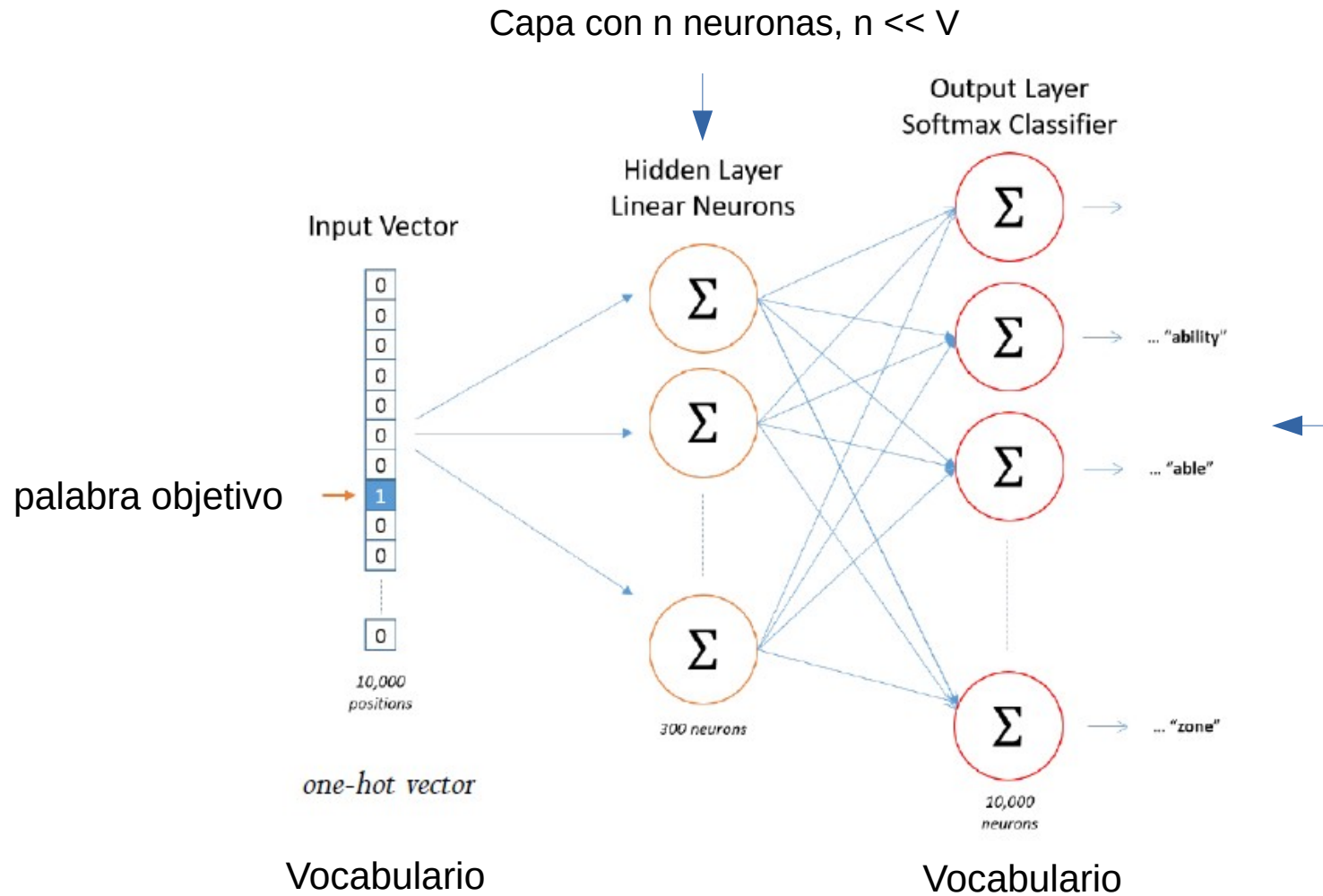


IIC3670 Procesamiento de Lenguaje Natural

<https://github.com/marcelomendoza/IIC3670>

- WORD2VEC -

Word vectorization

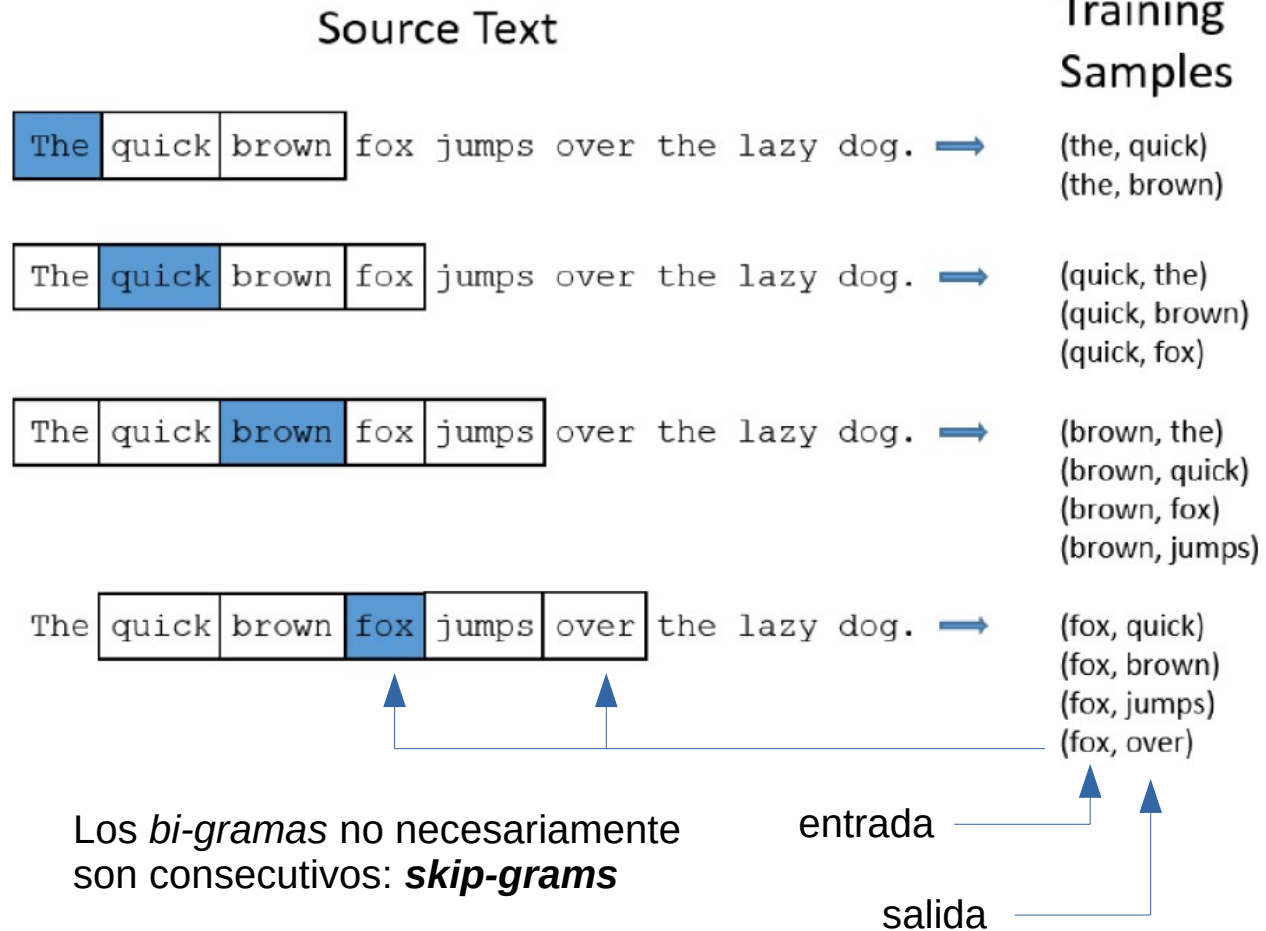


Word vectorization

- Skip-grams (a.k.a. Word2vec): entrenamiento de la red

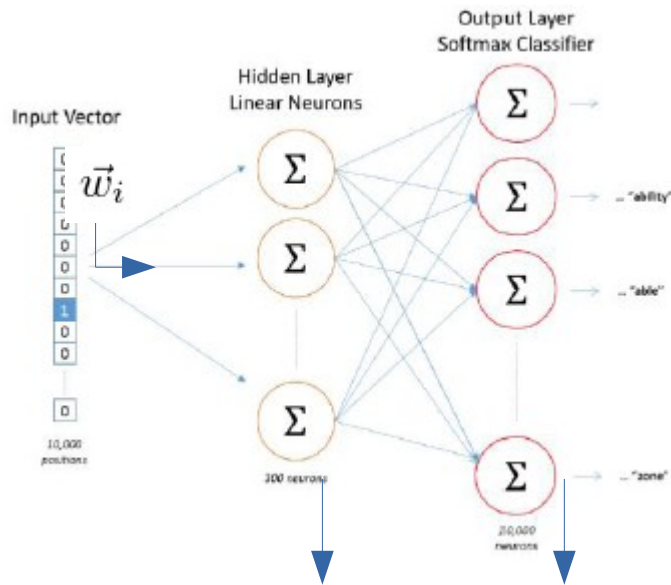
La red se entrena mostrando *bi-gramas*

Una ventana deslizando
recorre el texto →



Word vectorization

- Skip-grams (a.k.a. Word2vec): entrenamiento de la red



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

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

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

Word vectorization

- Skip-grams (a.k.a. Word2vec): Como generar el training set
 - 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:

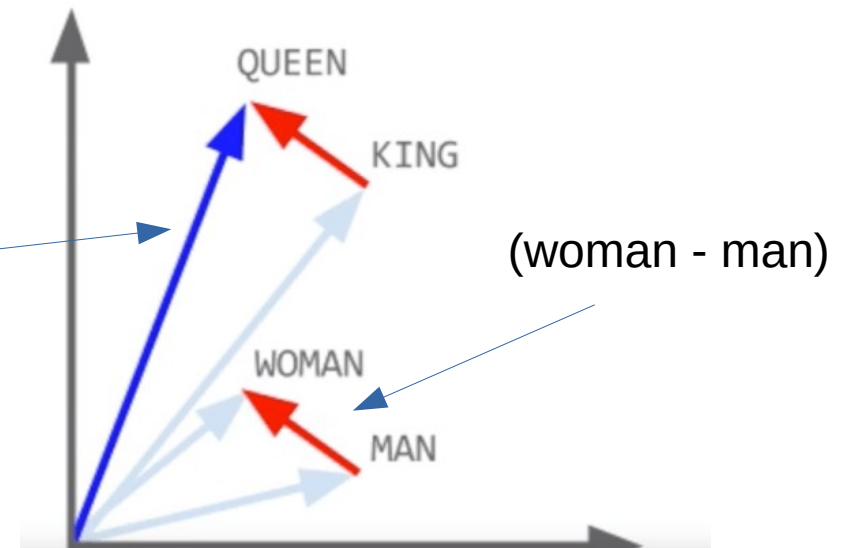
$$P(w_i) = \frac{f(w_i)^\beta}{\sum_{j=0}^n f(w_j)^\beta}$$

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

Word vectorization

- Operadores en word2vec: word analogies

king + (woman - man)



$$\arg \max_{b^* \in V} (\text{sim}(b^*, b - a + a^*))$$

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

Word vectorization

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

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



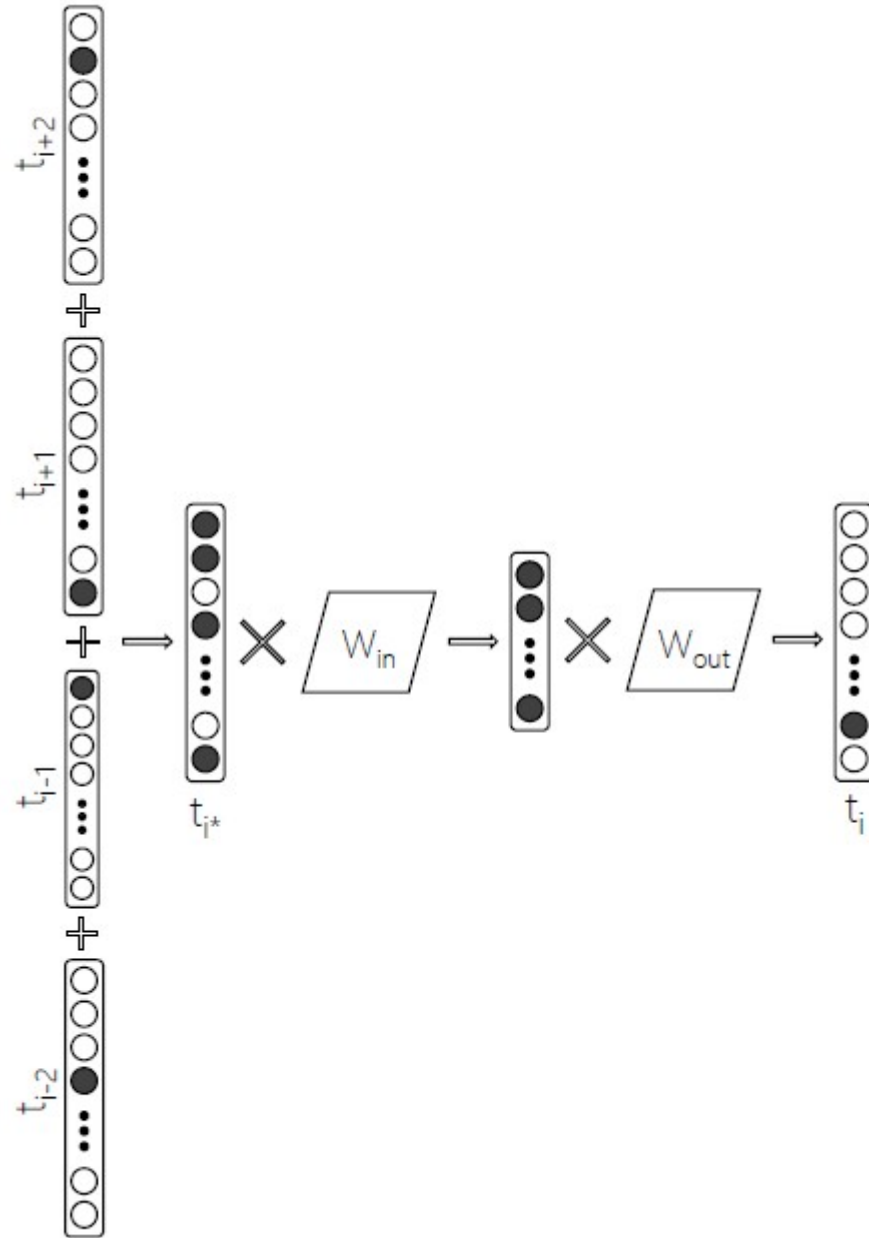
Word vectorization

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

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

↓
regularización

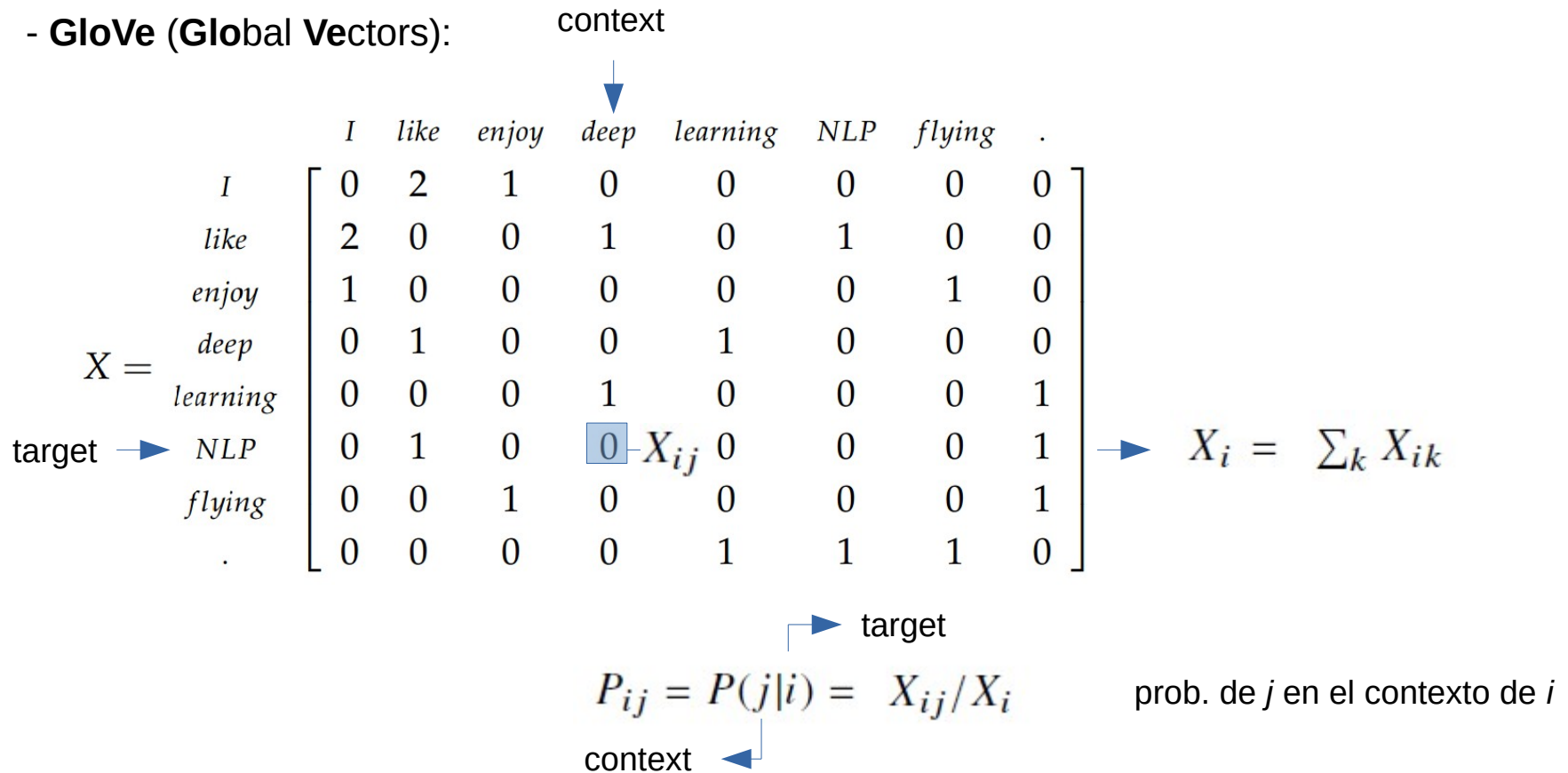
$$\mathcal{L} = \mathcal{L}_{CBOW} - \lambda \cdot \sum_V \|\vec{w}_i\|$$



- GLOVE -

Word vectorization

- GloVe (Global Vectors):



Word vectorization

- GloVe (Global Vectors):

una baja correlación da un ratio ~ 1

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k ice)/P(k steam)$	8.9	8.5×10^{-2}	1.36	0.96

solid correlaciona con *ice*

gas correlaciona con *steam*

modelo

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

Corpus

Word vectorization

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

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

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

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

→ $w_i^T w_k = w_k^T w_i$

Word vectorization

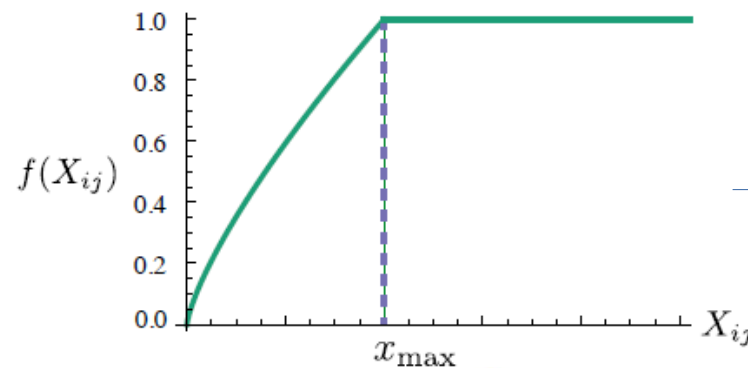
- GloVe (Global Vectors):

$$\underbrace{w_i^T \tilde{w}_k + b_i + \tilde{b}_k}_{\downarrow} = \log(X_{ik}) \uparrow \text{datos}$$

Modelo: Factorización de la matriz de co-ocurrencia

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

- X es sparse. Debemos compensar ese fenómeno:



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

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

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