IIC 3800 Tópicos en CC
NLP

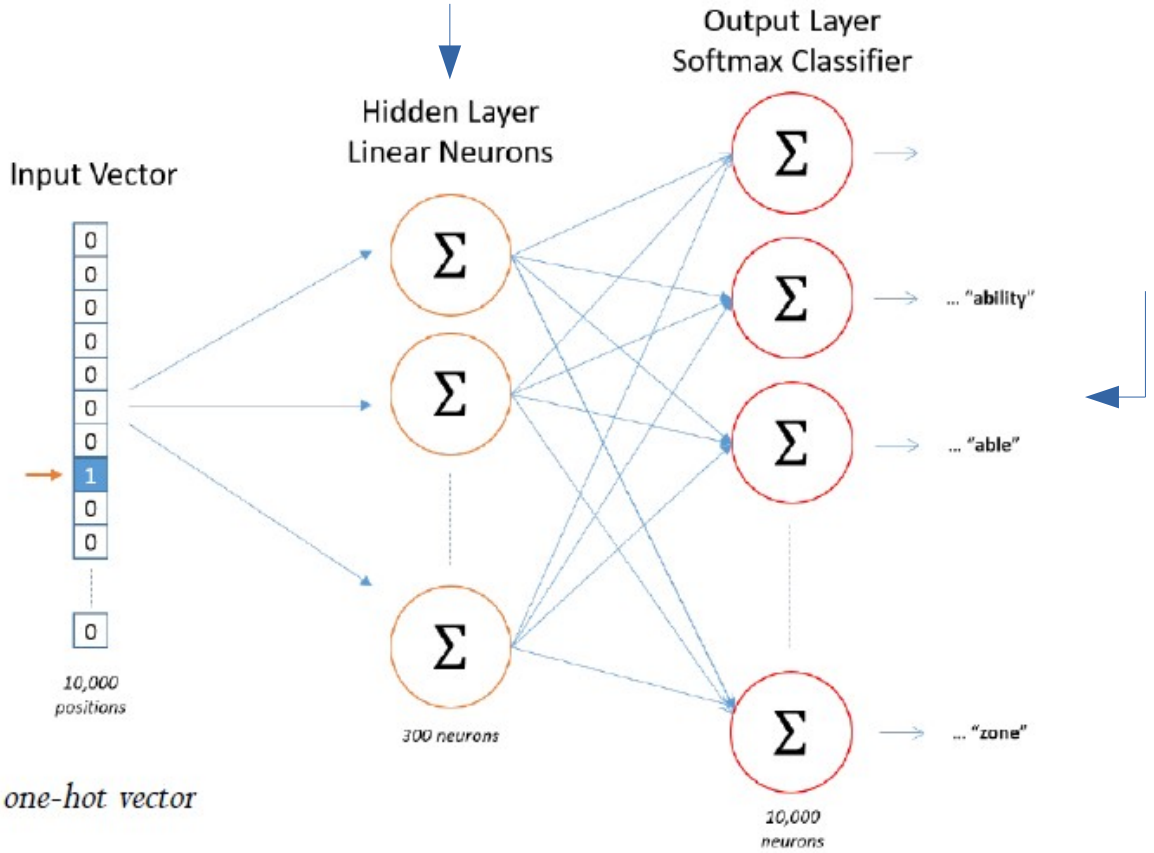
https://github.com/marcelomendoza/IIC3800

- WORD2VEC -

Word vectorization

Capa con n neuronas, n << V

palabra objetivo



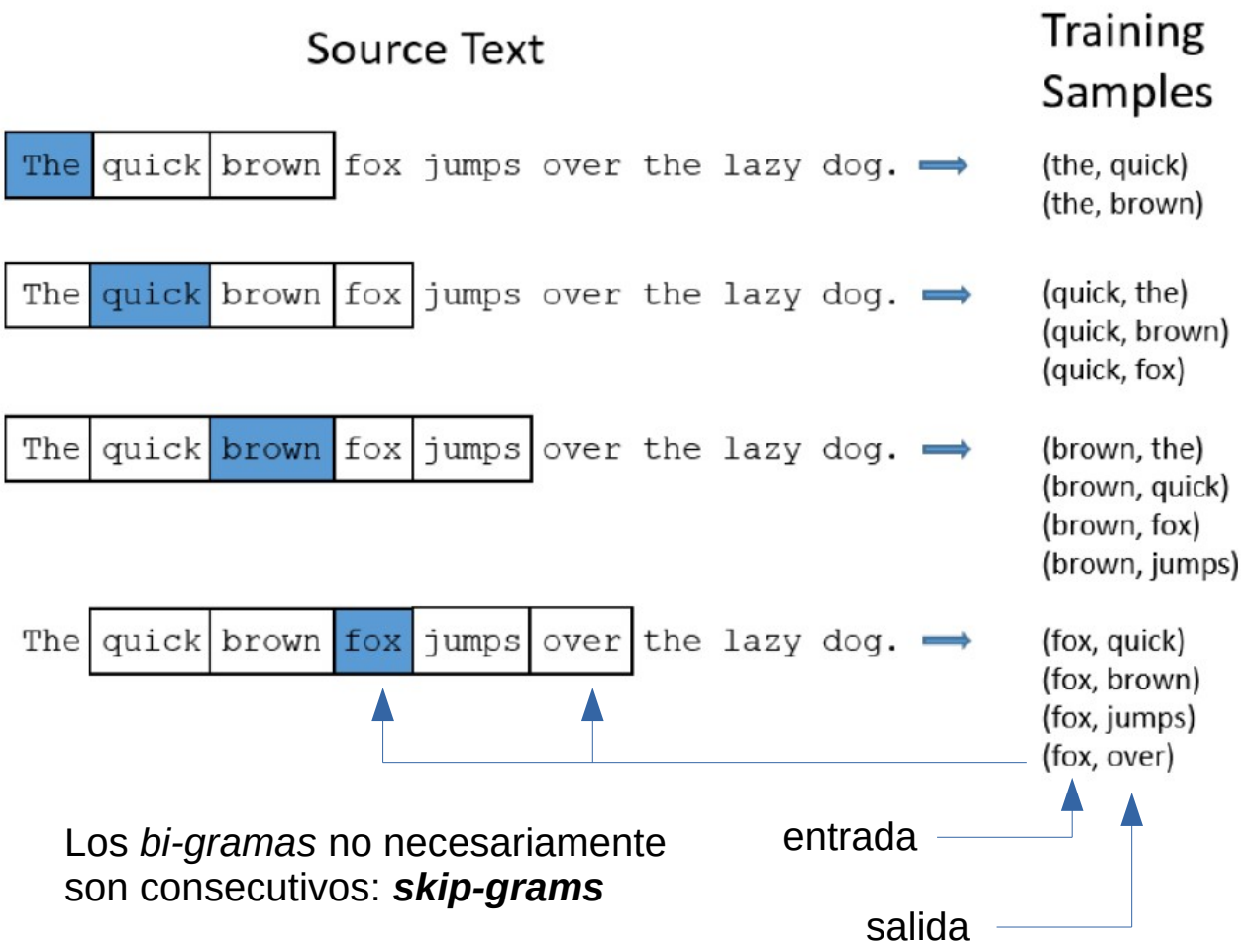
Vocabulario

Vocabulario

Word vectorization

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

La red se entrena mostrando *bi-gramas*



Una ventana deslizante recorre el texto

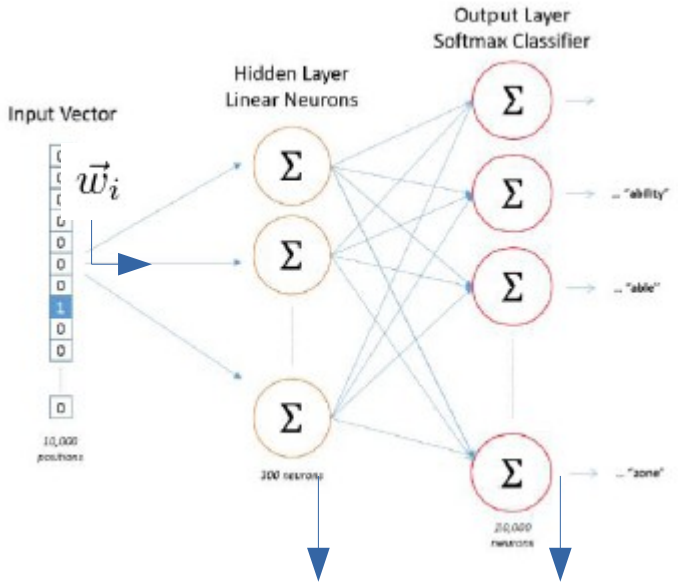
Los *bi-gramas* no necesariamente son consecutivos: ***skip-grams***

entrada

salida

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:

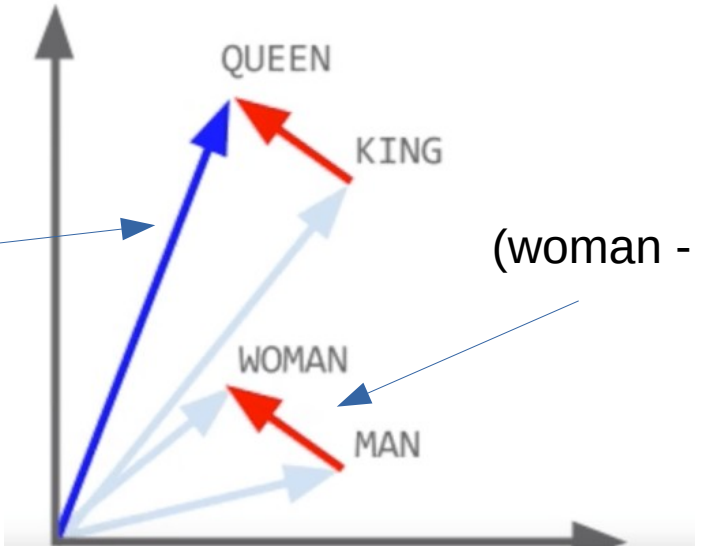
$$P(w_i) = \frac{f(w_i)^\beta}{\sum_{j=0}^{n} f(w_i)^\beta}$$

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

Word vectorization

- Operadores en word2vec: word analogies

king + (woman - man)



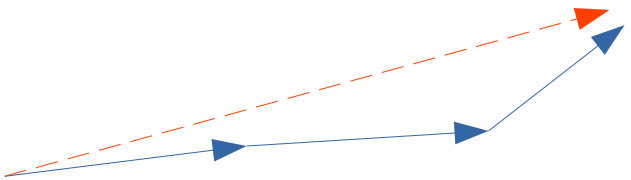
$$\arg\max_{b^* \in V} \left(sim\left(b^*, b - a + a^*\right) \right)$$

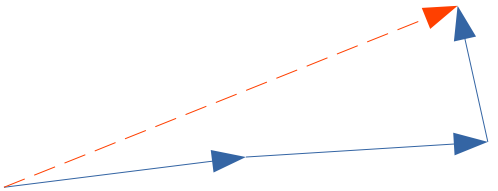
Levy & Goldberg, Linguistic Regularities in Sparse and Explicit Word Representations, ACL'14.

Word vectorization

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

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





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

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