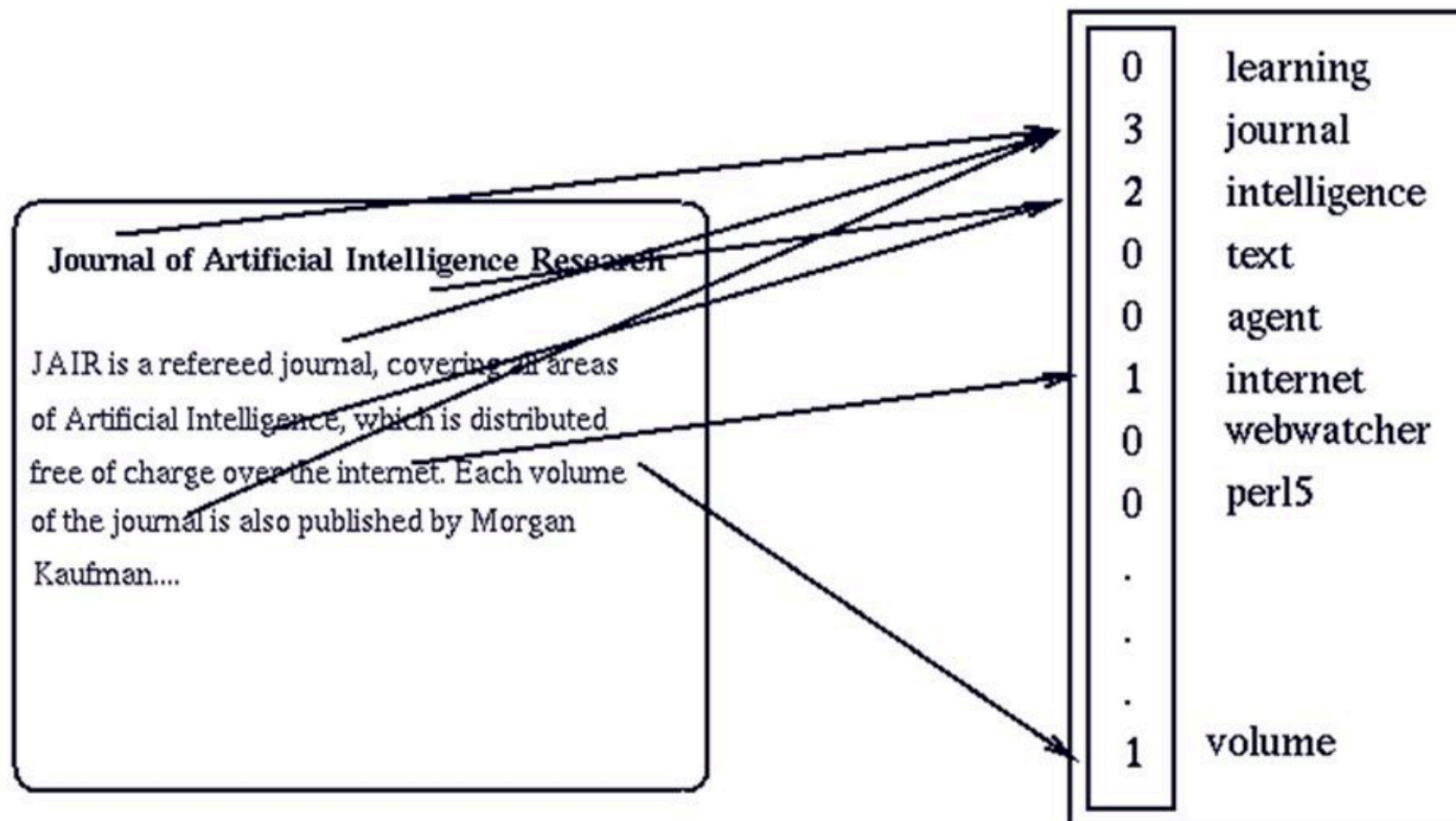


Vector space models

Аксенов Сергей
2019

Векторные представления документов: Bag of Words



Векторные представления документов: TF-IDF

$$\text{tf}(t, d) = \frac{n_t}{\sum_k n_k}$$

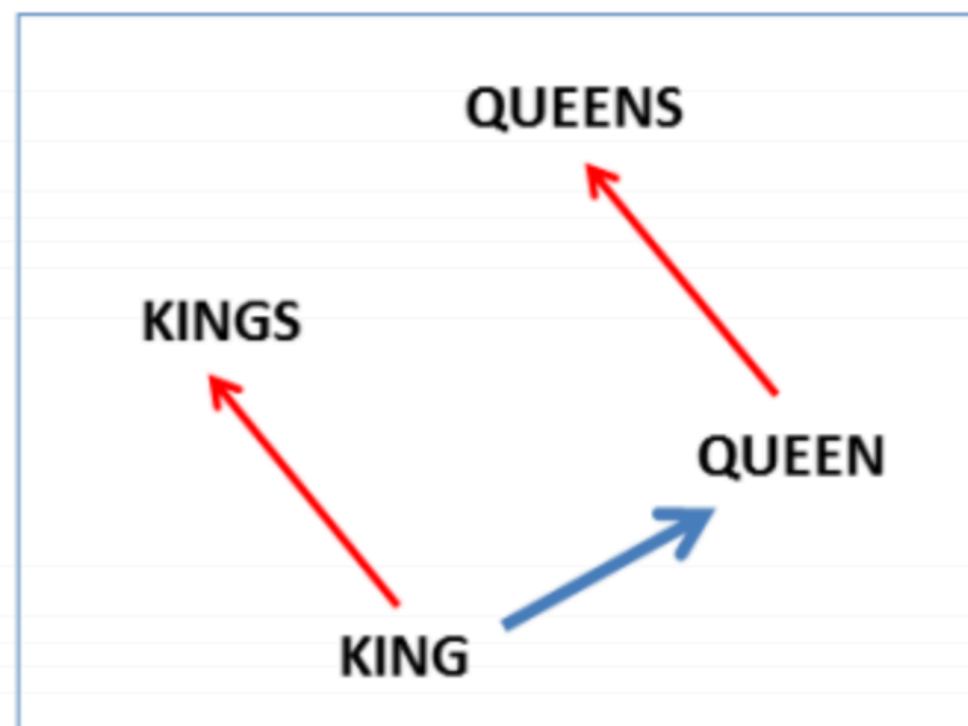
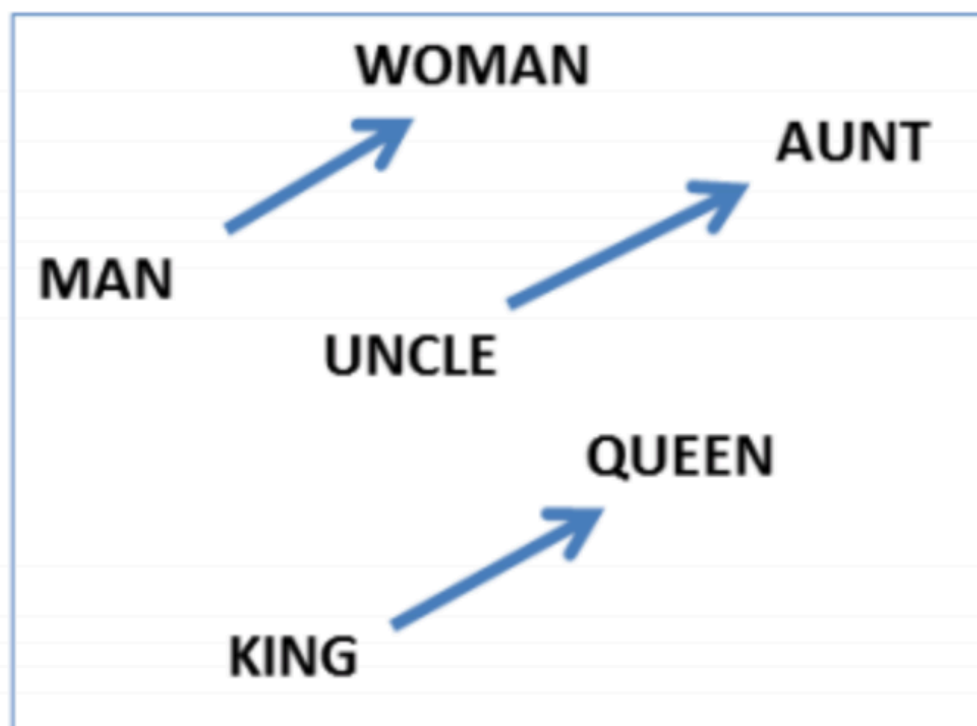
$$\text{idf}(t, D) = \log \frac{|D|}{|\{ d_i \in D \mid t \in d_i \}|}$$

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$

n_i — число слов в документе d_i

$|D|$ — число документов в корпусе D

Векторные представления слов



Векторные представления слов: RPMI

Дано:

- словарь V
- корпус документов D , состоящих из слов $w \in V$
- множество слов-контекстов C , $|C| = n_c$

Пусть f_{ij} — число вхождений слов w_i в документ d_j

F — матрица частот с n_r строками и n_c столбцами.

X — матрица RPMI с n_r строками и n_c столбцами.

$$p_{kl} = \frac{f_{kl}}{\sum_{i=1}^{n_r} \sum_{j=1}^{n_c} f_{ij}}$$

$$p_{k*} = \frac{\sum_{j=1}^{n_c} f_{kj}}{\sum_{i=1}^{n_r} \sum_{j=1}^{n_c} f_{ij}}$$

$$p_{*l} = \frac{\sum_{i=1}^{n_r} f_{il}}{\sum_{i=1}^{n_r} \sum_{j=1}^{n_c} f_{ij}}$$

$$pmi_{ij} = \log \frac{p_{ij}}{p_{i*} p_{*j}}$$

$$x_{ij} = \max(pmi_{ij}, 0)$$

Word2Vec — CBOW

X – one-hot представление входного слова w

w' – выходное слово

$$y_i = p(w' = w_i)$$

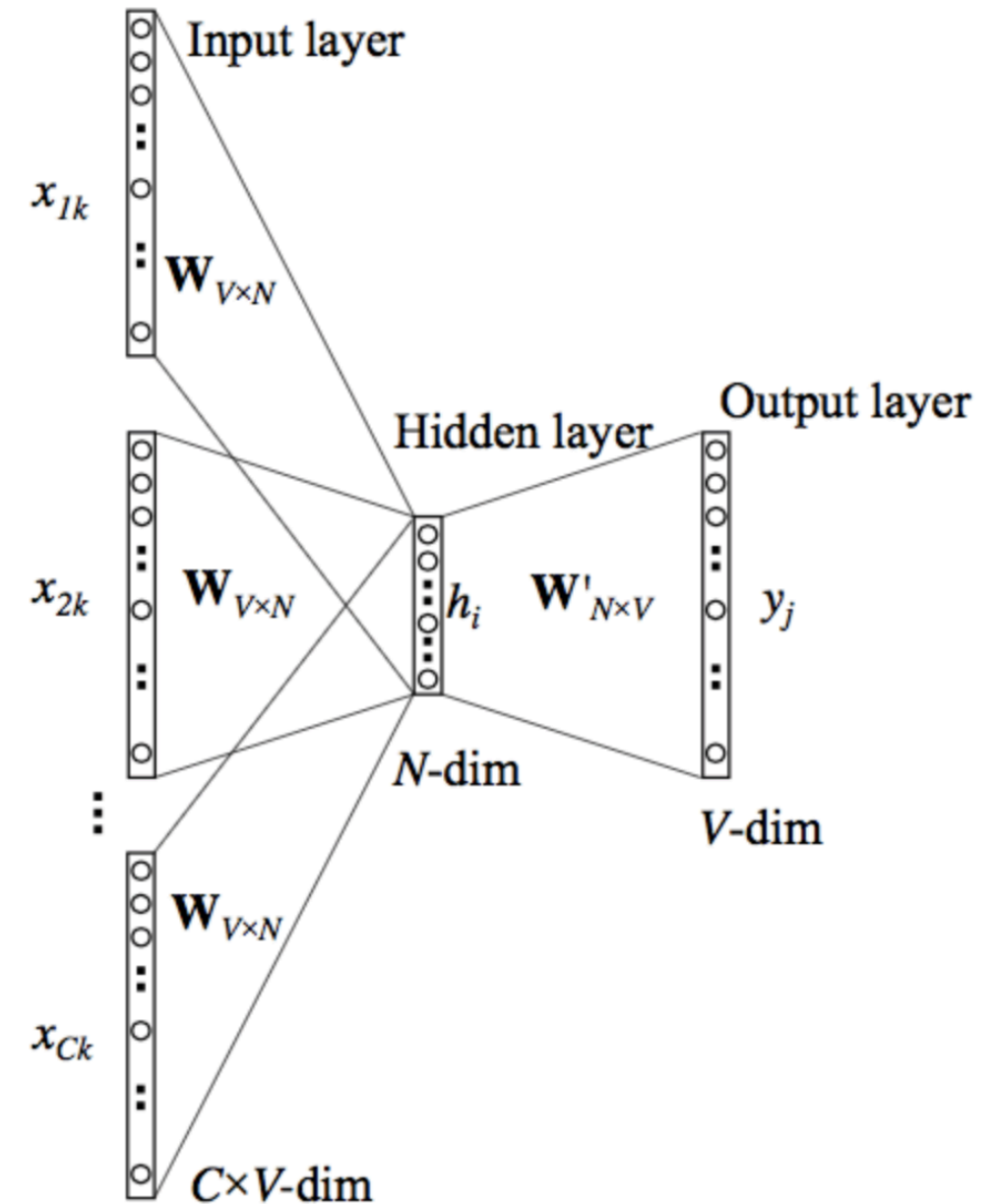
$W = |V \times N|$ – матрица весов между входным и скрытым слоем

$h = x^T W$ – скрытый слой – выбирает одну строку из матрицы W

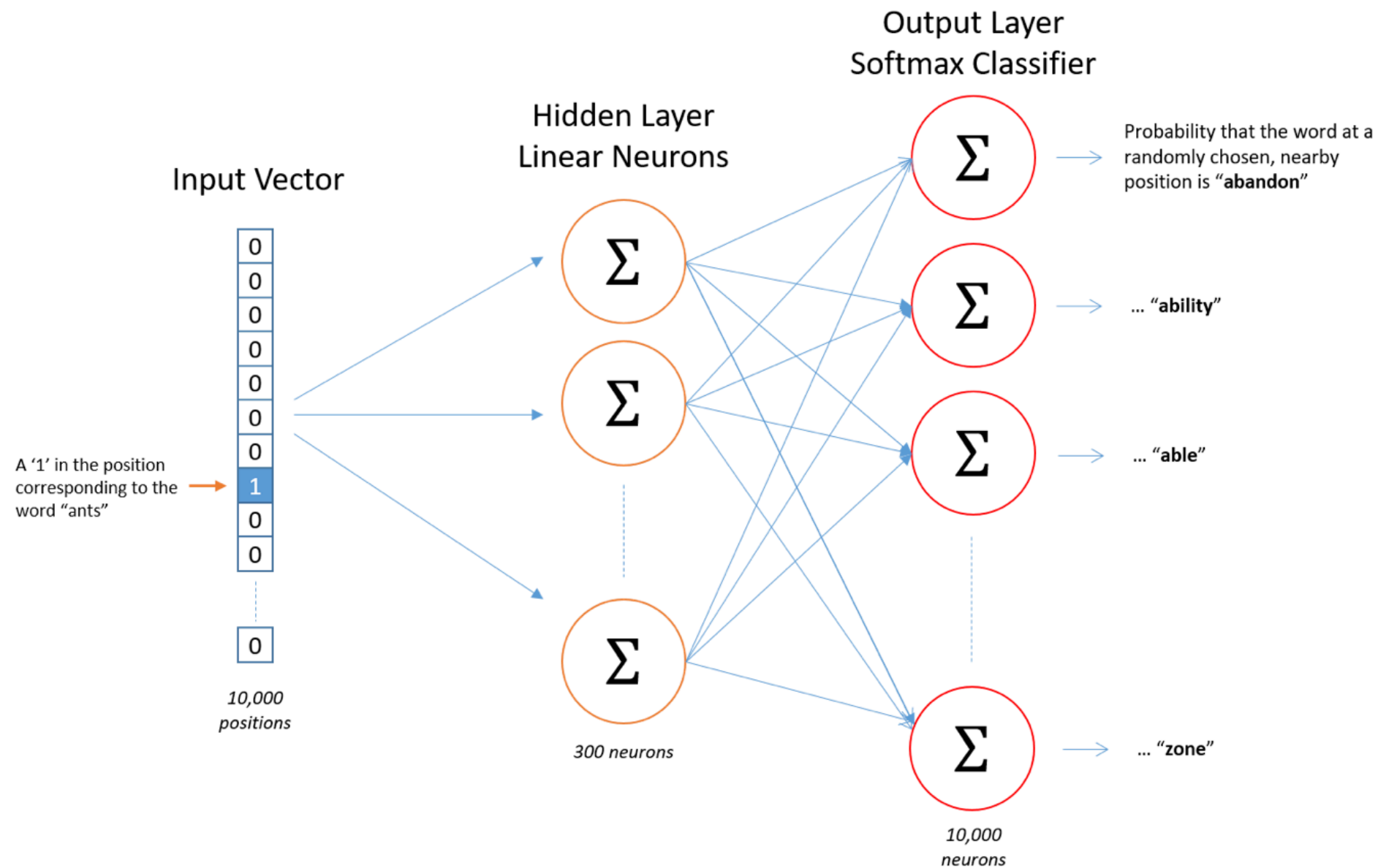
$W' = |N \times V|$ – матрица весов между скрытым слоем и выходным

$u_j = W' h$ – выходной слой

$$p(w_j) = y_j = \frac{\exp(u_j)}{\sum_i \exp(u_i)} \text{ – искомая вероятность}$$



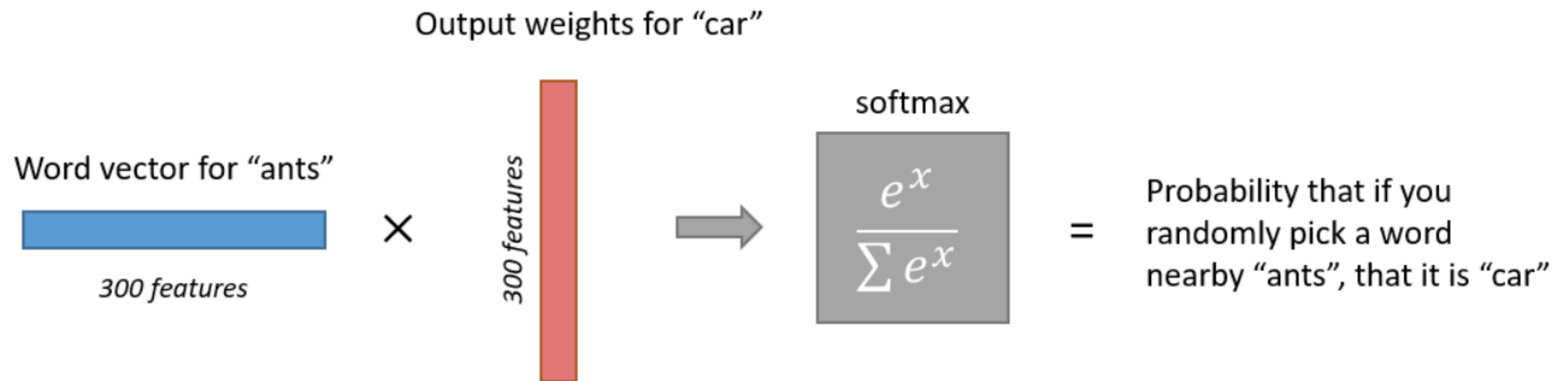
Word2Vec — skip-gram



Word2Vec — skip-gram

| Source Text | Training Samples | | | | | |
|---|------------------|-------|-------|--|------|---|
| <table><tr><td>The</td><td>quick</td><td>brown</td></tr></table> fox jumps over the lazy dog. ➡ | The | quick | brown | (the, quick) (the, brown) | | |
| The | quick | brown | | | | |
| The <table><tr><td>quick</td><td>brown</td><td>fox</td></tr></table> jumps over the lazy dog. ➡ | quick | brown | fox | (quick, the) (quick, brown) (quick, fox) | | |
| quick | brown | fox | | | | |
| The quick <table><tr><td>brown</td><td>fox</td><td>jumps</td></tr></table> over the lazy dog. ➡ | brown | fox | jumps | (brown, the) (brown, quick) (brown, fox) (brown, jumps) | | |
| brown | fox | jumps | | | | |
| The <table><tr><td>quick</td><td>brown</td><td>fox</td><td>jumps</td><td>over</td></tr></table> the lazy dog. ➡ | quick | brown | fox | jumps | over | (fox, quick) (fox, brown) (fox, jumps) (fox, over) |
| quick | brown | fox | jumps | over | | |

Word2Vec — skip-gram



Word2Vec — skip-gram

