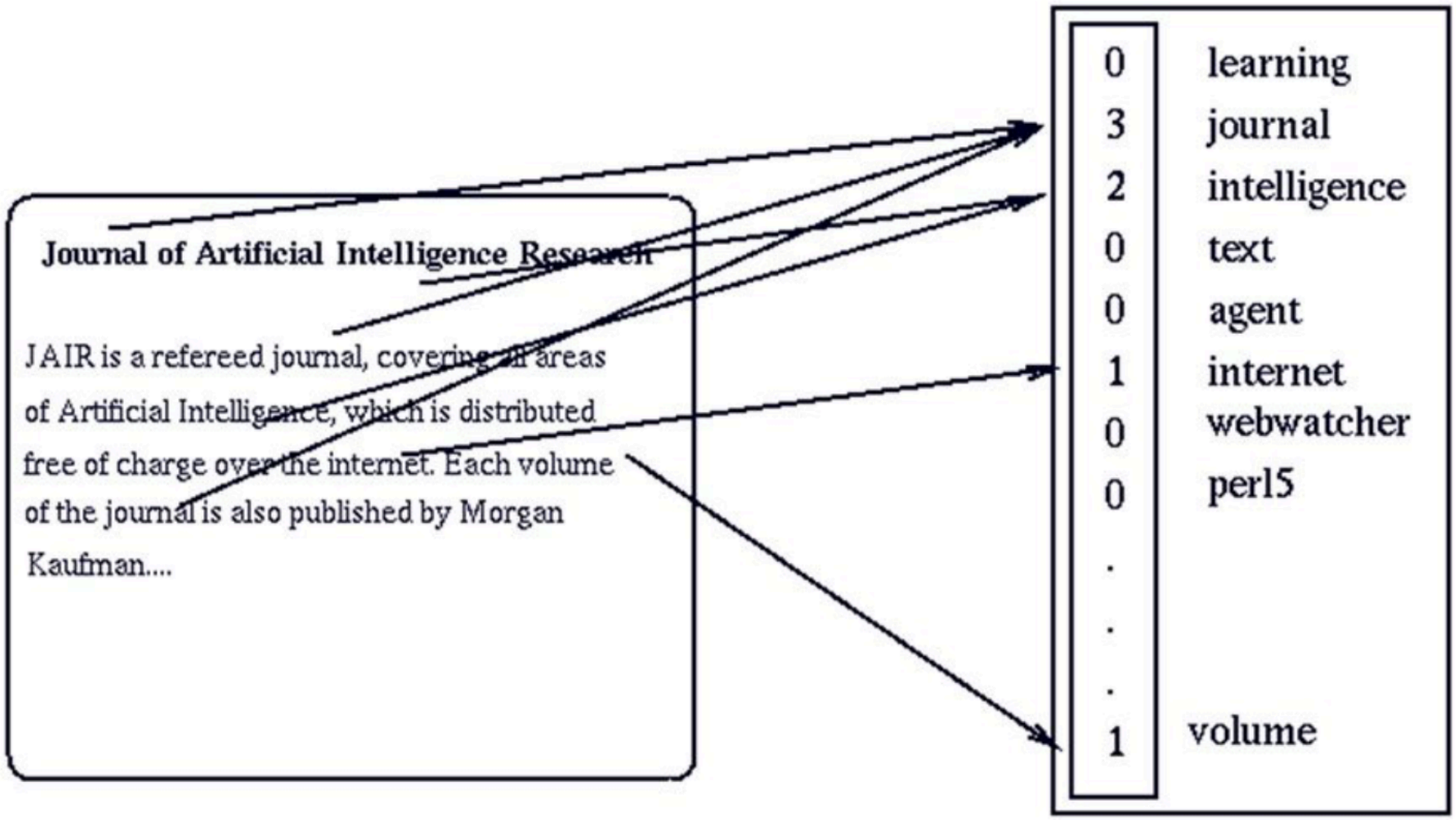
Vector space models

Аксенов Сергей
2019

Векторные представления документов: Bag of Words



Векторные представления документов: TF-IDF

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

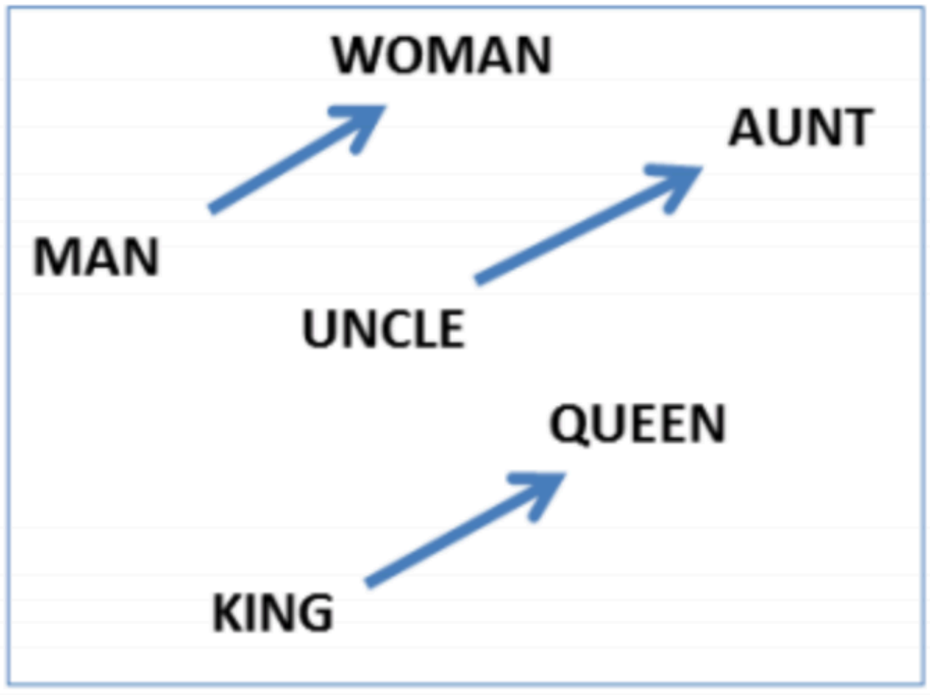
$$\mathrm{idf}(t, D) = \log \frac{|D|}{|\{\, d_i \in D \mid t \in d_i \,\}|}$$

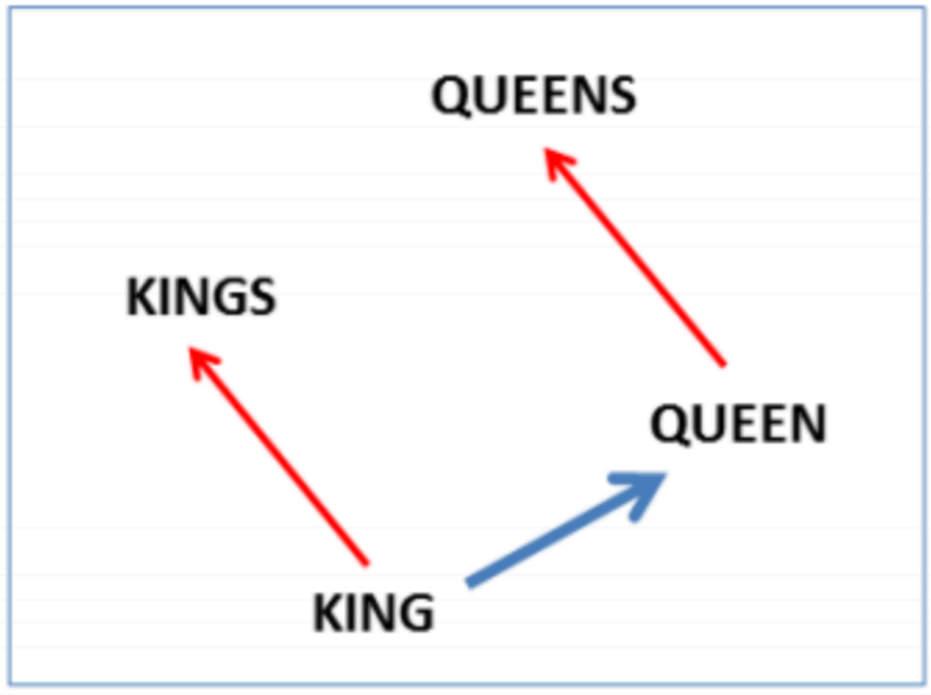
$$\text{tf-idf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

n_i — число слов в документе d_i

$|D|$ — число документов в корпусе D

Векторные представления слов





Векторные представления слов: PPMI

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 PPMI 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

$y_i = p(w' = w_i)$

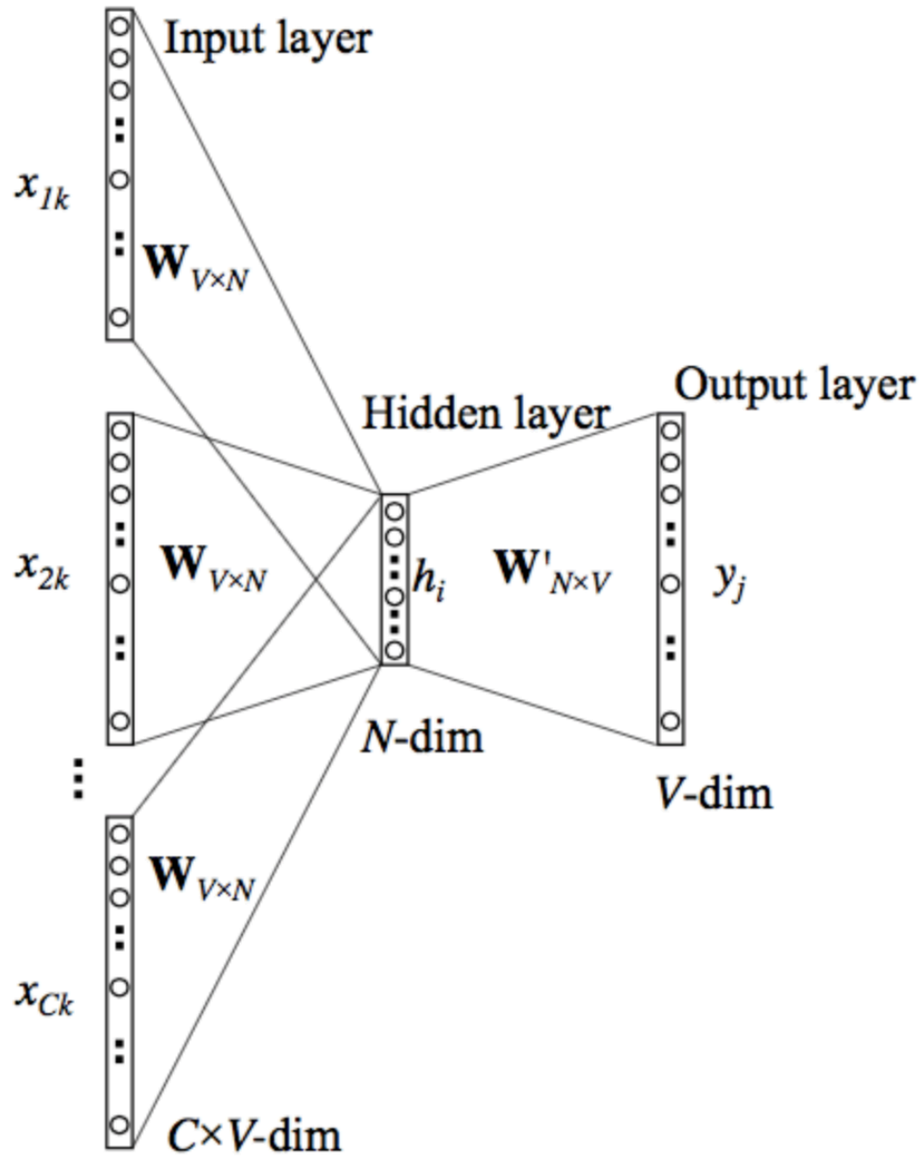
$W = |V \times N|$ – матрица весов между входным и скрытым слоем

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

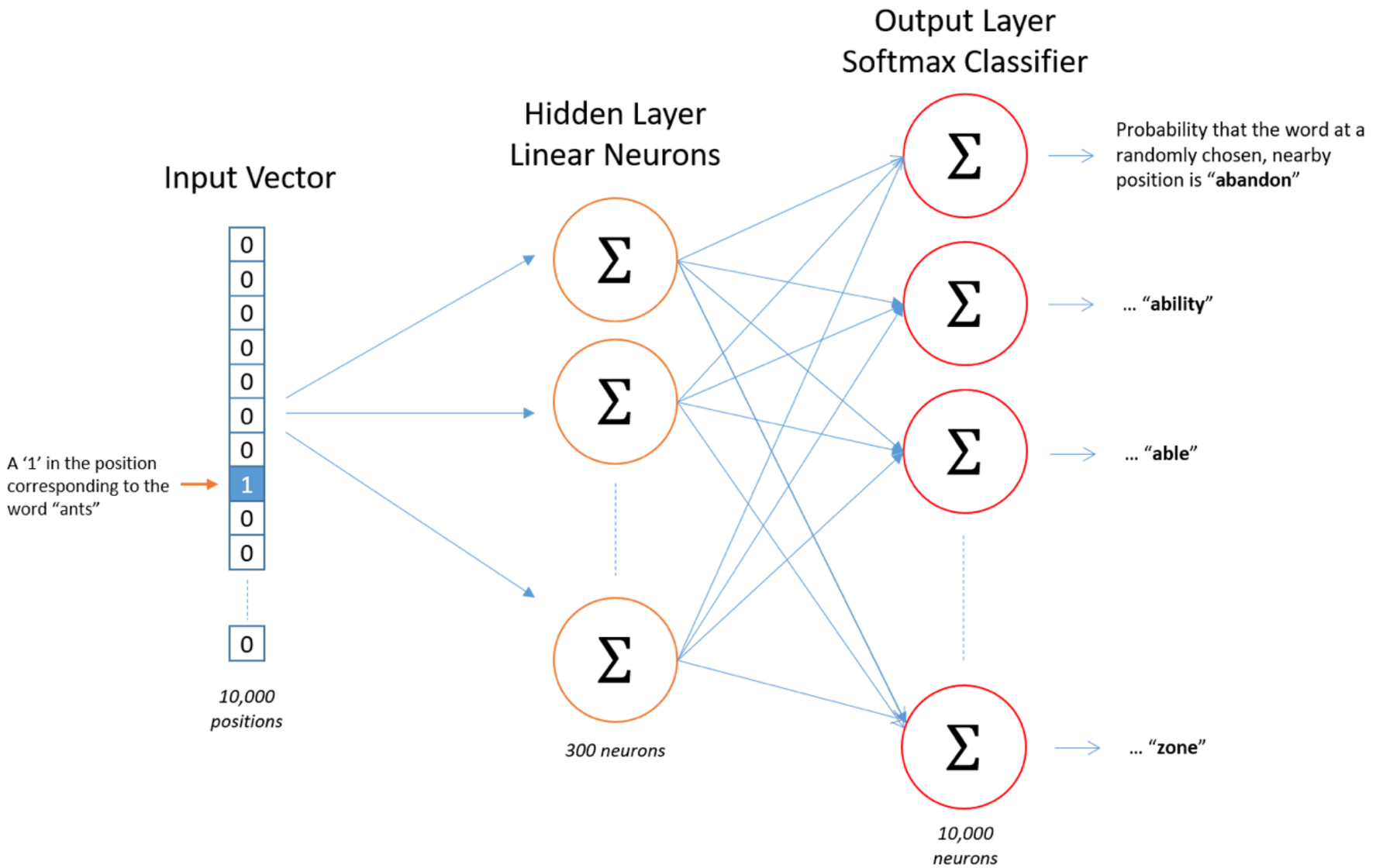
$W' = |N \times V|$ – матрица весов между скрытым слоем и выходным

$u_j = W'h$ – выходной слой

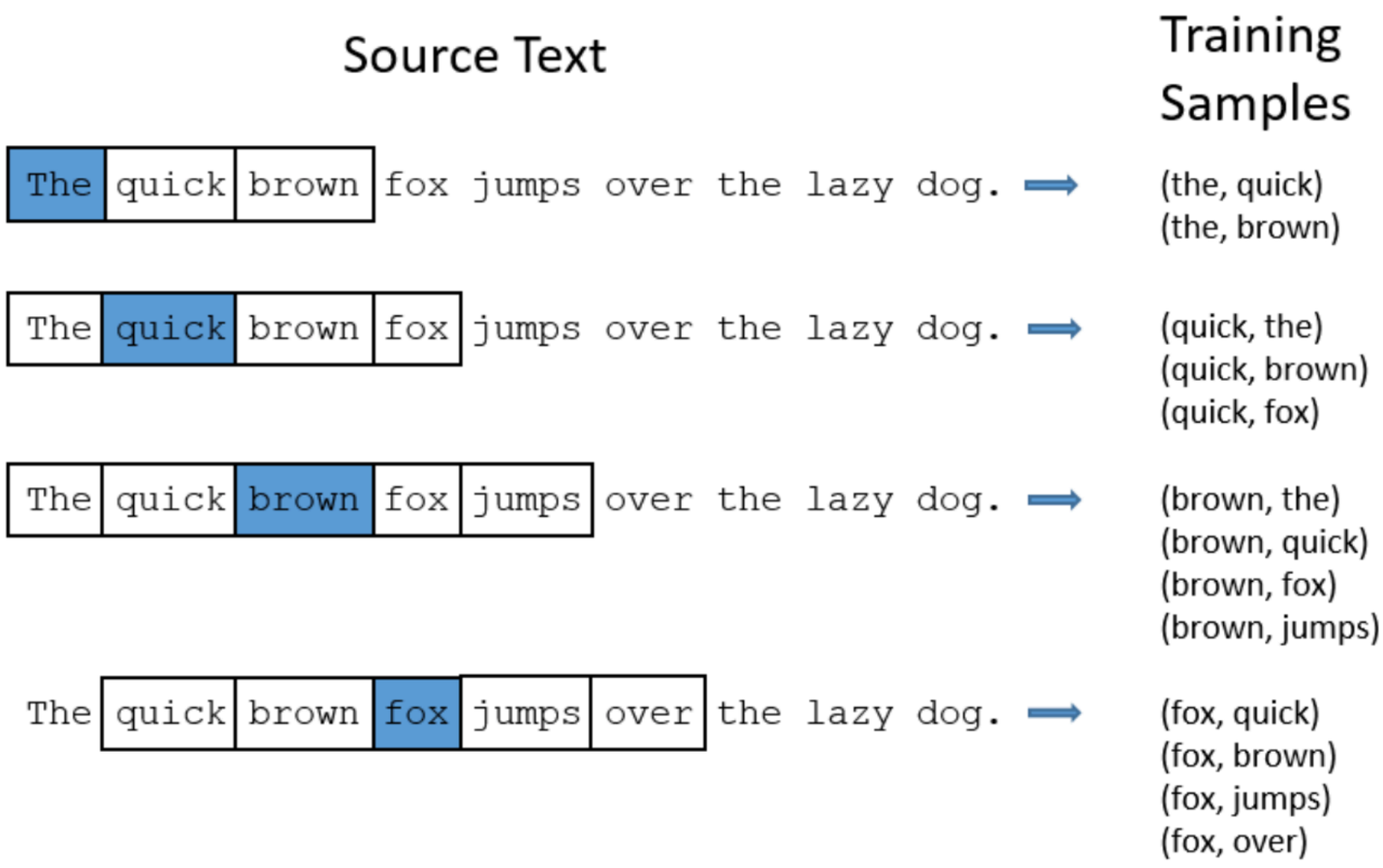
$p(w_j) = y_j = \frac{exp(u_j)}{\sum_i exp(u_i)}$ – искомая вероятность



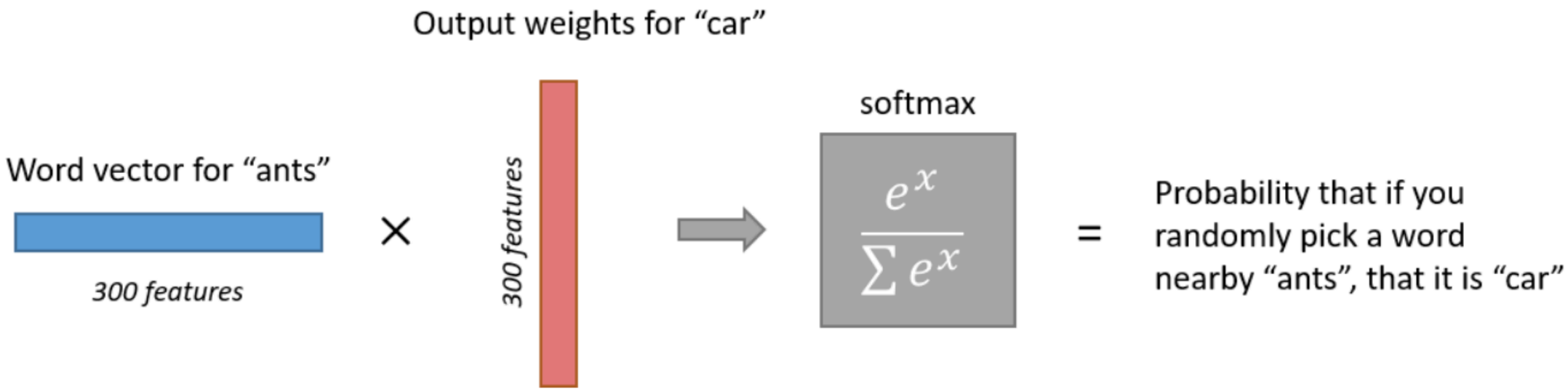
Word2Vec — skip-gram



Word2Vec — skip-gram



Word2Vec — skip-gram



Word2Vec — skip-gram