Word Embeddings

29.10.2020

Plan

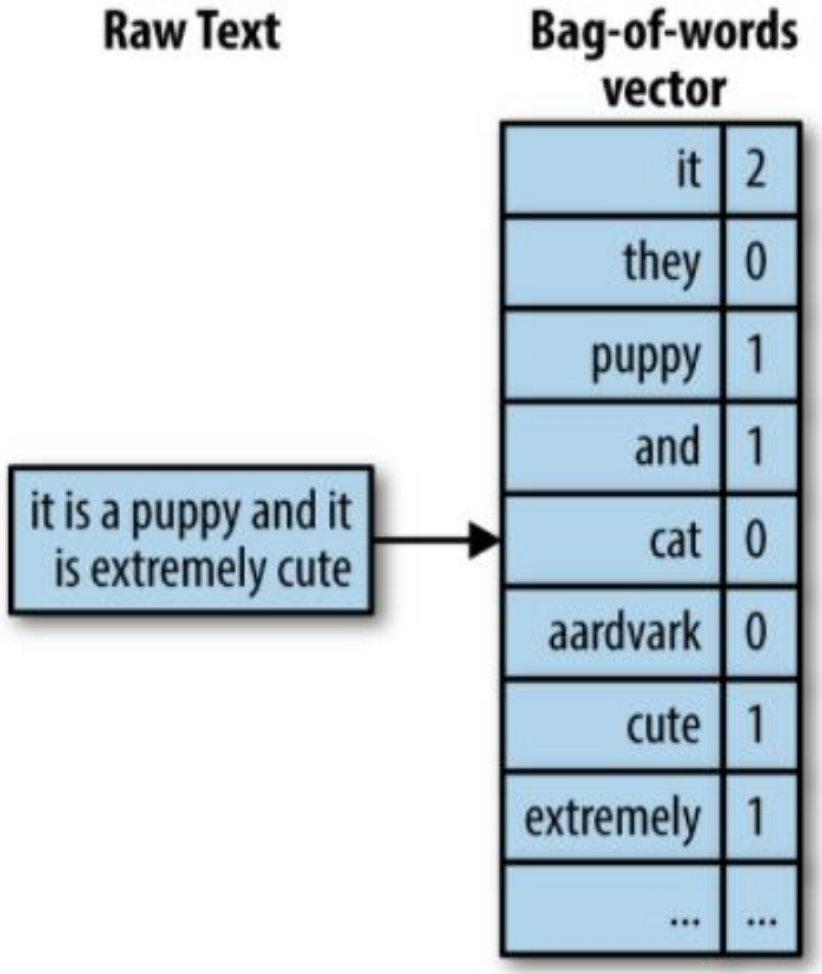
- One-hot encoding (OHE)
- Bag of words (BOW)
- TF-IDF
- Word2vec
- Doc2vec

One-hot encoding (OHE)

id	color
1	red
2	blue
3	green
4	blue

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Bag of words (BOW)



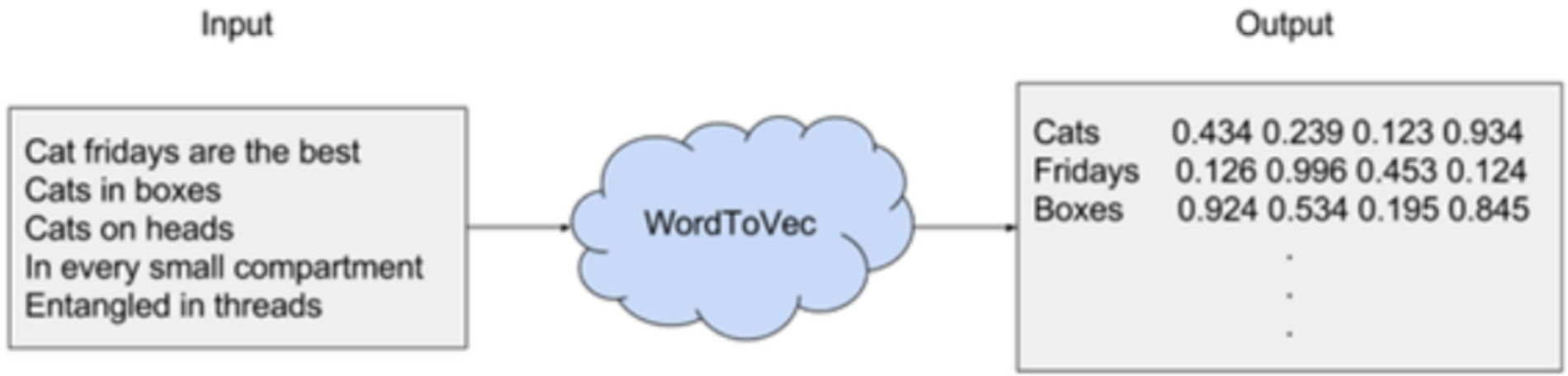
TF-IDF

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

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

Word2vec

Word2vec is a combination of models used to represent distributed representations of words in a corpus.



Word2vec

$$P(w_o|w_c) = \frac{e^{s(w_o,w_c)}}{\sum_{w_i \in V} e^{s(w_i,w_c)}}$$

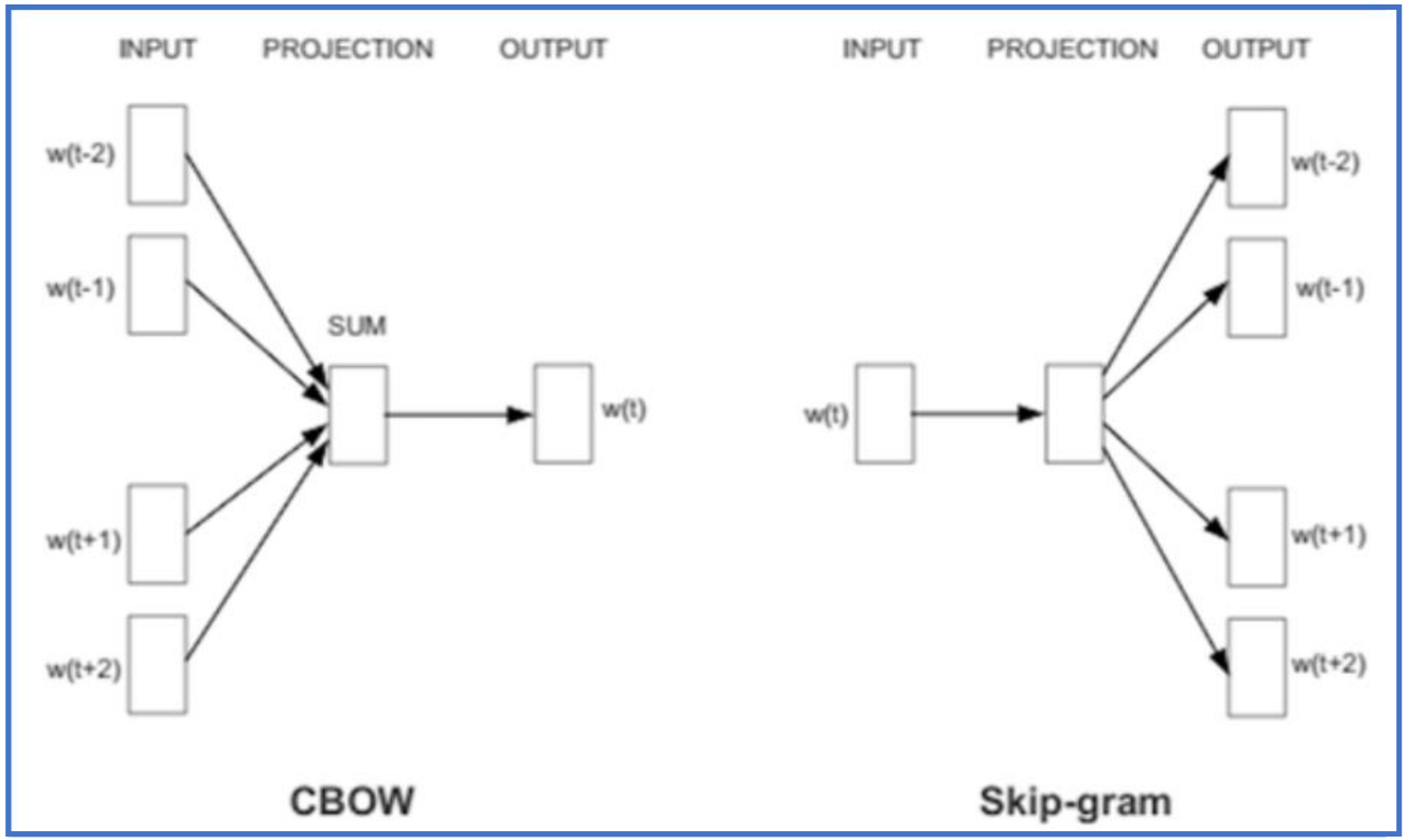
ω_o — вектор целевого слова

ω_c — это некоторый вектор контекста, вычисленный (например, путем усреднения) из векторов окружающих нужное слово других слов.

$s(\omega_o, \omega_c)$— это функция, которая двум векторам сопоставляет одно число. Например, это может быть упоминавшееся выше косинусное расстояние.

Word2vec

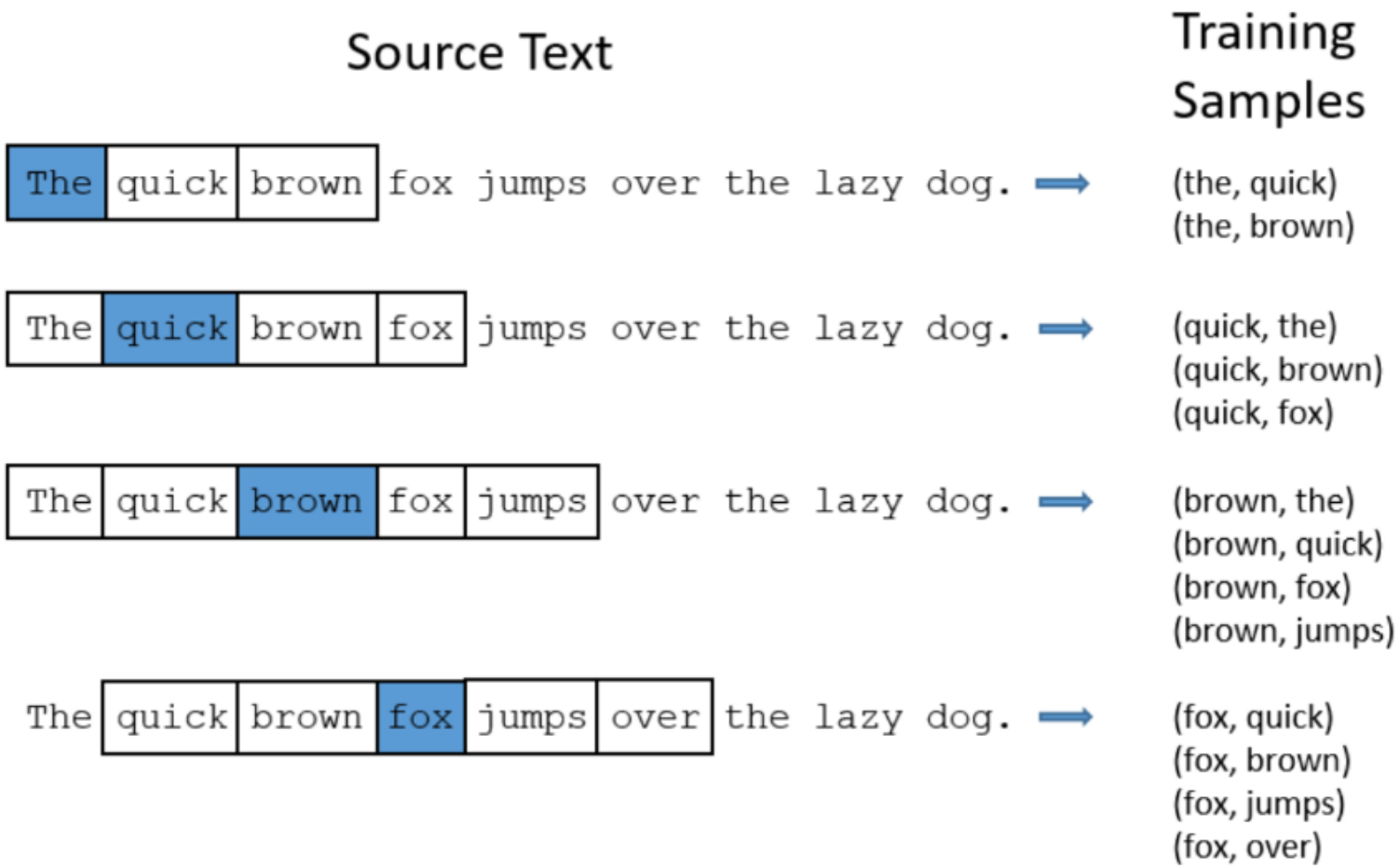
- CBOW
- Skipgram



Word2vec

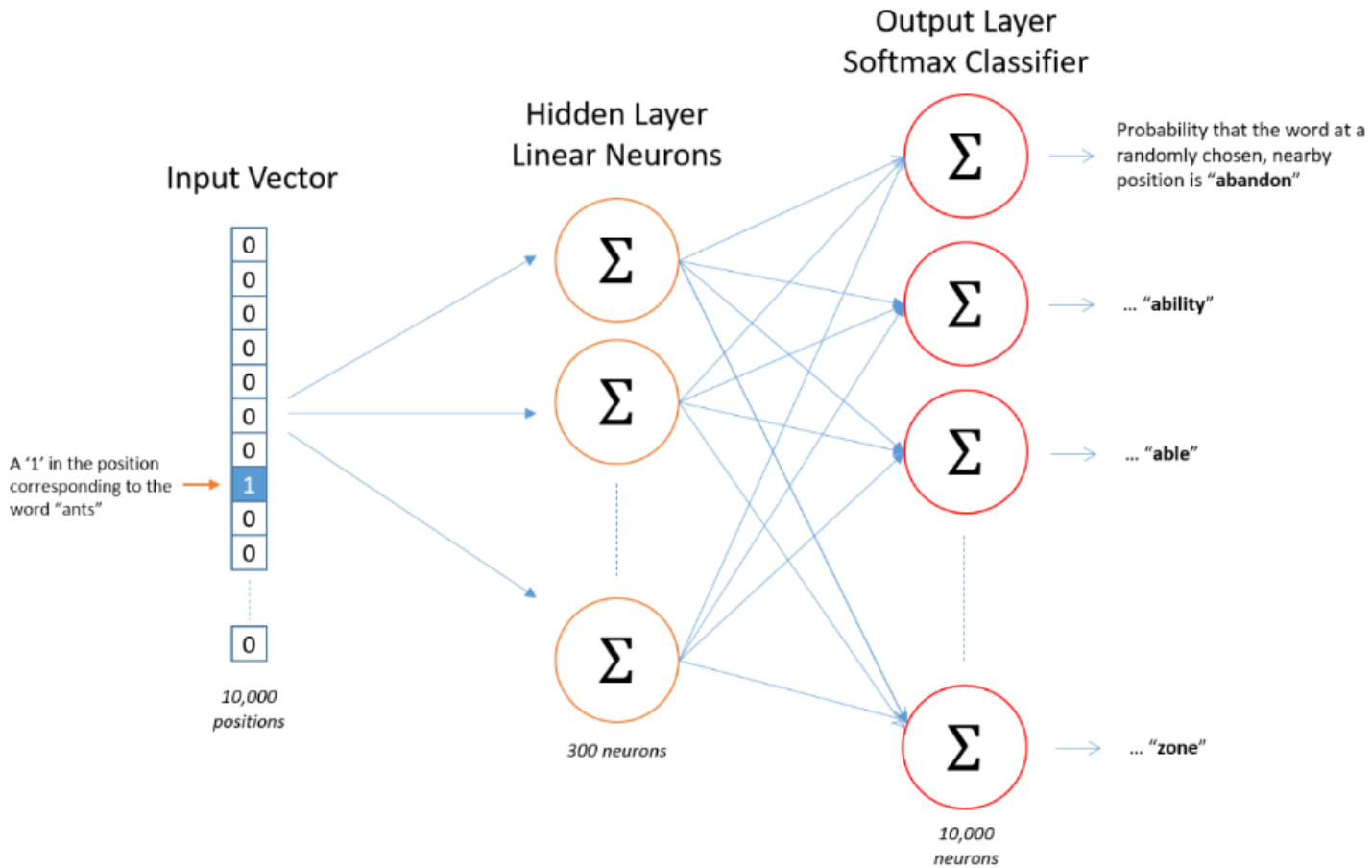
Example:
«The quick brown fox jumps over the lazy dog.»

Word2vec


Source Text
Training Samples
The quick brown fox jumps over the lazy dog.
(the, quick)
(the, brown)
The quick brown fox jumps over the lazy dog.
(quick, the)
(quick, brown)
(quick, fox)
The quick brown fox jumps over the lazy dog.
(brown, the)
(brown, quick)
(brown, fox)
(brown, jumps)
The quick brown fox jumps over the lazy dog.
(fox, quick)
(fox, brown)
(fox, jumps)
(fox, over)

Word2vec

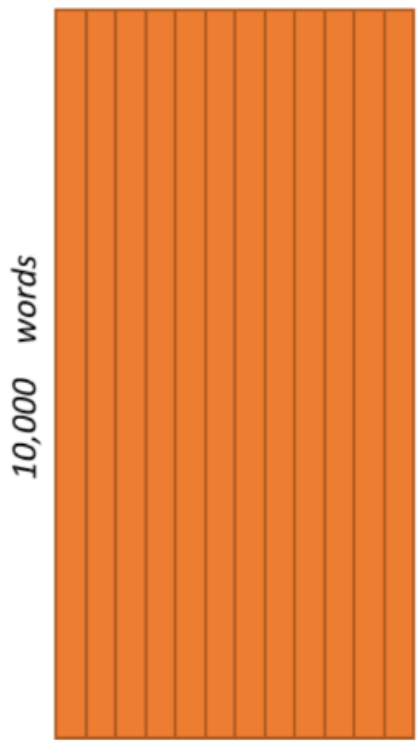
Architecture.



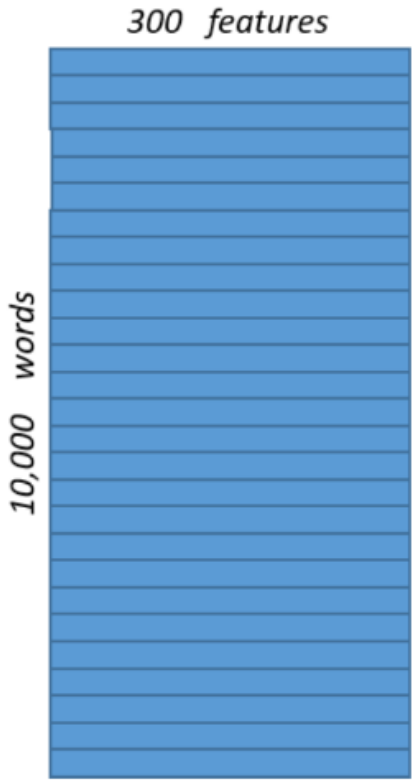
Word2vec

Output.

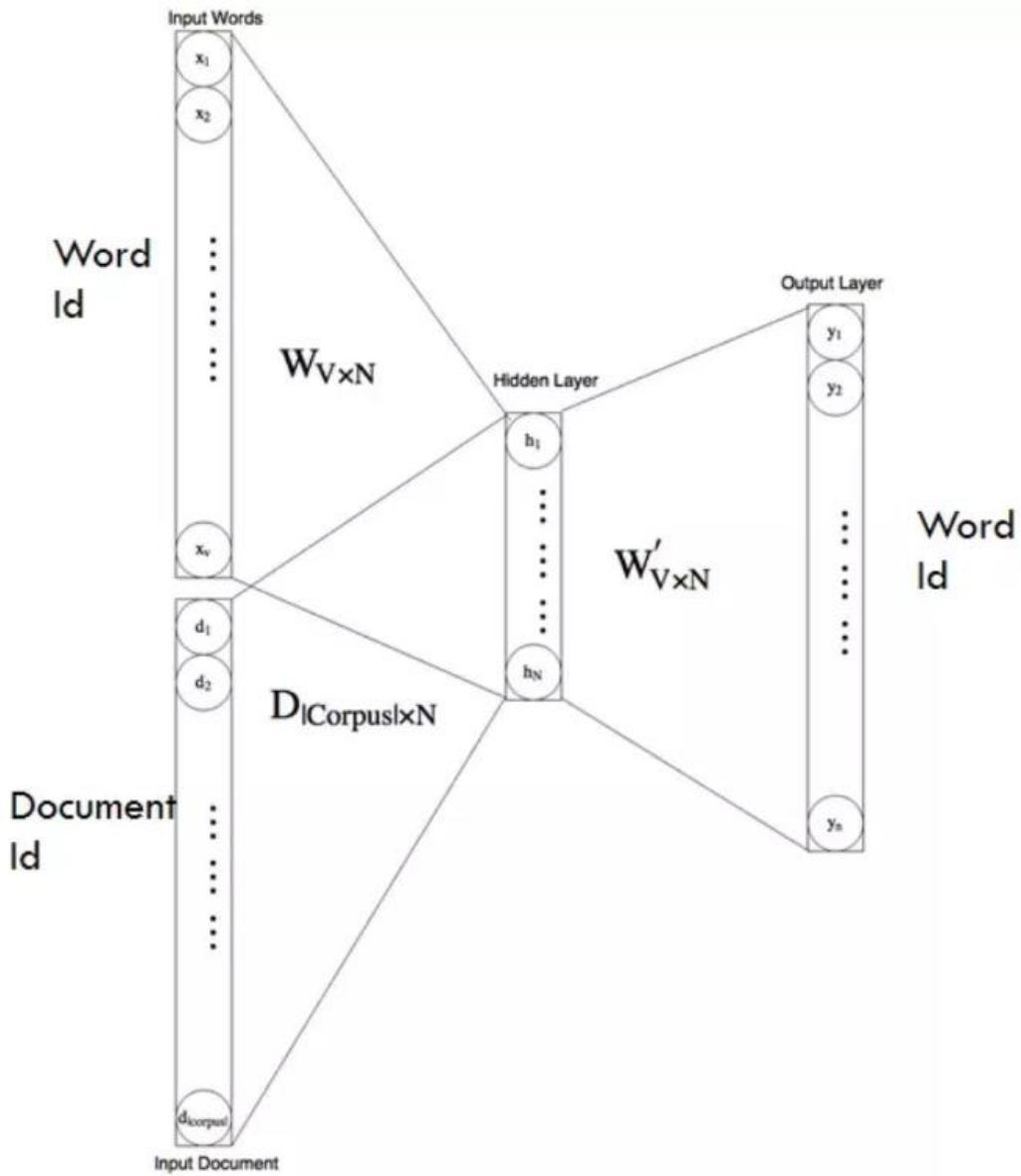
Hidden Layer
Weight Matrix



Word Vector
Lookup Table!



Doc2vec


Input Words
Word Id
$W_{V\times N}$
Hidden Layer
Output Layer
$W'_{V\times N}$
Word Id
Document Id
$D_{|Corpus|\times N}$
Input Document

Links

Word Embeddings
Word2vec
FastText