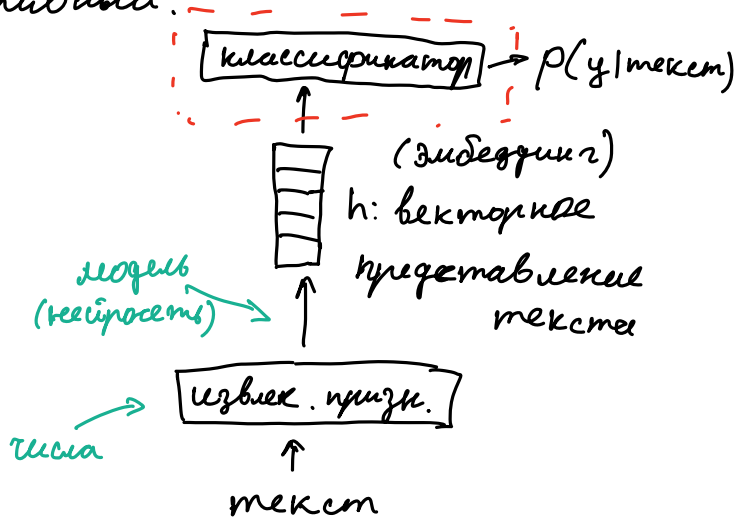
Глубинное обучение в текстах

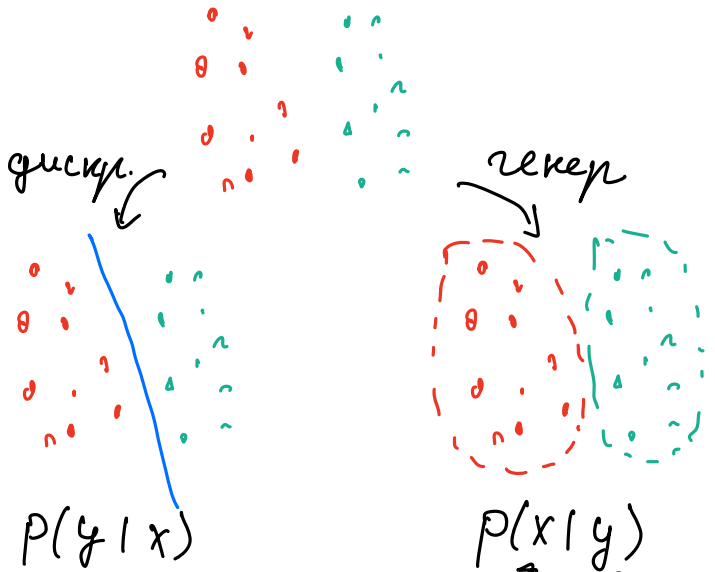
Спам ⇔ не спам?

Позитивный отзыв ⇔ негативный?

Генерация текста.



Дискриминативные и генеративные модели



$$y^* = \arg\max_y p(y|x)$$

$$p(y|x) = \frac{p(x|y)\,p(y)}{p(x)}$$

$$y^* = \arg\max_y p(y|x) = \arg\max_y p(x|y)\,p(y)$$

Наивный Байес

$$y^* = \arg\max_y p(y|x) = \arg\max_y \underbrace{p(x|y)\,p(y)}_{?}$$

$$p(y=k) = \frac{1}{\ell}\sum_{i=1}^{\ell}[y_i = k]$$

$$p(x \mid y=k) = p(x_1, \dots, x_n \mid y=k) \approx \prod_{i=1}^{n} p(x_i \mid y=k)$$

предположение

$$p(x_i \mid y=k) = \frac{N(x_i, y=k)}{\sum_{j=1}^{|V|} N(x_j, y=k)}$$

сколько раз x_i был в тексте класса k

размер выборки

длина текста

$$N(x_i, y=k) = \sum_{j=1}^{\ell} \sum_{s=1}^{n} [x_{js} = x_i][y_j = k]$$

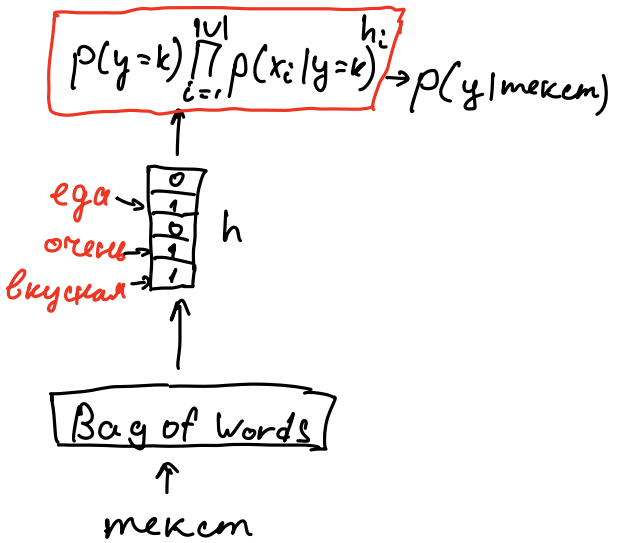
Почему это работает?

x: Еда | очень | вкусная

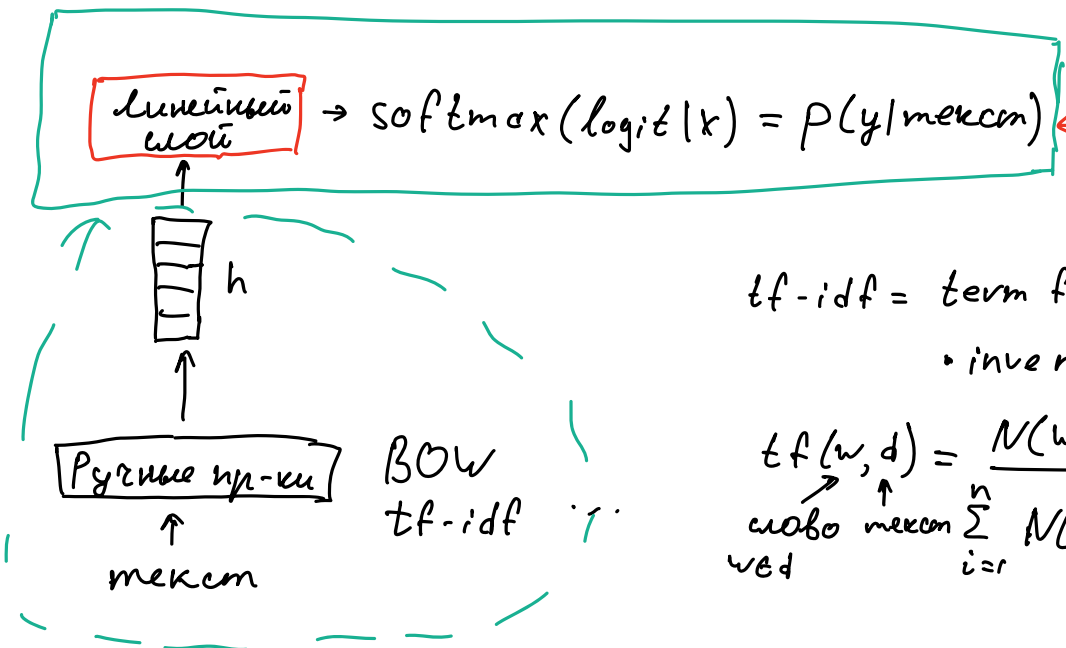
$p(x_1 \mid y=+)$ $p(x_1 \mid y=-)$ $p(x_3 \mid y=+) > p(x_3 \mid y=-)$

$$p(y \mid x) \propto \prod_{i=1}^{n} p(x_i \mid y=k) \cdot p(y)^{\frac{1}{n}}$$

$$BOW(\underbrace{x_1, \dots, x_n}_{x}) = \{N(w_j, x)\}_{j=1}^{|V|}$$



Логистическая регрессия



tf-idf = term frequency (tf) · inverse document frequency (idf)

$$tf(w, d) = \frac{N(w, d)}{\sum_{i=1}^{n} N(w_i, d)}$$

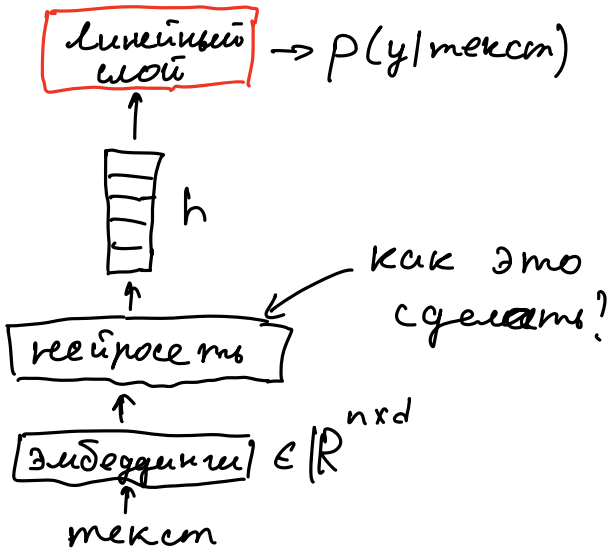
слово w, текст d

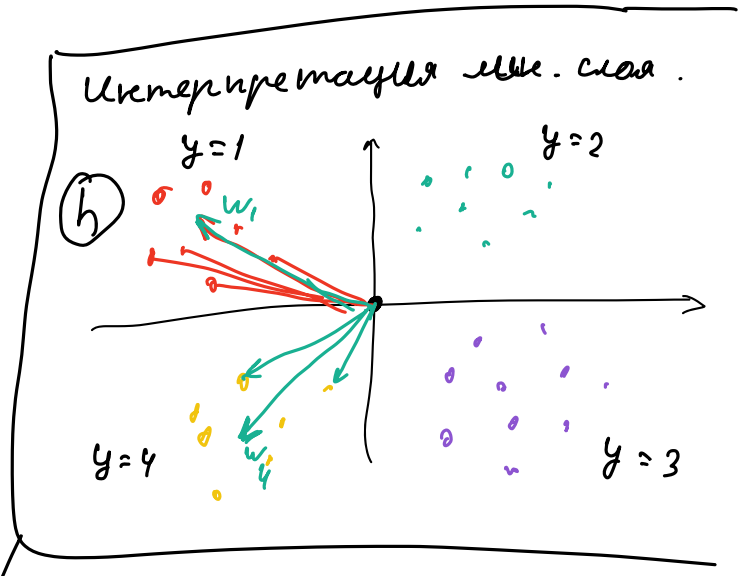
$$idf(w, d, D) = \log \frac{|D|}{|\{d \in D : w \in d\}|}$$

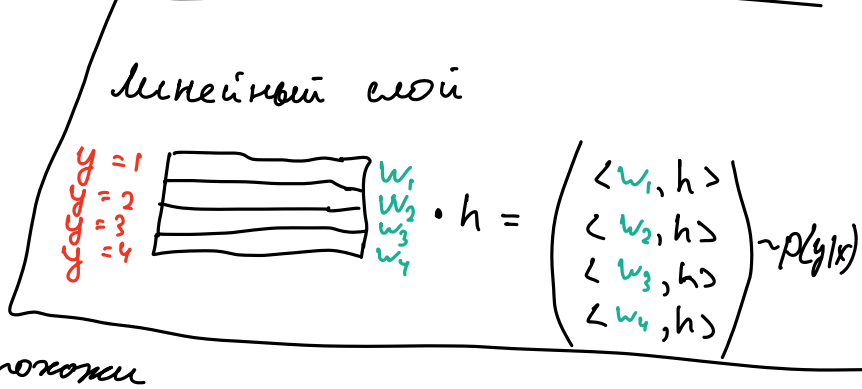
(D — датасет)

$$tf\text{-}idf(\underbrace{x_1, \dots, x_n}_{x}) = \{tf(w_j, x) \cdot idf(w_j, x, D)\}_{j=1}^{|V|}$$

Нейросетевые методы







Эмбеддинги

Сопоставляют каждому слову свой вектор.

еда ↔ $\mathbb{R}^d$

эмбеддинги похожих слов похожи

$$\rho(\text{еда}, \text{хлеб}) = \|\text{еда} - \text{хлеб}\|_2^2 \to 0$$

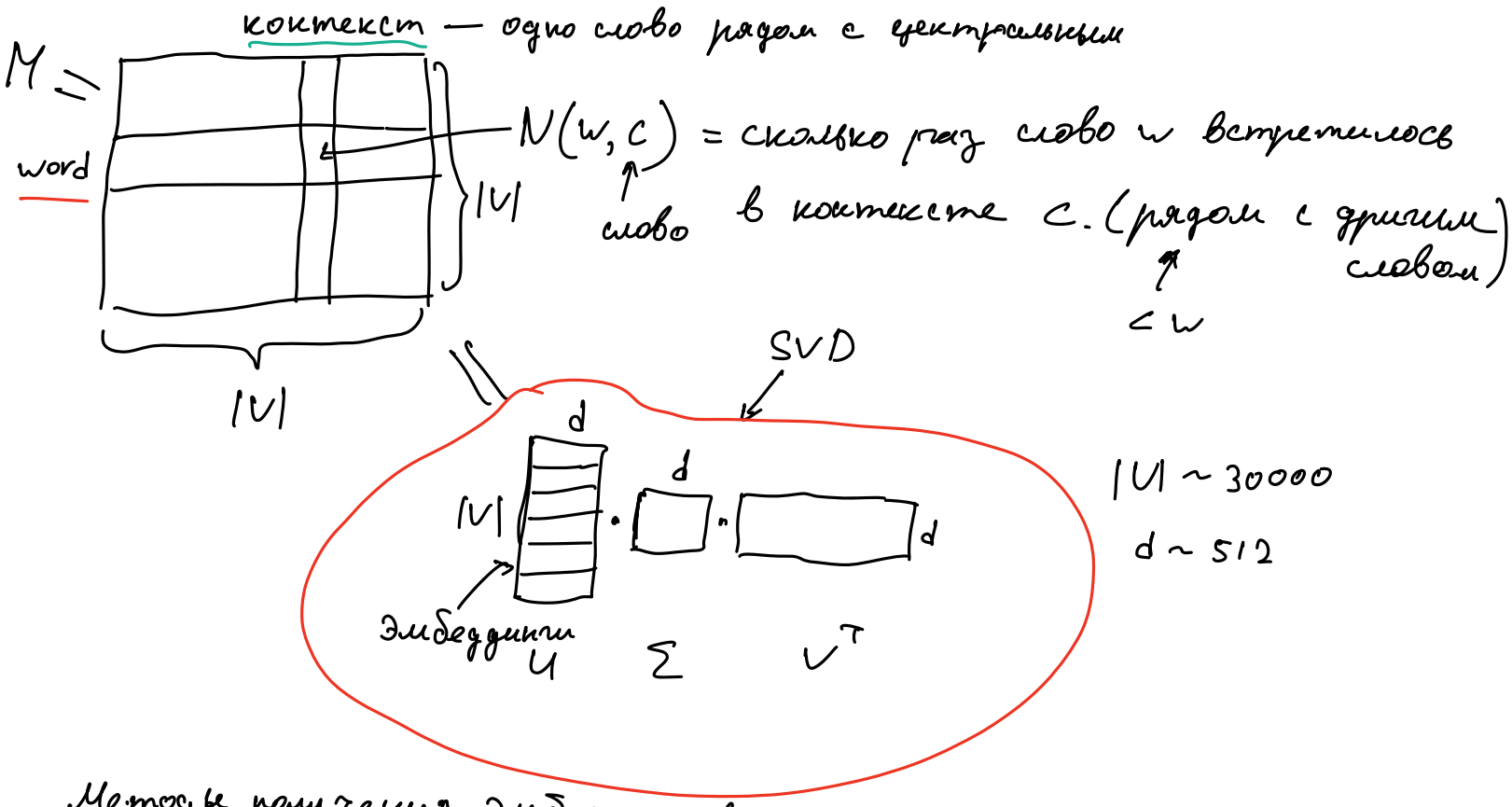
$$\rho(\text{еда}, \text{очки}) = \|\text{еда} - \text{очки}\|_2^2 > \varepsilon$$

Как получить эмбеддинги?

Count-based метод:

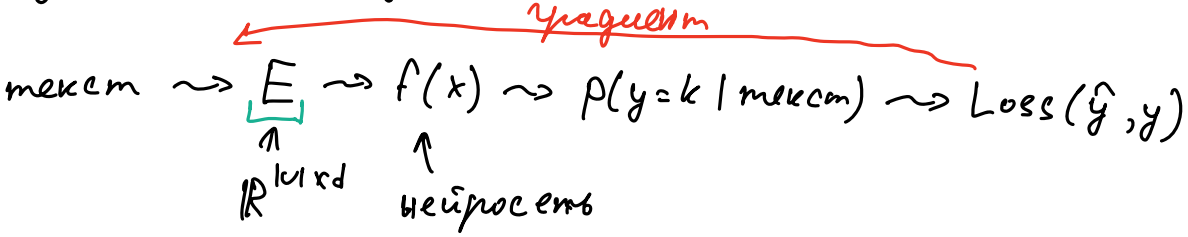
в лесу родилась елочка в лесу она ...

$w = 1$



Методы получения эмбеддингов:

- Word2Vec
- Glove
- Обучаемые эмбеддинги ← современный подход



Пример