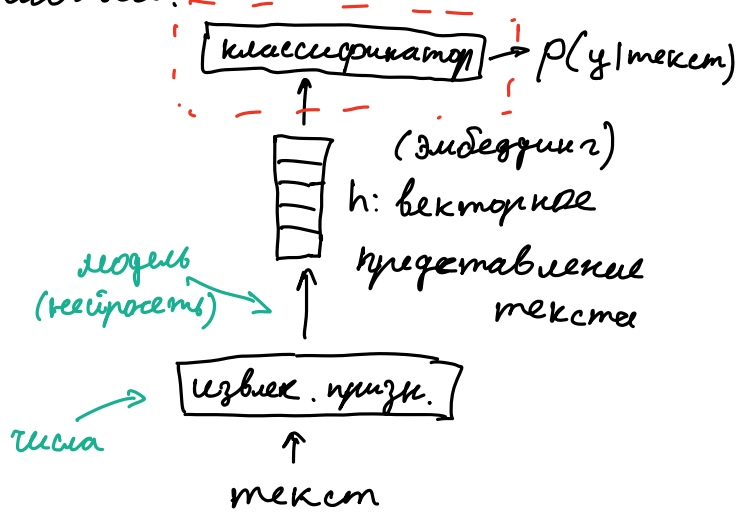


Глубокое обучение в текстах

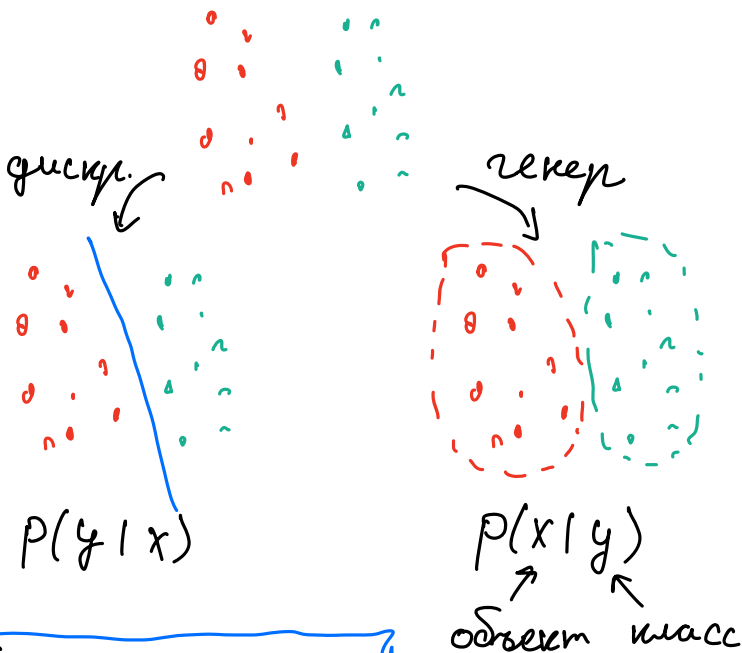
Спам ↔ не спам?

Позитивный отзыв ↔ негативный?

Генерация текста.



Дискриминативные и генеративные модели



$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

$$y^* = \underset{y}{\operatorname{argmax}} P(y|x)$$

$$y^* = \underset{y}{\operatorname{argmax}} P(y|x) = \underset{y}{\operatorname{argmax}} P(x|y)P(y)$$

Наивный Байес

$$y^* = \underset{y}{\operatorname{argmax}} P(y|x) = \underset{y}{\operatorname{argmax}} \underbrace{P(x|y)}_{?} \underbrace{P(y)}$$

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

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

сколько раз x_i было в тексте класса k

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

предположение \leftarrow размер выборки \leftarrow длина текста

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

Почему это работает?

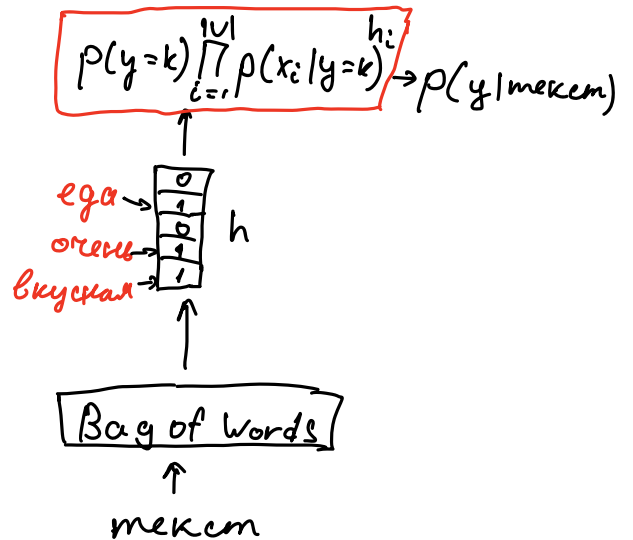
x : Еда | оцень | вкусная

$$p(y|x) \propto \prod_{i=1}^n p(x_i | y=k) \cdot p(y) \quad \frac{1}{22}$$

$p(x_1 | y=+)$ $p(x_1 | y=-)$ $p(x_3 | y=+)$ $p(x_3 | y=-)$

$p(x_3 | y=+) > p(x_3 | y=-)$

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



Логистическая регрессия

линейный $\rightarrow \text{softmax}(\text{logit}(x)) = p(y | \text{текст})$ \leftarrow всегда max



Ручные пр-ки

текст

BOW
tf-idf

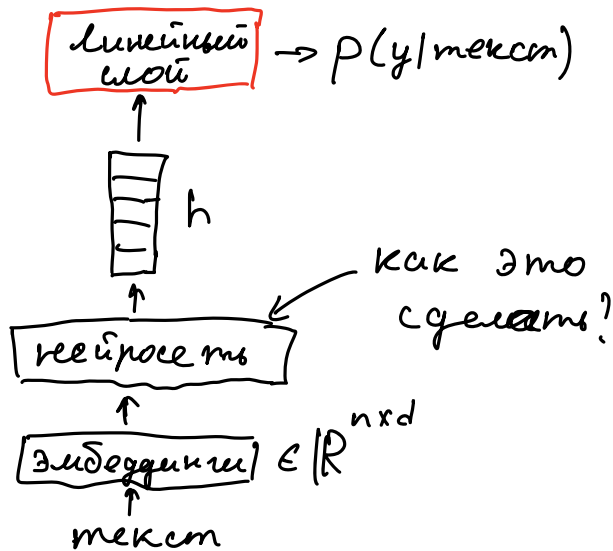
tf-idf = term frequency (tf) •

inverse document frequency (idf)

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

\leftarrow слово w в документе d

Нейросетевые методы



Эмбединги

Сопоставляют каждому слову свой вектор.

$$\text{еда} \leftrightarrow \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}_d$$

эмбединги похожих слов похожи

$$\rho(\text{еда}, \text{клуб}) = \left\| \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}_{\text{еда}} - \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}_{\text{клуб}} \right\|_2^2 \rightarrow 0$$

$$\rho(\text{еда}, \text{орехи}) = \left\| \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}_{\text{еда}} - \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}_{\text{орехи}} \right\|_2^2 > \epsilon$$

Как получить эмбединги?

Count-based метод:

в лесу родился елочка в лесу она ...

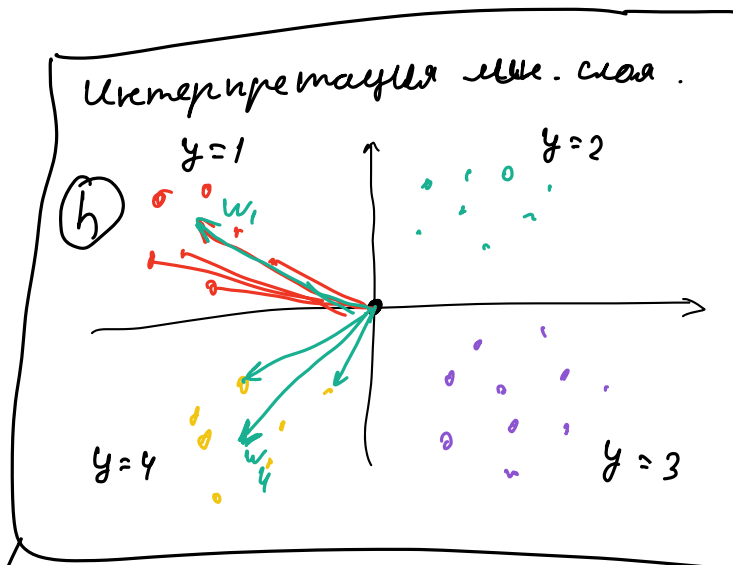
$$[w=1]$$

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

↑
гемма

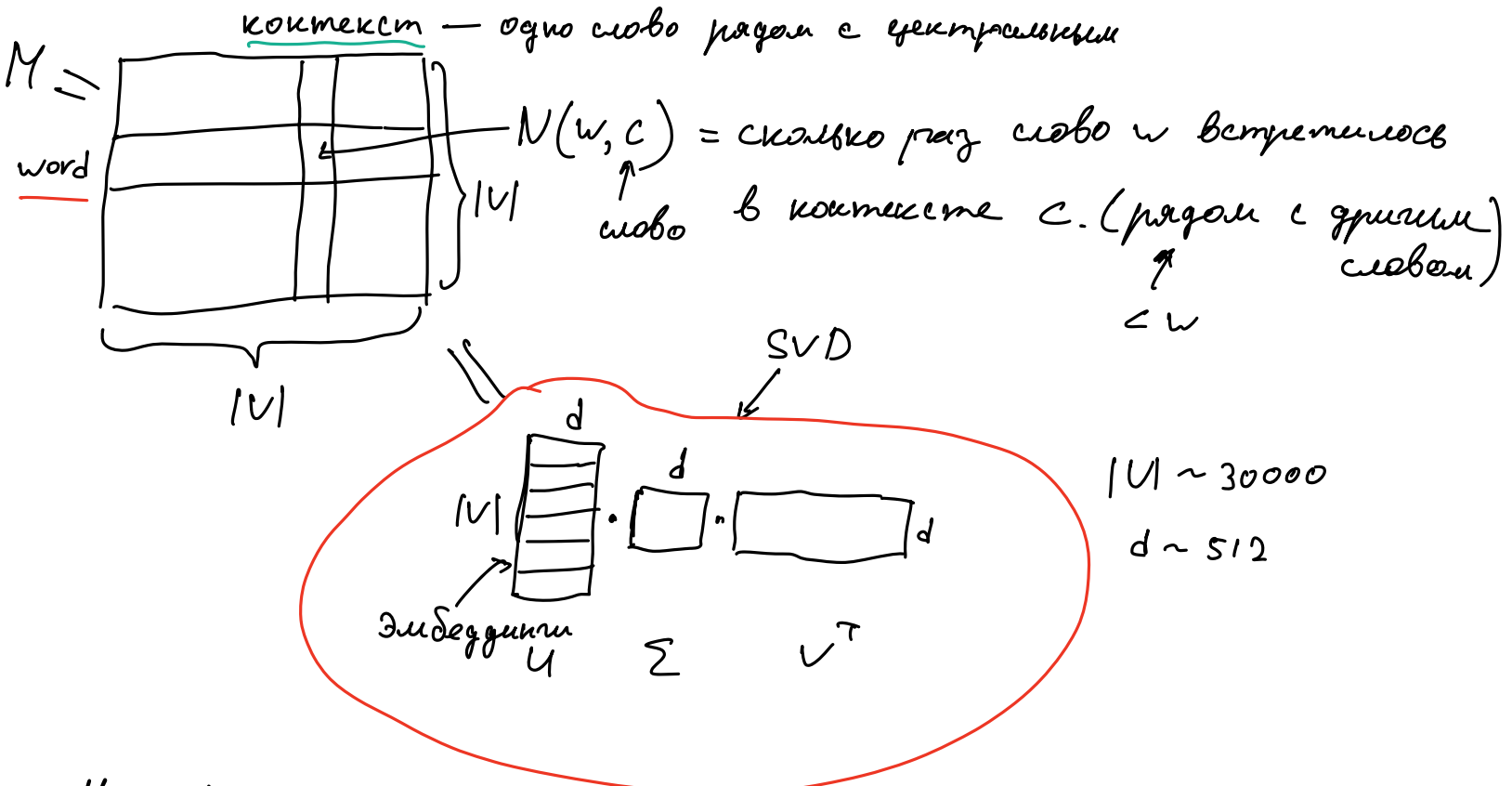
$$\text{tf-idf}(x_1, \dots, x_n) = \left\{ \text{tf}(w_i, x) \cdot \text{idf}(w_i, x, D) \right\}_{i=1}^{|V|}$$

x



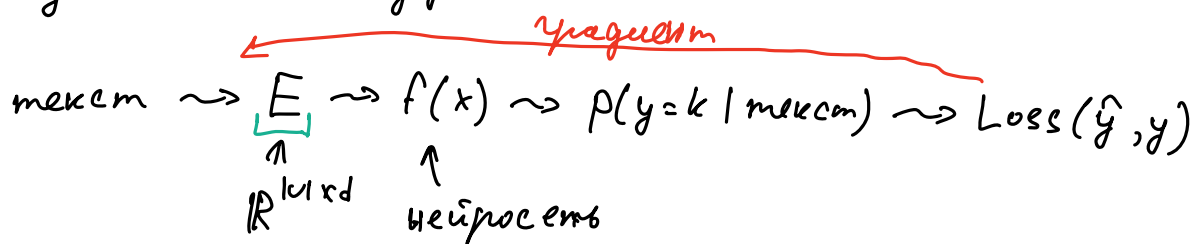
линейный слой

$$\begin{matrix} y=1 \\ y=2 \\ y=3 \\ y=4 \end{matrix} \begin{bmatrix} \vdots \\ \vdots \\ \vdots \\ \vdots \end{bmatrix} \cdot h = \begin{pmatrix} \langle w_1, h \rangle \\ \langle w_2, h \rangle \\ \langle w_3, h \rangle \\ \langle w_4, h \rangle \end{pmatrix} \sim p(y|x)$$



Методы получения эмбедингов:

- Word2Vec
- GloVe
- Обучаемые эмбединги ← современный подход



Пример

