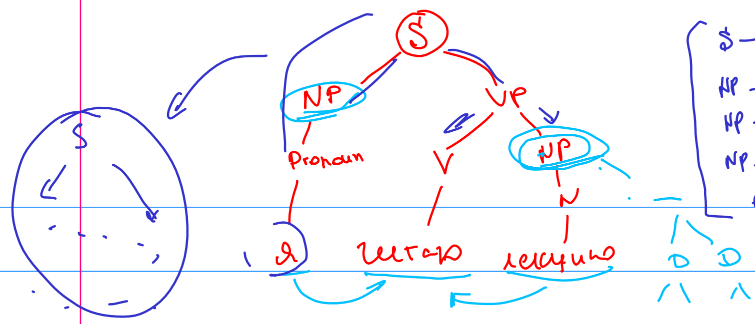
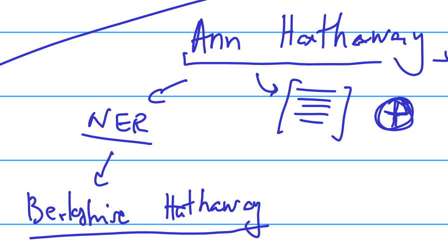
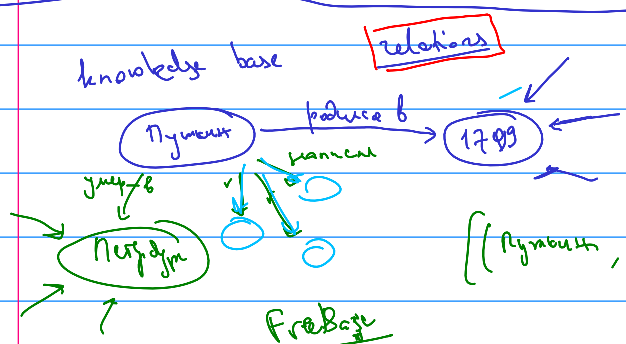


S → NP VP
NP → N
NP → Pronoun
NP → N PP
PP → ...

big round table
~~round big table~~



knowledge base



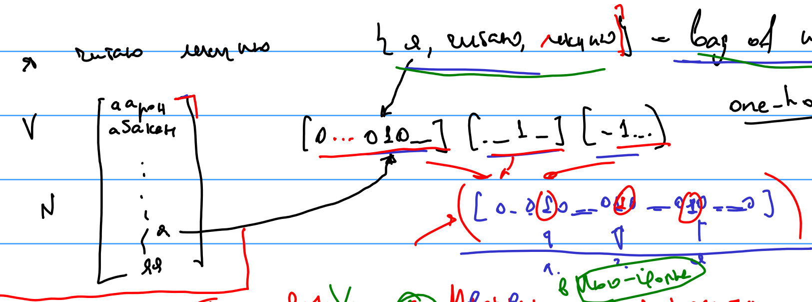
[(Пушкин, родился_в, 1799)]

FreeBase

читаю лекцию

{я, читаю, лекцию} – bag of words, BoW

one-hot



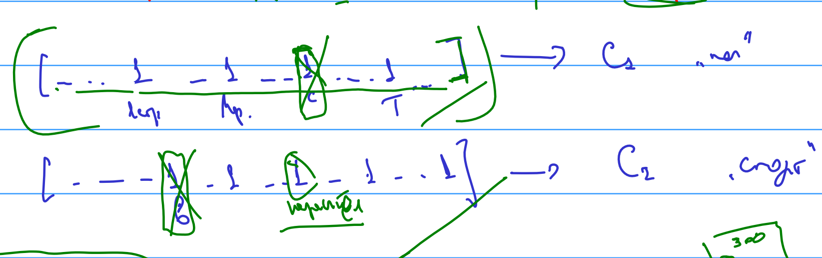
лекцию
~~лента~~
лекция

лекцию
morphology
(лекция, вин.п., ед.ч.)
[... 1 ...][... 1 ...][100]

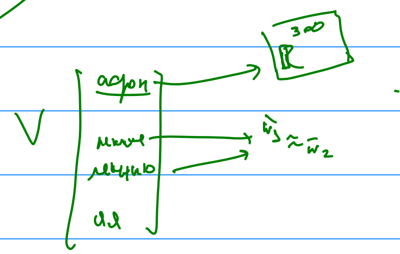
Трамп встретился с Меркель в Нью-Йорке → политика

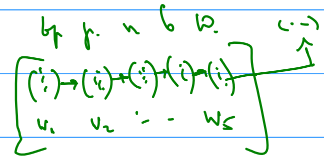
Криштиану Роналду передал в „Ювентус" → спорт

стоп-слова



tf-idf веса





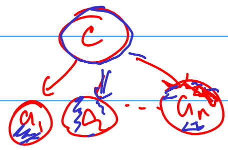
Категор. текстов

Naive Bayes

BoW $[0 \dots 0 \, 1 \, 0 \dots 0]$

$$p(C|\bar{x}) = \frac{p(C)\,p(\bar{x}|C)}{p(\bar{x})} \propto p(C) \cdot \prod_{i=1}^{n} p(a_i|C)$$

$\bar{x} = (a_1, a_2, \dots, a_n)$



$p(C) = \frac{|C|}{|D|}$

$p(w|C) = \frac{\#\{w \in C\}}{|C|}$

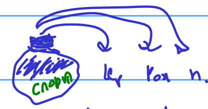
Кр. Рон. нап. в Wiki

$w_1 \; — \; \dots \; w_5$

$p(\text{Спорт}|\bar{s}) \propto p(\text{Спорт}) \cdot p(\text{Крым.}|\text{сп}) \, p(\text{Рон}|\text{сп}) \dots p(\text{Wiki}|\text{сп})$

$p(\text{Пол.}|\bar{s}) \propto p(\text{Пол.}) \; p \dots (\text{пол.}) \dots p \dots (\text{пол.})$

Multinomial NB



k_1 Рон n.

Мат. ожидание

$p(\text{обуч}|\text{мат}) \gg p(\text{обуч})$

Multivariate NB

$[0 \; 0 \; \underset{Кр.}{1} \; 0 \; \underset{Рон}{1} \; 0 \; \underset{Муж.}{0} \; 0] \in \{0,1\}^{|V|}$

Со. – tf-idf

$p(s|C) = \prod_{w \in V} p(w|C)$

$p(\text{Спорт}|\bar{s}) \propto p(\text{Спорт})\,p(\text{Кр}|\text{сп}) \dots [1 - p(\text{Муж.}|\text{сп})] \dots [\dots]$

Кр. Рон. нап. с Wiki контр. на 30 или евро

↘ физ. ↘ физ.

$p(\text{Кр}|\text{ф.}) \cdot p(\text{Рон}|\text{ф.})$

0.001 × 0.001

0.001

$\bar{s} \rightarrow p(\text{спорт}|\bar{s})$ – 75%   0.999 | 0.99 | 0.95 | 0.05

$p(\text{физ.}|\bar{s})$ – 25%   0.001 | 0.01 | 0.05 | 0.95

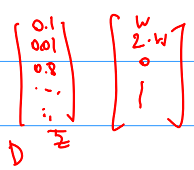
---

1 – BoW ✓

1 – Naive assumption ✓

– есть размеченный датасет

– ≥ 1 ~~doc~~ – 1 метка



1) Нет разметки ⇒ кластеризация

$D = \{s_n\}$

$p(C|s) \propto p(C) \cdot \prod_{w \in s} p(w|C)$

$\pi_t = p(C_t), \quad \varphi_{wt} = p(w|C_t)$

– M-шаг: $\pi_t := \frac{\mathbb{E}\,\#(s \in C_t)}{|D|}$ $\quad \varphi_{wt} := \frac{\mathbb{E}\,\#\{w \in d \in C_t\}}{\mathbb{E} \sum_{d \in C_t} |d|}$ $\quad n_{dw}$

tf-idf

– E-шаг: $p(s_n \in C_t) \propto p(C_t) \prod p(w|C_t) = \pi_t \cdot \prod_{w \in s} \varphi_{wt}$

$\forall d \, \forall t \quad z_{dt} = p(t|d) = \frac{\pi_t \prod \varphi_{wt}}{\sum (\pi_{t'} \prod \varphi_{wt'})}$

$$\mathbb{E}\,\#\{d \in C_t\} = \sum_{d \in D} z_{dt}$$

$$\mathbb{E}\,\#\{w \in d \in C_t\} = \sum_{d \in D} z_{dt} \cdot n_{wd}$$

---

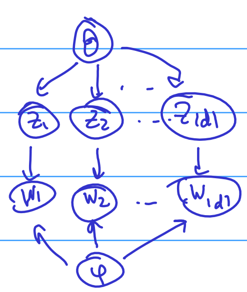
2) **Topic Modeling**

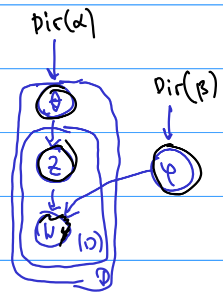
d – распр. вер. на темах $\theta_{dt} = p(t|d)$

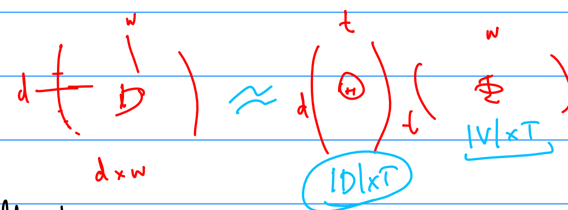
$\varphi_{tw} = p(w|t)$ $\qquad$ pLSA

$d \times t \qquad t \times w$

$$p(d|\Theta, \Phi) = \prod_{w \in d} p(w|\theta, \Phi) = \prod_{w \in d} \sum_{t \in T} \theta_{dt} \varphi_{tw}$$







LDA – Latent Dirichlet allocation