

Тематическое моделирование

Лектор: Алтухов Никита Александрович
Аналитик данных Сбербанк

Мешок слов

Токены

and beautiful blue is king love old queen sky the this

Документы*

0	1	1	1	1	0	0	0	0	1	1	0
1	1	1	0	2	1	0	1	1	0	2	0
2	0	1	1	0	0	1	0	0	1	0	1
3	1	1	0	0	1	0	1	1	0	2	0

*Документ - набор текстовой информации

TF-IDF

Токены

Документы*

	adore	cats	dogs	don	hate	like	love	spiders
0	0.000000	0.000000	0.556451	0.000000	0.000000	0.000000	0.830881	0.000000
1	0.000000	0.000000	0.462208	0.690159	0.000000	0.556816	0.000000	0.000000
2	0.707107	0.707107	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	0.000000	0.000000	0.000000	0.000000	0.707107	0.000000	0.000000	0.707107

$$w_{x,y} = \text{tf}_{x,y} \times \log \left(\frac{N}{\text{df}_x} \right)$$

TF-IDF

Term x within document y

$\text{tf}_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

n-граммы

Токен это: слог, слово, n-грамма

This is Big Data AI Book

Uni-Gram

This	Is	Big	Data	AI	Book
------	----	-----	------	----	------

Bi-Gram

This is	Is Big	Big Data	Data AI	AI Book
---------	--------	----------	---------	---------

Tri-Gram

This is Big	Is Big Data	Big Data AI	Data AI Book
-------------	-------------	-------------	--------------

Тематическое моделирование

- Тема - набор ключевых слов (термов / терминов / токенов), совместно часто встречающихся в документах
- Тема - условное распределение на множестве терминов. Вероятность термина w в теме t , $p(w|t)$
- Тематика документа - условное распределение на множестве тем. Вероятность темы t в документе d , $p(t|d)$
- Тематическая модель автоматически выявляет латентные темы по наблюдаемым частотам термов $p(w|d)$

Задачи тематического моделирования

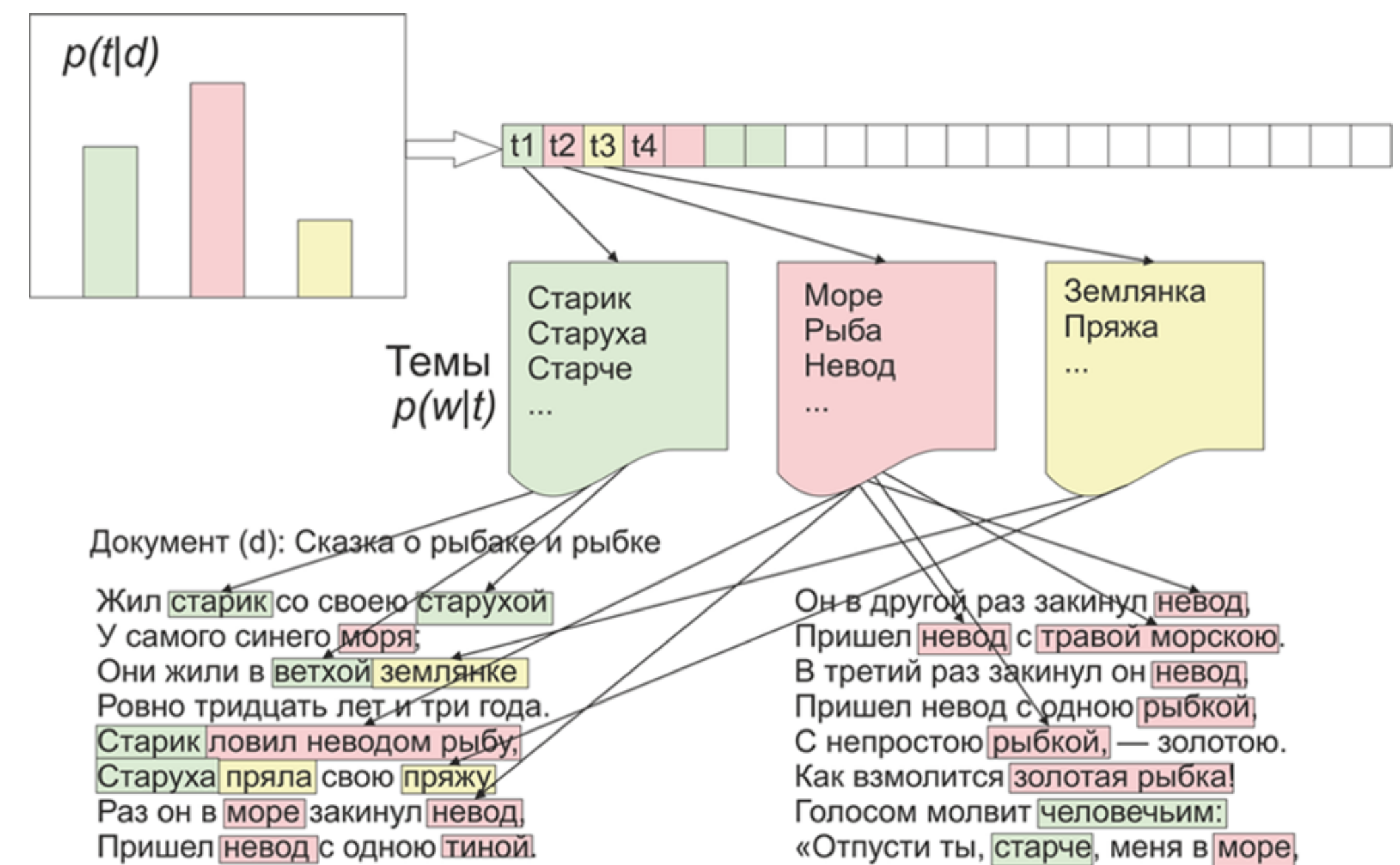
- Классификация и категоризация документов
- Автоматическое аннотирование документов
- Суммаризация коллекции текстов
- Сегментация (разделение большого документа по темам)
- Рекомендательная система

Базовые предположения

- Порядок документов в коллекции не важен
- Порядок слов в документе не важен (Bag-of-Words)
- Каждая пара (d, w) связана с некоторой темой $t \in T$. Тема латентна
- Гипотеза условной независимости: слова в документе зависят только от темы и не зависят от самого документа $p(w|t, d) = p(w|t)$

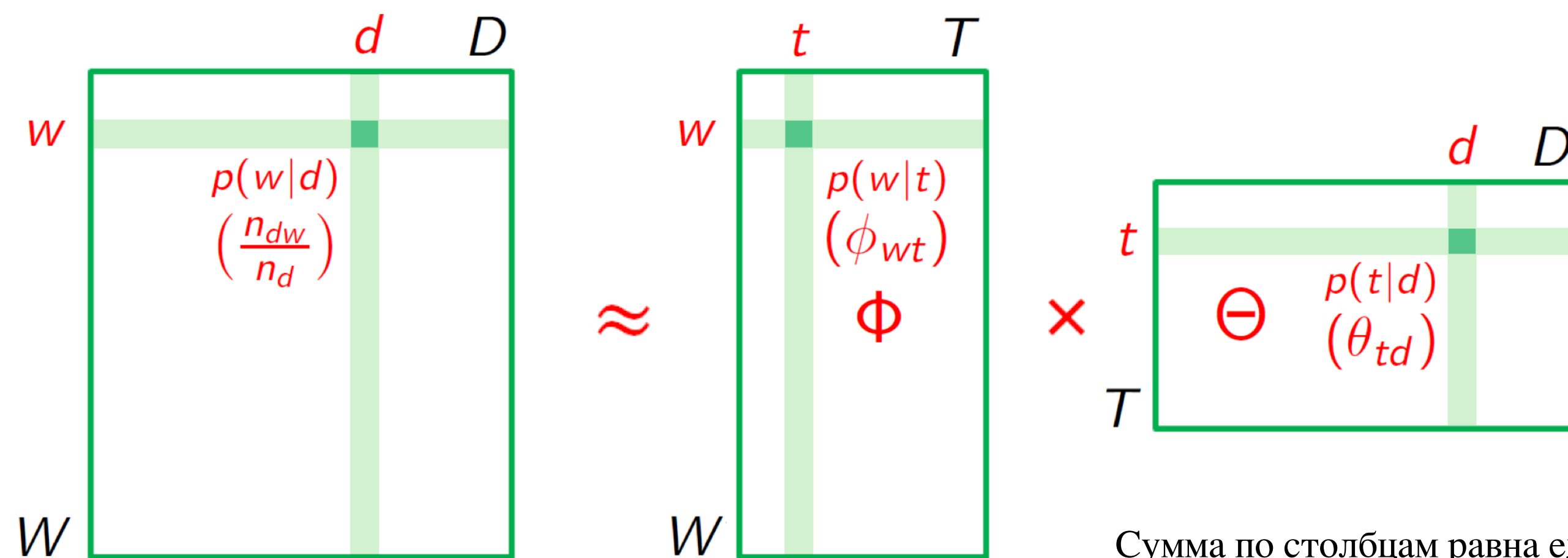
Вероятностный процесс порождения текста

- Документ d - это смесь распределений $p(w|t)$ с весами $p(t|d)$
- Для каждой позиции определяется тема, к которой слово будет относиться по условному распределению тем в документе
- Берем распределение слов в теме и генерируем слово для позиции в документе



Формальная постановка задачи

- Дано:
 - W - словарь термов (слов или словосочетаний)
 - D - коллекция текстовых документов
 - n_{dw} - сколько раз термин w встретился в документе d
- Найти параметры вероятностной тематической модели:



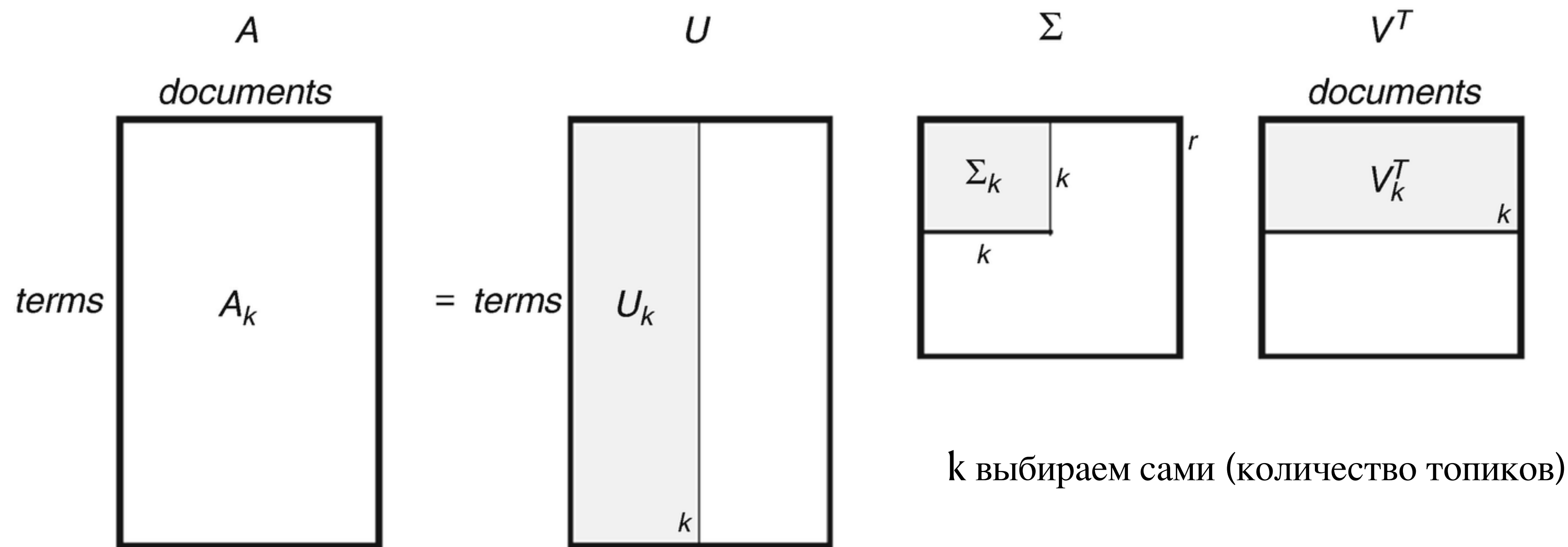
$$p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$$

$$p(w|d) = \sum_t p(w|t) p(t|d),$$

Сумма по столбцам равна единице => распределение вероятностей

LSA

Латентно-семантический анализ

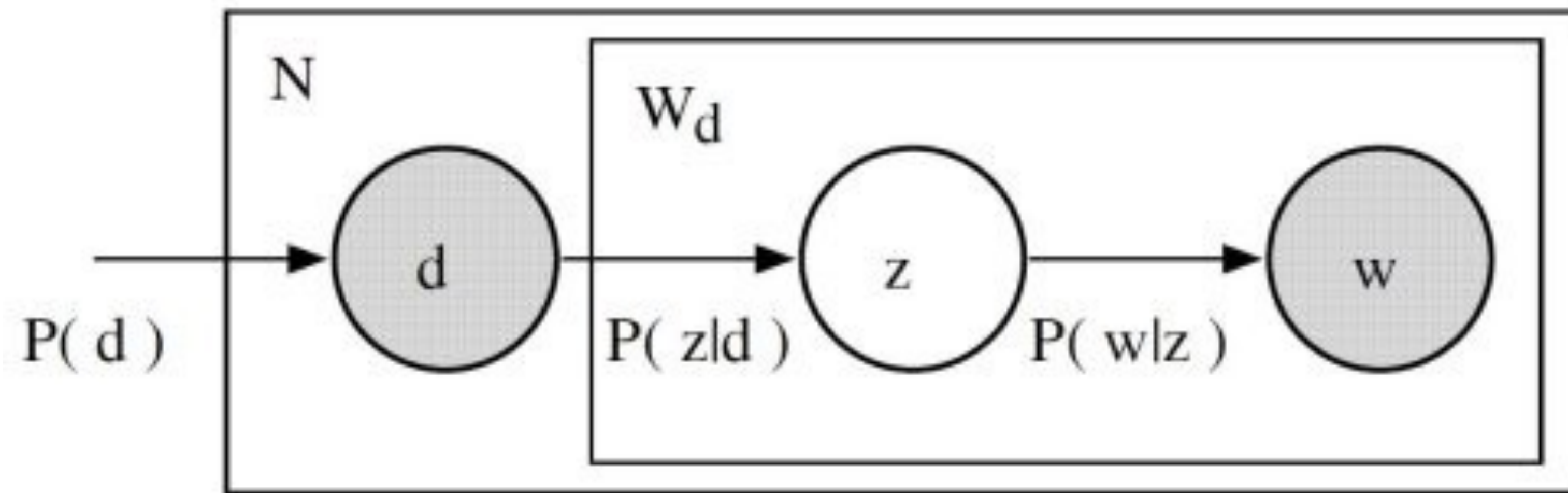


k выбираем сами (количество топиков)

SVD - разложение
(Сингулярное разложение)

pLSA

Вероятностный латентно-семантический анализ



В итоге совместная вероятность
увидеть документ и слово:

$$P(D, W) = P(D) \sum_Z P(Z|D) P(W|Z)$$

<https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05>

pLSA

Вероятностный латентно-семантический анализ

$$P(D, W) = P(D) \sum_Z P(Z|D) P(W|Z)$$

Правая часть этого уравнения сообщает нам, насколько вероятно, увидеть какой-то документ, а затем на основе распределения тем этого документа, насколько вероятно найти определенное слово в этом документе

pLSA

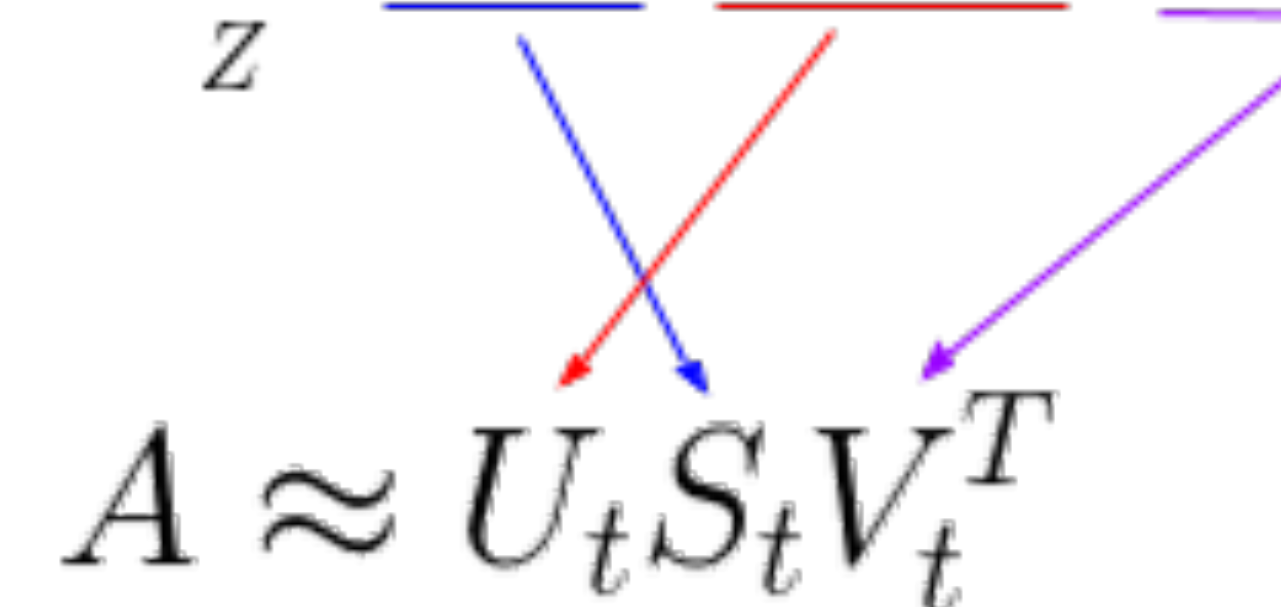
Вероятностный латентно-семантический анализ

$$P(D, W) = \sum_Z P(Z) P(D|Z) P(W|Z)$$

Если раскрутить логику в обратную сторону, начав с темы, то получим эквивалентное уравнение. Начнем с темы $P(z)$, а затем независимо сгенерируем документ с $P(d|z)$ и слово с $P(w|z)$.

Связь LSA и pLSA

$$P(D, W) = \sum_Z \underbrace{P(Z)}_{\text{blue}} \underbrace{P(D|Z)}_{\text{red}} \underbrace{P(W|Z)}_{\text{purple}}$$

$$A \approx \underbrace{U_t}_{\text{red}} \underbrace{S_t}_{\text{blue}} \underbrace{V_t^T}_{\text{purple}}$$


ЕМ-алгоритм ([англ. *Expectation-maximization \(EM\) algorithm*](#)) — алгоритм, используемый в [математической статистике](#) для нахождения оценок [максимального правдоподобия](#) параметров вероятностных моделей, в случае, когда модель зависит от некоторых [скрытых переменных](#).

ЕМ-алгоритм

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

ЕМ-алгоритм: метод простой итерации для системы уравнений

Вычисление условных распределений тем
для каждого слова в каждом документе

Частотные вероятности оценки слов в
темах и тем в документах

$$\begin{aligned} \text{Е-шаг:} & \begin{cases} p_{tdw} = \text{norm}_{t \in T}(\phi_{wt} \theta_{td}) \end{cases} \\ \text{М-шаг:} & \begin{cases} \phi_{wt} = \text{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} \right) \\ \theta_{td} = \text{norm}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} \right) \end{cases} \end{aligned}$$

где $\text{norm}_{t \in T}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$ — операция нормировки вектора.

Минусы pLSA

- Количество параметров в модели pLSA растет линейно с количеством документов, следовательно, очень часто переобучается, нужно использовать регуляризацию
- Нет единственного решения у задачи, следовательно без регуляризации решение неустойчиво
- Медленно сходится
- Нет управления разреженностью Φ и Θ

<https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05>

<http://www.machinelearning.ru/wiki/images/o/o2/Voron20mipt-ptm-emlda.pdf>

LDA

Латентное размещение Дирихле

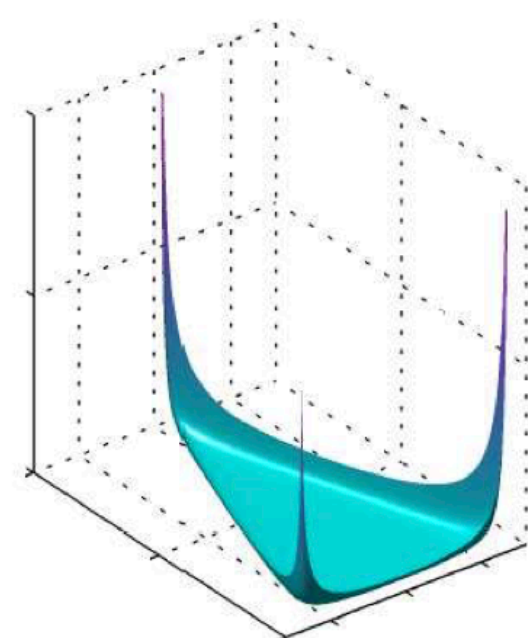
- Вектор-столбцы $\phi_t = (\phi_{wt})_{w \in W}$ и $\theta_d = (\theta_{td})_{t \in T}$ порождаются из распределения Дирихле с параметрами $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$

$$\text{Dir}(\phi_t | \beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \phi_{wt} > 0; \quad \beta_0 = \sum_w \beta_w, \quad \beta_w > 0;$$

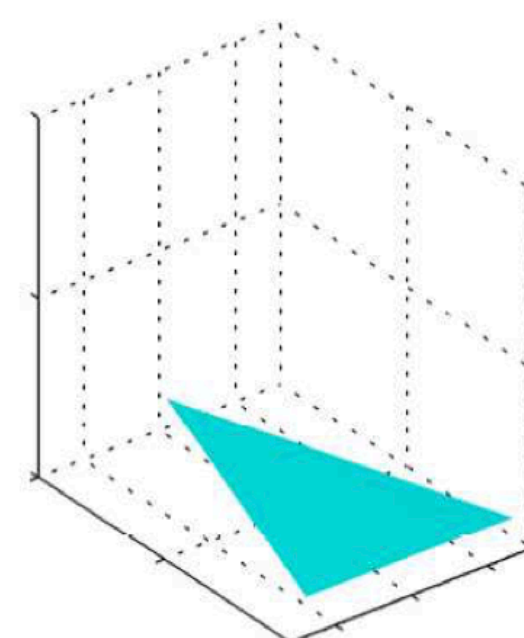
$$\text{Dir}(\theta_d | \alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \theta_{td} > 0; \quad \alpha_0 = \sum_t \alpha_t, \quad \alpha_t > 0;$$

Пример:

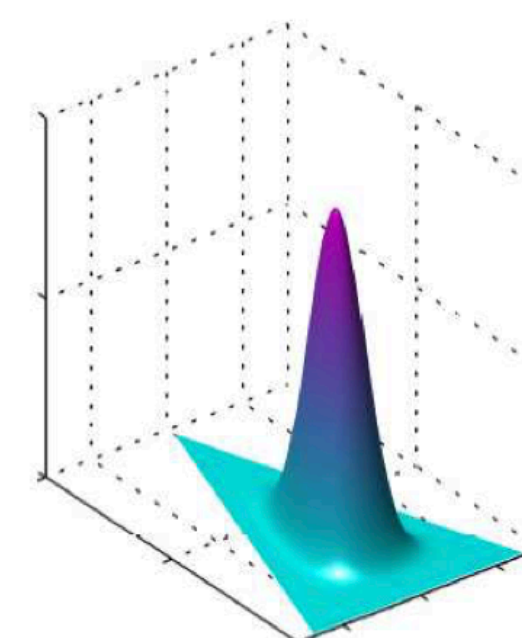
$\text{Dir}(\theta | \alpha)$,
 $|T| = 3$,
 $\theta, \alpha \in \mathbb{R}^3$



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$
Разреженные вектора!!



$\alpha_1 = \alpha_2 = \alpha_3 = 1$



$\alpha_1 = \alpha_2 = \alpha_3 = 10$

LDA

Тематическая модель

$$p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}, \quad \phi_t \sim \text{Dir}(\phi|\beta), \quad \theta_d \sim \text{Dir}(\theta|\alpha)$$

Вход: векторы гиперпараметров β, α ;

Выход: коллекция документов;

выбрать вектор ϕ_t из $\text{Dir}(\phi|\beta)$ для каждой темы $t \in T$;

выбрать вектор θ_d из $\text{Dir}(\theta|\alpha)$ для каждого документа $d \in D$;

для всех документов $d \in D$

для всех позиций термов $i = 1, \dots, n_d$ в документе d

 выбрать тему t_i из $p(t|d) \equiv \theta_{td}$;

 выбрать терм w_i из $p(w|t_i) \equiv \phi_{wt_i}$;

Выводы

- Можно так же рассматривать через Дивергенцию Кульбака-Лейблера, это позволяет снять ограничения $\beta_w > 0, \alpha_t > 0$
- LDA и pLSA почти не отличаются на больших данных
- LDA имеет больше параметров по сравнению с pLSA
- Популярность LDA немного переоценена, робастные pLSA (с регуляризацией и разными примочками) почти не отличаются по перплексии от LDA

Список источников

GitHub: <https://github.com/nikitosl/spbu-nlp-2020>

<https://medium.com/technovators/topic-modeling-art-of-storytelling-in-nlp-4dc83e96a987>

<http://www.machinelearning.ru/wiki/images/b/bc/Voron-2015-BigARTM.pdf>

LSA: <https://habr.com/ru/post/110078/>

<https://habr.com/ru/post/230075/>

<https://habr.com/ru/post/240209/>

pLSA: <https://towardsdatascience.com/topic-modelling-with-plsa-728b92043f41>

<http://www.machinelearning.ru/wiki/index.php?title=PLSA>

SVD: <https://habr.com/ru/company/surfingbird/blog/139863/>

<https://habr.com/ru/company/yandex/blog/313892/>

LDA: <https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd>

<https://habr.com/ru/company/surfingbird/blog/150607/>

<https://habr.com/ru/post/417167/>

<https://logic.pdmi.ras.ru/~sergey/teaching/mlkfu14/17-lda.pdf>

<https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24>