

# Вероятностные тематические модели

## Лекция 2.

### Онлайновый EM-алгоритм и аддитивная регуляризация

К. В. Воронцов  
k.vorontsov@iai.msu.ru

Этот курс доступен на странице вики-ресурса  
<http://www.MachineLearning.ru/wiki>  
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

МФТИ – ФИЦ ИУ РАН • 3 октября 2024

- 1 Часто используемые регуляризаторы**
  - Сглаживание и разреживание
  - Декоррелирование
  - Разреживающий регуляризатор для отбора тем
- 2 Алгоритмическая реализация ARTM**
  - Рациональный и онлайн-EM-алгоритм
  - Комбинирование регуляризаторов
  - Библиотеки BigARTM и TopicNet
- 3 Эксперименты с тематическими моделями**
  - Производительность BigARTM
  - Измерение качества тематических моделей
  - Комбинирование регуляризаторов

## Напоминание. Задача тематического моделирования

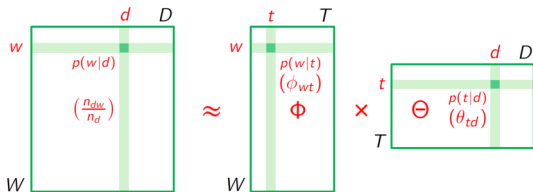
**Дано:** коллекция текстовых документов,  $p(w|d) = \frac{n_{dw}}{n_d}$

Вероятностная тематическая модель:

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

**Найти:** параметры модели  $\phi_{wt} = p(w|t)$ ,  $\theta_{td} = p(t|d)$

Это задача стохастического матричного разложения:



Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. NIPS-2001. JMLR 2003.

## Напоминание. ARTM — аддитивная регуляризация

Максимизация log правдоподобия с регуляризатором  $R$ :

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} \equiv p(t|d, w) = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \operatorname{norm}_{w \in W} \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), & n_{wt} = \sum_{d \in D} n_{dw} p_{tdw} \\ \theta_{td} = \operatorname{norm}_{t \in T} \left( n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), & n_{td} = \sum_{w \in W} n_{dw} p_{tdw} \end{cases} \end{cases}$$

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

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014.

## Напоминание. Комбинирование регуляризаторов в ARTM

Максимизация  $\log$  правдоподобия с  $k$  регуляризаторами  $R_i$ :

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + \sum_{i=1}^k \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

где  $\tau_i$  — коэффициенты регуляризации.

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \mathop{\text{norm}}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \mathop{\text{norm}}_{w \in W} \left( \sum_{d \in D} n_{dw} p_{tdw} + \phi_{wt} \sum_{i=1}^k \tau_i \frac{\partial R_i}{\partial \phi_{wt}} \right) \\ \theta_{td} = \mathop{\text{norm}}_{t \in T} \left( \sum_{w \in W} n_{dw} p_{tdw} + \theta_{td} \sum_{i=1}^k \tau_i \frac{\partial R_i}{\partial \theta_{td}} \right) \end{cases} \end{cases}$$

Vorontsov K., Potapenko A. Additive regularization of topic models.  
 Machine Learning, 2015.

## Дивергенция Кульбака–Лейблера и её свойства

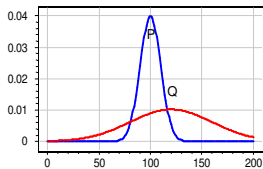
Функция расстояния между распределениями  $P = (p_i)_{i=1}^n$  и  $Q = (q_i)_{i=1}^n$ :

$$KL(P\|Q) \equiv KL_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

1.  $KL(P\|Q) \geq 0$ ;  $KL(P\|Q) = 0 \Leftrightarrow P = Q$ ;
2. Минимизация  $KL$  эквивалентна максимизации правдоподобия:

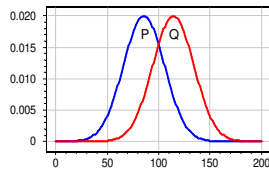
$$KL(P\|Q(\alpha)) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i(\alpha)} \rightarrow \min_{\alpha} \iff \sum_{i=1}^n p_i \ln q_i(\alpha) \rightarrow \max_{\alpha}.$$

3. Если  $KL(P\|Q) < KL(Q\|P)$ , то  $P$  сильнее вложено в  $Q$ , чем  $Q$  в  $P$ :



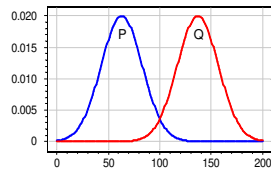
$$KL(P\|Q) = 0.44$$

$$KL(Q\|P) = 2.97$$



$$KL(P\|Q) = 0.44$$

$$KL(Q\|P) = 0.44$$



$$KL(P\|Q) = 2.97$$

$$KL(Q\|P) = 2.97$$

## Регуляризатор сглаживания

**Гипотеза** сглаженности:

распределения  $\phi_{wt}$  близки к заданному распределению  $\beta_w$ ;  
 распределения  $\theta_{td}$  близки к заданному распределению  $\alpha_t$ .

$$\sum_{t \in T} \text{KL}(\beta_w \parallel \phi_{wt}) \rightarrow \min_{\Phi}; \quad \sum_{d \in D} \text{KL}(\alpha_t \parallel \theta_{td}) \rightarrow \min_{\Theta}.$$

Максимизируем сумму регуляризаторов:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \rightarrow \max.$$

Подставляем, получаем формулы М-шага, похожие на LDA  
 (однако в LDA есть ограничения  $\beta_0 \beta_w > -1$ ,  $\alpha_0 \alpha_t > -1$ ):

$$\phi_{wt} = \text{norm}_{w \in W}(n_{wt} + \beta_0 \beta_w), \quad \theta_{td} = \text{norm}_{t \in T}(n_{td} + \alpha_0 \alpha_t).$$

## Регуляризатор разреживания

**Гипотеза** разреженности: среди  $\phi_{wt}$ ,  $\theta_{td}$  много нулей;  
 распределения  $\phi_{wt}$  **далеки** от заданного распределения  $\beta_w$ ;  
 распределения  $\theta_{td}$  **далеки** от заданного распределения  $\alpha_t$ .

$$\sum_{t \in T} \text{KL}(\beta_w \| \phi_{wt}) \rightarrow \max_{\Phi}; \quad \sum_{d \in D} \text{KL}(\alpha_t \| \theta_{td}) \rightarrow \max_{\Theta}.$$

Максимизируем сумму регуляризаторов:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \rightarrow \max.$$

Это обобщение LDA, снимающее ограничения на  $\alpha_t, \beta_w$ :

$$\phi_{wt} = \text{norm}_{w \in W}(n_{wt} - \beta_0 \beta_w), \quad \theta_{td} = \text{norm}_{t \in T}(n_{td} - \alpha_0 \alpha_t).$$

---

*Varadarajan J., Emonet R., Odoñez J.-M.* A sparsity constraint for topic models — application to temporal activity mining. NIPS-2010.



## Объединение сглаживания и разреживания

Общий вид регуляризаторов сглаживания и разреживания:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \rightarrow \max,$$

где  $\beta_0 > 0$ ,  $\alpha_0 > 0$  — коэффициенты регуляризации,  
 $\beta_{wt}$ ,  $\alpha_{td}$  — параметры, задаваемые пользователем:

- $\beta_{wt} > 0$ ,  $\alpha_{td} > 0$  — сглаживание
- $\beta_{wt} < 0$ ,  $\alpha_{td} < 0$  — разреживание

**Возможные применения** сглаживания и разреживания:

- скорректировать состав термов и документов темы
- задать предметные темы со специальной лексикой
- задать фоновые темы с общей лексикой языка
- задать псевдо-документ с ключевыми терминами темы

## Частичное обучение (semi-supervised learning)

Общий вид регуляризаторов сглаживания и разреживания:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \rightarrow \max,$$

**Идея:** в построенной модели можно скорректировать темы, добавляя и удаляя в них термы и документы.

Разреживание по «чёрным спискам»:

- $\beta_{wt} = -\frac{1}{|W_t|} [w \in W_t]$  — термов из  $W_t$  не должно быть в  $t$
- $\alpha_{td} = -\frac{1}{|T_d|} [t \in T_d]$  — тем из  $T_d$  не должно быть в  $d$

Сглаживание по «белым спискам»:

- $\beta_{wt} = \frac{1}{|W_t|} [w \in W_t]$  — термы из  $W_t$  должны быть в  $t$
- $\alpha_{td} = \frac{1}{|T_d|} [t \in T_d]$  — темы из  $T_d$  должны быть в  $d$

## Проблема $\ln 0$ в дивергенции Кульбака–Лейблера

В регуляризаторе сглаживания/разреживания

$$R(\Phi) = \beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln \phi_{wt} \rightarrow \max$$

не возникает ли проблема с  $\ln \phi_{wt}$  при  $\phi_{wt} = 0$  или  $\phi_{wt} \rightarrow 0$ ?

Подправим регуляризатор, при сколь угодно малом  $\varepsilon$ :

$$R(\Phi) = \beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln(\phi_{wt} + \varepsilon) \rightarrow \max.$$

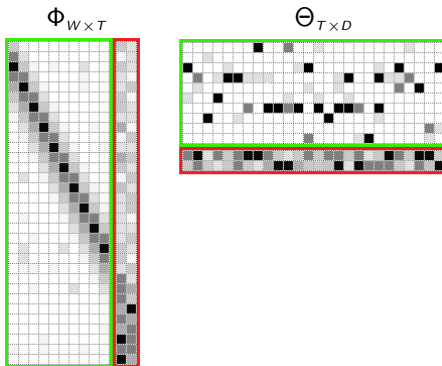
Подставив в формулу M-шага, получим для всех  $t \in S$ :

$$\phi_{wt} = \operatorname{norm}_{w \in W} \left( n_{wt} + \beta_0 \beta_w \frac{\phi_{wt}}{\phi_{wt} + \varepsilon} \right).$$

Поправка не нужна, т.к. практически не влияет на результат

## Разделение тем на предметные и фоновые

*Предметные темы*  $S$  содержат термины предметной области,  
 $p(w|t)$ ,  $p(t|d)$ ,  $t \in S$  — разреженные, существенно различные  
*Фоновые темы*  $B$  содержат слова общей лексики,  
 $p(w|t)$ ,  $p(t|d)$ ,  $t \in B$  — существенно отличные от нуля



## Регуляризатор декоррелирования тем

**Цель:** сделать темы как можно более различными, выделить для каждой темы *лексическое ядро* — набор термов, отличающий её от других тем.

Минимизируем ковариации между вектор-столбцами  $\phi_t$ :

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \rightarrow \max.$$

Подставляем в формулы M-шага, получаем ещё один вариант разреживания — контрастирование строк матрицы  $\Phi$  (малые вероятности  $\phi_{wt}$  в строке становятся ещё меньше):

$$\phi_{wt} = \text{norm}_{w \in W} \left( n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \right).$$

## Разреживающий регуляризатор для отбора тем

**Цель:** избавиться от незначимых тем (topic selection).

Разреживаем распределение  $p(t) = \sum_d p(d)\theta_{td}$ , максимизируя кросс-энтропию между  $p(t)$  и равномерным распределением:

$$R(\Theta) = -\tau \sum_{t \in T} \ln \sum_{d \in D} p(d)\theta_{td} \rightarrow \max.$$

Подставляем, получаем:

$$\theta_{td} = \operatorname{norm}_{t \in T} \left( n_{td} - \tau \frac{n_d}{n_t} \theta_{td} \right), \text{ вариант: } \theta_{td} = \operatorname{norm}_{t \in T} \left( n_{td} \left( 1 - \frac{\tau}{n_t} \right) \right).$$

**Эффект:** обнуляются строки матрицы  $\Theta$  с малыми  $n_t$ , заодно (неожиданно) удаляются зависимые и расщеплённые темы.

---

*Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization. SLDS 2015.*

## Рациональный EM-алгоритм

**Идея:** E-шаг встраивается внутрь M-шага для каждого  $d \in D$ , чтобы не хранить трёхмерный массив значений  $n_{dwt}$ .

**Вход:** коллекция  $D$ , число тем  $|T|$ , число итераций  $i_{\max}$ ;

**Выход:** матрицы термов тем  $\Phi$  и термов документов  $\Theta$ ;

инициализация  $\phi_{wt}, \theta_{td}$  для всех  $d \in D, w \in W, t \in T$ ;

**для всех** итераций  $i = 1, \dots, i_{\max}$

$n_{wt}, n_{td} := 0$  для всех  $d \in D, w \in W, t \in T$ ;

**для всех** документов  $d \in D$  и всех термов  $w \in W$

$n_{tdw} := n_{dw} \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td})$  для всех  $t \in T$ ;

$n_{wt} += n_{tdw}; n_{td} += n_{tdw}$  для всех  $t \in T$ ;

$\phi_{wt} := \operatorname{norm}_{w \in W} \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$  для всех  $w \in W, t \in T$ ;

$\theta_{td} := \operatorname{norm}_{t \in T} \left( n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$  для всех  $d \in D, t \in T$ ;

## Онлайновый EM-алгоритм

**Вход:** коллекция  $D$ , число тем  $|T|$ , параметры  $j_{\max}$ ,  $\gamma$ ;

**Выход:** матрицы термов тем  $\Phi$  и термов документов  $\Theta$ ;

инициализировать  $n_{wt} := 0$ ;  $\tilde{n}_{wt} := 0$ ;  $\phi_{wt} := \text{random}$ ;

**для всех** документов  $d \in D$

инициализировать  $\theta_{td} := \frac{1}{|T|}$ ;

**для всех**  $j = 1, \dots, j_{\max}$  (итерации по документу)

$n_{tdw} := n_{dw} \text{norm}_{t \in T}(\phi_{wt} \theta_{td})$  для всех  $w \in d$ ;

$\theta_{td} := \text{norm}_{t \in T} \left( \sum_w n_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$ ;

$\tilde{n}_{wt} := \tilde{n}_{wt} + n_{tdw}$  для всех  $w \in d$ ;

**если** пора обновить матрицу  $\Phi$  **то**

$n_{wt} := \gamma n_{wt} + \tilde{n}_{wt}$ ;  $\tilde{n}_{wt} := 0$ ;

$\phi_{wt} := \text{norm}_{w \in W} \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ ;



## Пакетный онлайнный EM-алгоритм в BigARTM

Коллекция  $D$  разбивается на пакеты  $D_b$ ,  $b = 1, \dots, B$ , которые могут обрабатываться параллельно и/или распределённо.

**Вход:** коллекция документов  $D$ , число тем  $|T|$ ,  
параметры  $\delta \equiv \text{decay\_weight}$ ,  $\alpha \equiv \text{apply\_weight}$ ;

**Выход:** матрица  $\Phi$ ;

инициализировать  $n_{wt} := 0$ ,  $\tilde{n}_{wt} := 0$ ,  $\phi_{wt} := \text{random}$ ;

**для всех** пакетов  $D_b$ ,  $b = 1, \dots, B$

$(\tilde{n}_{wt}) := (\tilde{n}_{wt}) + \text{ProcessBatch}(D_b, \Phi)$ ;

**если** пора обновить матрицу  $\Phi$  **то**

$n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}$ ,  $\tilde{n}_{wt} := 0$ ;

$\phi_{wt} := \text{norm}_{w \in W} \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ ;

---

*Oleksandr Frei, Murat Apishev.* Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016.

## Пакетный онлайнный EM-алгоритм: функция ProcessBatch

Функция **ProcessBatch** обрабатывает пакет документов  $D_b$ , не меняя матрицу  $\Phi$ , и выдаёт счётчики термов в темах  $\tilde{n}_{wt}$ .

**Вход:** пакет  $D_b$ , матрица  $\Phi = (\phi_{wt})$ , параметр  $j_{\max}$ ;

**Выход:** матрица счётчиков  $(\tilde{n}_{wt})_{W \times T}$ ;

инициализировать  $\tilde{n}_{wt} := 0$ ;

**для всех**  $d \in D_b$

инициализировать  $\theta_{td} := \frac{1}{|T|}$ ;

**для всех**  $j = 1, \dots, j_{\max}$  (итерации по документу)

$p_{tdw} := \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td})$ ;

$\theta_{td} := \operatorname{norm}_{t \in T} \left( \sum_{w \in d} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$ ;

$\tilde{n}_{wt} := \tilde{n}_{wt} + n_{dw} p_{tdw}$ ;

## Сравнение оффлайнного и онлайнного алгоритмов

### Оффлайн EM-алгоритм:

- 1 многократное итерирование по коллекции
- 2 однократный проход по документу
- 3 хранение матрицы  $\Theta$
- 4 обновление  $\Phi$  в конце каждого прохода по коллекции
- 5 применяется при обработке небольших коллекций

### Онлайн EM-алгоритм:

- 1 однократный проход по коллекции
- 2 многократное итерирование по каждому документу
- 3 нет необходимости хранить матрицу  $\Theta$
- 4 обновление  $\Phi$  через заданное число пакетов
- 5 применяется при потоковой обработке больших коллекций

## Альтернативная матричная реализация EM-алгоритма

EM-алгоритм (результат E-шага  $p(t|d, w)$  встроен в M-шаг):

$$\phi_{wt} = \operatorname{norm}_{w \in W} \left( \phi_{wt} \sum_{d \in D} n_{dw} \frac{\theta_{td}}{(\Phi \Theta)_{wd}} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$$

$$\theta_{td} = \operatorname{norm}_{t \in T} \left( \theta_{td} \sum_{w \in W} n_{dw} \frac{\phi_{wt}}{(\Phi \Theta)_{wd}} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$$

Матричная запись (norm — нормировка по столбцам):

$$\Phi := \operatorname{norm}(\Phi \otimes (N \oslash \Phi \Theta) \Theta^T + \Phi \otimes \nabla_{\Phi} R)$$

$$\Theta := \operatorname{norm}(\Theta \otimes \Phi^T (N \oslash \Phi \Theta) + \Theta \otimes \nabla_{\Theta} R)$$

где  $N = (n_{dw})$  —  $W \times D$ -матрица исходных данных,

$\otimes$  и  $\oslash$  — покомпонентное умножение и деление матриц.

Илья Ирхин. Реализация ARTM: [https://github.com/ilirhin/python\\_artm](https://github.com/ilirhin/python_artm)

M. Shashanka et al. Probabilistic latent variable models as nonnegative factorizations. 2008.

## Улучшение сходимости

В формулах M-шага вместо  $\phi_{wt}$  и  $\theta_{td}$  можно подставлять несмещённые частотные оценки (PLSA)  $\hat{\phi}_{wt} = \frac{n_{wt}}{n_t}$  и  $\hat{\theta}_{td} = \frac{n_{td}}{n_d}$ :

$$\phi_{wt} = \operatorname{norm}_{w \in W} \left( n_{wt} + \hat{\phi}_{wt} \frac{\partial R(\hat{\Phi}, \hat{\Theta})}{\partial \phi_{wt}} \right)$$
$$\theta_{td} = \operatorname{norm}_{t \in T} \left( n_{td} + \hat{\theta}_{td} \frac{\partial R(\hat{\Phi}, \hat{\Theta})}{\partial \theta_{td}} \right)$$

**Доказано**, что в результате такой модификации

- увеличивается значение регуляризованного правдоподобия
- монотонный рост регуляризованного правдоподобия начинается быстрее — как правило, со второй итерации
- чем больше  $\tau$ , тем заметнее улучшение сходимости
- не требуется дополнительных затрат времени или памяти

---

И.А.Ирхин, К.В.Воронцов. Сходимость алгоритма аддитивной регуляризации тематических моделей. 2020.

## Включение и отключение регуляризаторов

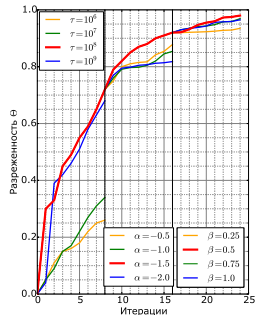
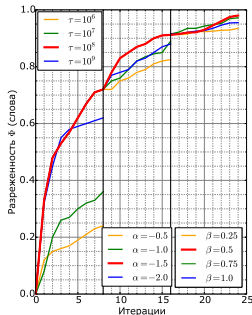
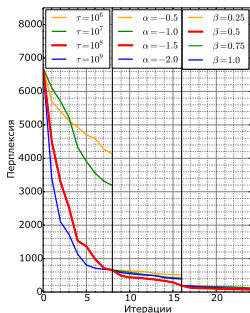
1. Регуляризация ведёт итерационный процесс к матричному разложению с требуемыми свойствами, но даёт смещённые оценки матриц  $\Phi, \Theta$ . По окончании процесса можно возвращать несмещённые PLSA-оценки:

$$\phi_{wt} = \operatorname{norm}_{w \in W}(n_{wt})$$
$$\theta_{td} = \operatorname{norm}_{t \in T}(n_{td})$$

2. Коэффициенты регуляризации можно менять в итерациях.
3. Регуляризаторы можно включать не сразу или по очереди.
4. Регуляризаторы можно отключать по достижению эффекта.
5. Одни регуляризаторы могут выполнять подготовительную работу для применения следующих регуляризаторов.

## Управление траекторией регуляризации

- 1 задать диапазон и сетку значений каждого  $\tau_i$   
(удобно использовать относительные коэффициенты  $\tilde{\tau}_i$ )
- 2 задать последовательность подключения регуляризаторов  
(имеются эмпирические рекомендации)
- 3 визуализировать несколько критериев качества (спойлер):



## Относительные коэффициенты регуляризации

Формула M-шага со взвешенной суммой регуляризаторов  $R_i$ :

$$\phi_{wt} = \operatorname{norm}_{w \in W} \left( n_{wt} + \sum_{i=1}^k \tau_i \phi_{wt} \frac{\partial R_i}{\partial \phi_{wt}} \right).$$

Суммарное воздействие  $r_{it}$  регуляризатора  $R_i$  на тему  $t$  и суммарное воздействие  $r_i$  регуляризатора  $R_i$  на все темы:

$$r_{it} = \sum_{w \in W} \left| \phi_{wt} \frac{\partial R_i}{\partial \phi_{wt}} \right|, \quad r_i = \sum_{t \in T} r_{it}.$$

Относительный коэффициент регуляризации  $\tilde{\tau}_i$ :

$$\tau_i = \tilde{\tau}_i \frac{n}{r_i} \quad \text{или} \quad \tau_i = \tilde{\tau}_i \left( \gamma_i \frac{n_t}{r_{it}} + (1 - \gamma_i) \frac{n}{r_i} \right),$$

где  $\gamma_i$  — индивидуализация воздействия  $R_i$  на темы.



## BigARTM: библиотека тематического моделирования

### Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Пакетный онлайнный параллельный ARTM
- Встроенная библиотека регуляризаторов и мер качества

### Сообщество:

- Открытый код <https://github.com/bigartm>  
(discussion group, issue tracker, pull requests)
- Документация <http://bigartm.org>



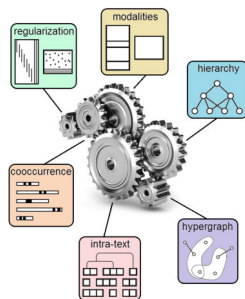
### Лицензия и среда разработки:

- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Windows, Linux, MacOS (32/64 bit)
- Интерфейсы API: command-line, C++, and Python

## Ключевые возможности библиотек BigARTM и TopicNet

### BigARTM

- библиотека регуляризаторов
- мультимодальные модели
- иерархические модели
- гиперграфовые модели
- модели связности текста



### TopicNet

- Перебор сценариев регуляризации для выбора моделей
- Автоматическое протоколирование экспериментов
- Построение «банка тем» из множества моделей
- Визуализация результатов тематического моделирования

*V. Bulatov, E. Egorov, E. Veselova, D. Polyudova, V. Alekseev, A. Goncharov, K. Vorontsov.*  
TopicNet: making additive regularisation for topic modelling accessible. LREC-2020

## Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

3.7М статей Википедии, 100К слов:

время min (перплексия)

проц.	$ T $	Gensim	Vowpal Wabbit	BigARTM	BigARTM асинхрон
1	50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
1	100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
1	200	637m (3241)	154m (3960)	83m (3347)	53m (3362)
2	50	89m (5056)		22m (5092)	13m (5160)
2	100	143m (4012)		29m (4107)	19m (4144)
2	200	325m (3297)		47m (3347)	28m (3380)
4	50	88m (5311)		12m (5216)	7m (5353)
4	100	104m (4338)		16m (4233)	10m (4357)
4	200	315m (3583)		26m (3520)	16m (3634)
8	50	88m (6344)		8m (5648)	5m (6220)
8	100	107m (5380)		10m (4660)	6m (5119)
8	200	288m (4263)		15m (3929)	10m (4309)

*D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov.*

Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.

## Правдоподобие и перплексия (perplexity)

*Правдоподобие* языковой модели  $p(w|d)$  (чем выше, тем лучше):

$$\mathcal{L}(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d), \quad p(w|d) = \sum_t \phi_{wt} \theta_{td}$$

*Перплексия* языковой модели  $p(w|d)$  (чем меньше, тем лучше):

$$\mathcal{P}(D) = \exp\left(-\frac{1}{n} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d)\right), \quad n = \sum_{d \in D} \sum_{w \in d} n_{dw}$$

### Интерпретация перплексии:

- если распределение  $p(w|d) = \frac{1}{|W|}$  равномерное, то  $\mathcal{P} = |W|$
- мера «удивлённости» модели словам текста
- коэффициент ветвления (branching factor) текста
- известные оценки человеческой перплексии: 8–12

## Измерение интерпретируемости тем

Тема интерпретируемая, если по топовым словам темы эксперт может определить, о чём эта тема, и дать ей название.

- *Экспертные оценки:*
  - интерпретируемость темы по балльной шкале;
  - каждую тему оценивают несколько экспертов.
- *Метод интрузий (intrusion):*
  - в список топовых слов внедряется лишнее слово;
  - измеряется доля ошибок экспертов при его определении

**Задача:** найти внутренний критерий интерпретируемости, наиболее коррелирующий с экспертными оценками

**Решение:** *когерентность* (согласованность) тем (topic coherence)

---

*Newman D., Lau J.H., Grieser K., Baldwin T.* Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

## Эксперимент. Связь когерентности и интерпретируемости

Измерялась ранговая корреляция Спирмена каждой из 15 метрик и экспертными оценками интерпретируемости.

PMI — лучшая метрика.

Gold-standard — средняя корреляция Спирмена между оценками разных экспертов.

Resource	Method	Median	Mean
WordNet	HSO	0.15	0.59
	JCN	-0.20	0.19
	LCH	-0.31	-0.15
	LESK	0.53	0.53
	LIN	0.09	0.28
	PATH	0.29	0.12
	RES	0.57	0.66
	VECTOR	-0.08	0.27
	WuP	0.41	0.26
	Wikipedia	RACO	0.62
MiW		0.68	0.70
DOC SIM		0.59	0.60
<b>PMI</b>		<b>0.74</b>	<b>0.77</b>
Google	TITLES	0.51	
	LOGHITS	-0.19	
Gold-standard	IAA	<b>0.82</b>	<b>0.78</b>

**Вывод:** когерентность близка к «золотому стандарту».

*Newman D., Lau J.H., Grieser K., Baldwin T.* Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

## Когерентность как внутренний критерий интерпретируемости

*Когерентность (согласованность) темы  $t$  по  $k$  топовым словам:*

$$\text{coh}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^k \text{PMI}(w_i, w_j)$$

где  $w_i$  —  $i$ -е слово в порядке убывания  $\phi_{wt}$ ,

$\text{PMI}(u, v) = \ln \frac{P_{uv}}{P_u P_v}$  — *поточечная взаимная информация*  
(pointwise mutual information),

$P_{uv}$  — доля документов, в которых слова  $u, v$  хотя бы один раз встречаются рядом (в одном предложении или в окне 10 слов),

$P_u$  — доля документов, в которых  $u$  встретился хотя бы 1 раз,  
 $P_{uv}, P_u$  можно вычислять по другой коллекции (Википедии).

*Когерентность модели* = средняя когерентность всех тем.

---

*Newman D., Lau J.H., Grieser K., Baldwin T.* Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

## Лексическое ядро, чистота и контрастность темы

Лексическое ядро  $W_t$  темы  $t$ , варианты определения:

- $W_t$  — top- $k$  термов с наибольшими значениями  $p(w|t)$
- $W_t = \{w : p(w|t) > p(w)\}$
- $W_t = \{w : p(w|t) > \frac{1}{|W|}\}$  [Кольцов и др., 2014]
- $W_t = \{w : p(t|w) > 0.25\}$  [Воронцов, Потапенко, 2014]

Характеристики лексического ядра темы:

- $|W_t|$  — размер ядра темы, ориентировочно  $|W_t| \sim \frac{|W|}{|T|}$
- $\sum_{w \in W_t} p(w|t)$  — чистота темы, из  $[0, 1]$ , лучше больше
- $\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$  — контрастность темы,  $[0, 1]$ , лучше больше
- $\frac{1}{|W_t|} \sum_{w \in W_t} \log \frac{p(w|t)}{p(w)}$  — logLift, лучше больше [Taddy, 2012]

---

Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. AIST, 2014.



## Разреживание, сглаживание, декоррелирование, отбор тем

M-шаг при комбинировании 6 регуляризаторов:

$$\phi_{wt} = \text{norm}_w \left( n_{wt} + \tau_1 \underbrace{\beta_w[t \in B]}_{\substack{\text{сглаживание} \\ \text{фоновых} \\ \text{тем}}} - \tau_2 \underbrace{\beta_w[t \in S]}_{\substack{\text{разреживание} \\ \text{предметных} \\ \text{тем}}} - \tau_3 \underbrace{\phi_{wt} \sum_{s \in S \setminus t} \phi_{ws}}_{\text{декоррелирование}} \right)$$

$$\theta_{td} = \text{norm}_t \left( n_{td} + \tau_4 \underbrace{\alpha_t[t \in B]}_{\substack{\text{сглаживание} \\ \text{фоновых} \\ \text{тем}}} - \tau_5 \underbrace{\alpha_t[t \in S]}_{\substack{\text{разреживание} \\ \text{предметных} \\ \text{тем}}} - \tau_6 \underbrace{\frac{n_d}{n_t} \theta_{td}}_{\text{удаление} \\ \text{малых тем}} \right)$$

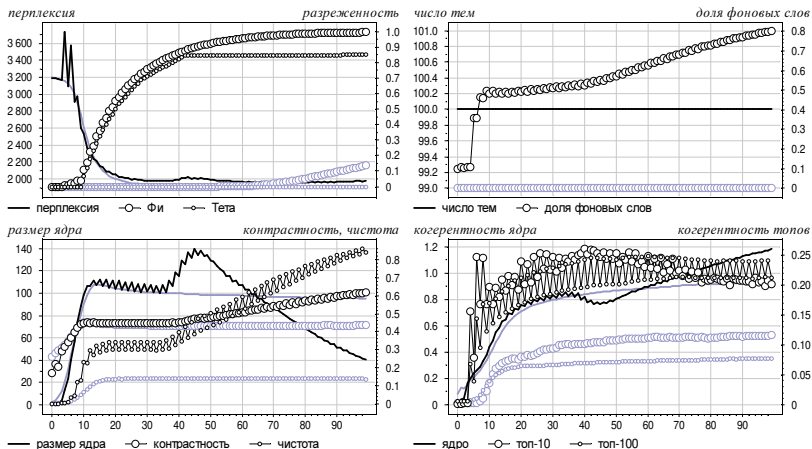
**Данные:** статьи NIPS (Neural Information Processing System)  
 $|D| = 1566$  статей,  $n = 2.3$  М,  $|W| = 13$  К,  
 контрольная коллекция:  $|D'| = 174$ .

---

Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. AIST'2014.

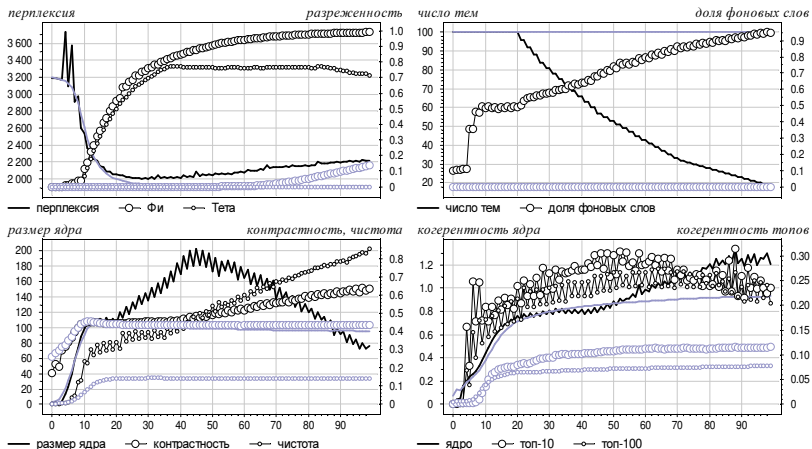
## Разреживание, сглаживание, декоррелирование

Зависимости критериев качества от итераций EM-алгоритма  
 (серый — PLSA, чёрный — ARTM)



## Те же регуляризаторы, плюс отбор тем

Зависимости критериев качества от итераций EM-алгоритма  
(серый — PLSA, чёрный — ARTM)



## Выводы по результатам экспериментов

**Одновременное улучшение многих критериев качества при незначительной деградации перплексии (правдоподобия):**

- *разреженность* выросла от 0 до 95%–98%
- *когерентность тем* выросла от 0.1 до 0.3
- *чистота тем* выросла от 0.15 до 0.8
- *контрастность тем* выросла от 0.4 до 0.6

**Рекомендации по выбору траектории регуляризации:**

- разреживание включать постепенно после 10-20 итераций
- сглаживание включать сразу
- декоррелирование включать сразу и как можно сильнее
- отбор тем включать постепенно,
- не совмещая с декоррелированием на одной итерации

- Регуляризация — стандартный приём для решения некорректно поставленных задач
- ARTM позволяет комбинировать регуляризаторы и строить тематические модели с требуемыми свойствами
- Онлайнный EM-алгоритм способен обрабатывать большую коллекцию за один проход
- BigARTM — эффективная реализация ARTM
- TopicNet — обёртка над BigARTM для экспериментов
- Сглаживание + разреживание + декоррелирование — часто используемая комбинация регуляризаторов
- Декоррелятор помогает при несбалансированности тем
- Другие регуляризаторы — в следующих лекциях

Упражнения на принцип максимума правдоподобия:

1. Униграммная модель документов:  $p(w|d) = \xi_{dw}$

Найти параметры модели  $\xi_{dw}$ .

2. Униграммная модель коллекции:  $p(w|d) = \xi_w$  для всех  $d$

Найти параметры модели  $\xi_w$ .

Подсказка: применить условия ККТ или основную лемму.

3. (более творческое задание)

Предложите модель, определяющую роли слов в текстах:

- тематические слова
- специфичные слова документа (шум)
- слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов  $p(r|w)$ ,  $r \in \{\text{т, ш, ф}\}$ .

Подсказка 2: можно разреживать  $p(r|w)$  для жёсткого определения ролей.

Подсказка 3: можно использовать документную частоту слов.

4. Заменяем  $\log$  другой монотонно возрастающей функцией  $\mu$ :

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \mu \left( \sum_{t \in T} \phi_{wt} \theta_{td} \right) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

Как изменится EM-алгоритм? Возможно ли подобрать функцию  $\mu$  так, чтобы сократился объём вычислений?

5. Заменяем  $\log$  монотонно возрастающей функцией  $\mu$  в регуляризаторе сглаживания–разреживания (модель LDA):

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_w \mu(\phi_{wt}) + \sum_{d \in D} \sum_{t \in T} \alpha_t \mu(\theta_{td}).$$

Как изменится M-шаг и воздействие регуляризатора на модель?

6. Какому регуляризатору соответствует формула M-шага

$$\phi_{wt} = \text{norm}_w(n_{wt} [n_{wt} > \gamma n_t])$$

Аналитик построил тематическую модель  $\Phi^0, \Theta^0$  и отметил среди столбцов матрицы  $\Phi^0$  темы двух типов: удачные  $T_+ \subset T$  и неудачные  $T_- \subset T$ .

Теперь он хочет построить модель ещё раз так, чтобы

- удачные темы остались в матрице  $\Phi$ ;
- остальные темы построились по-другому и были не похожи на каждую из неудачных тем  $t \in T_-$ .

7. Предложите регуляризаторы для этого.

8. Не получится ли так, что новые темы будут отдаляться от суммы неудачных тем  $\sum_{t \in T_-} \phi_{wt}^0$  вместо того, чтобы отдаляться от каждой из неудачных тем по отдельности? Почему это плохо и как этого избежать?

9. Предложите способ инициализации  $\Phi$  для новой модели.