Вероятностные тематические модели
Лекция 3.
Особенности EM-алгоритма и комбинирование регуляризаторов

Константин Вячеславович Воронцов
k.v.vorontsov@phystech.edu

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

МФТИ.ФПМИ.ИС.ИАД — ФИЦ ИУ РАН • 2025-10-02

Содержание

1. **Часто используемые регуляризаторы**
   - Сглаживание и разреживание
   - Декоррелирование
   - Разреживающий регуляризатор для отбора тем
2. **Особенности реализации EM-алгоритма в ARTM**
   - Подходы к ускорению EM-алгоритма
   - Комбинирование регуляризаторов
   - Библиотеки BigARTM и TopicNet
3. **Эксперименты с тематическими моделями**
   - Измерение качества тематических моделей
   - Эксперименты с комбинированием регуляризаторов
   - Эксперименты с оптимизацией числа тем

Напоминание. Задача тематического моделирования

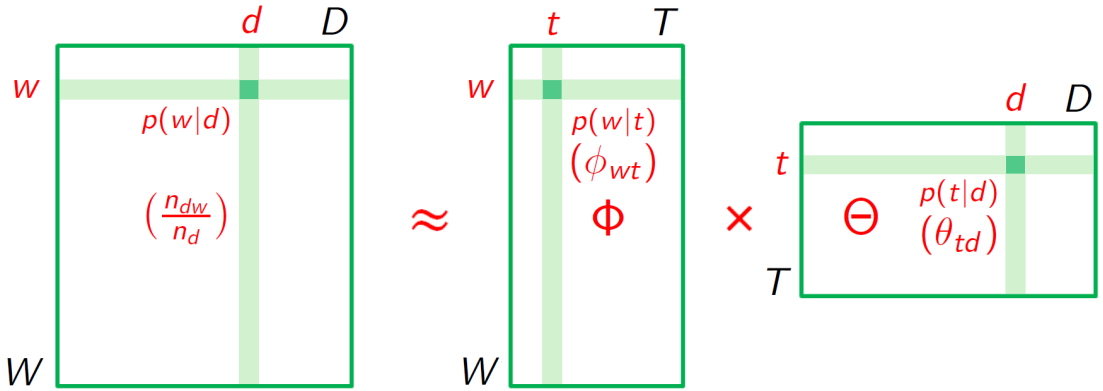
Дано: коллекция текстовых документов, $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

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

Критерий: максимум log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} \equiv p(t|d,w) = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}} \right), & n_{wt} = \sum\limits_{d\in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}} \right), & n_{td} = \sum\limits_{w\in d} n_{dw} p_{tdw} \end{cases}$$

где $\underset{t\in T}{\text{norm}}(x_t) = \dfrac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ — операция нормирования вектора.

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

Напоминание. Модель локального контекста (Attentive ARTM)

Дано: коллекция текстовых документов, $w_1, \ldots, w_n$

C_i — локальный контекст (окружение) терма w_i

α_{ui} — распределение весов термов u в контексте C_i

Найти: параметры ϕ_{wt}, p_t тематической модели

$$p(w|C_i) = \sum_{t \in T} \phi_{wt}\theta_{ti}, \quad \theta_{ti} = \sum_{u \in C_i} \alpha_{ui}\phi'_{tw}, \quad \phi'_{tw} = \underset{t \in T}{\operatorname{norm}}\left(\phi_{wt}p_t\right)$$

Критерий: максимум log правдоподобия с регуляризатором R:

$$\sum_{i=1}^{n} \ln p(w|C_i) + R(\Phi) \to \max_{\Phi}, \quad R(\Phi) = \sum_{i=1}^{k} \tau_i R_i(\Phi)$$

EM-алгоритм: $p_{ti} = \underset{t \in T}{\operatorname{norm}}\left(\phi_{w_i t}\theta_{ti}\right), \quad p_t = \frac{1}{n}\sum_{i=1}^{n} p_{ti}$

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(\sum_{i=1}^{n} [w_i = w] p_{ti} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)$$

Дивергенция Кульбака–Лейблера и её свойства

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

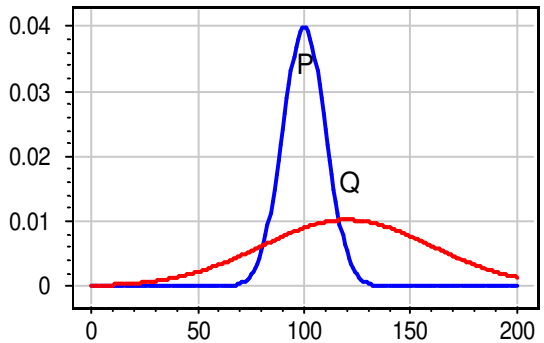
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

1. $\mathrm{KL}(P\|Q) \geqslant 0$; $\quad \mathrm{KL}(P\|Q) = 0 \;\Leftrightarrow\; P = Q$;

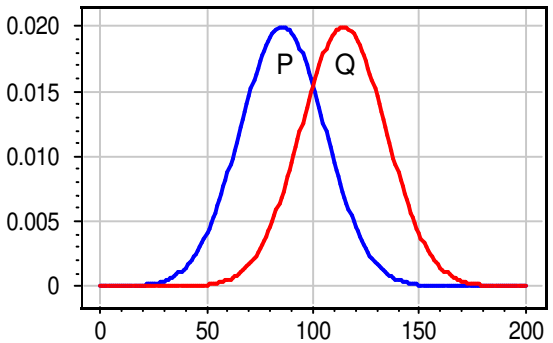
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

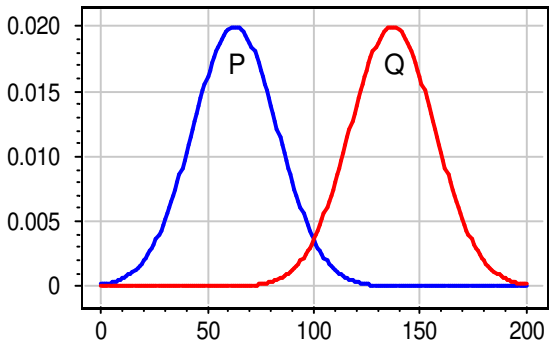
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



$\mathrm{KL}(P\|Q) = 0.44$
$\mathrm{KL}(Q\|P) = 2.97$



$\mathrm{KL}(P\|Q) = 0.44$
$\mathrm{KL}(Q\|P) = 0.44$



$\mathrm{KL}(P\|Q) = 2.97$
$\mathrm{KL}(Q\|P) = 2.97$

Регуляризатор сглаживания

Гипотеза сглаженности:
распределения ϕ_{wt} близки к заданному распределению β_w;
распределения θ_{td} близки к заданному распределению α_t.

$$\sum_{t\in T} \mathrm{KL}(\beta_w \| \phi_{wt}) \to \min_{\Phi}; \qquad \sum_{d\in D} \mathrm{KL}(\alpha_t \| \theta_{td}) \to \min_{\Theta}.$$

Максимизируем сумму регуляризаторов:

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

Подставляем, получаем формулы M-шага, похожие на LDA

$$\phi_{wt} = \underset{w\in W}{\mathrm{norm}}(n_{wt} + \beta_0\beta_w), \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}(n_{td} + \alpha_0\alpha_t),$$

однако в LDA есть ограничения $\beta_0\beta_w > -1$, $\alpha_0\alpha_t > -1$

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

Регуляризатор разреживания

Гипотеза разреженности: среди ϕ_{wt}, θ_{td} много нулей;
распределения ϕ_{wt} далеки от заданного распределения β_w;
распределения θ_{td} далеки от заданного распределения α_t.

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

Максимизируем сумму регуляризаторов:

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

Это обобщение LDA, снимающее ограничения на α_t, β_w:

$$\phi_{wt} = \underset{w\in W}{\mathrm{norm}}\big(n_{wt} - \beta_0\beta_w\big), \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}\big(n_{td} - \alpha_0\alpha_t\big).$$

Varadarajan J., Emonet R., Odobez 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:

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

где $\beta_0 > 0$, $\alpha_0 > 0$ — коэффициенты регуляризации,
β_{wt}, α_{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 термами темы

Частичное обучение (semi-supervised learning)

Общий вид регуляризаторов сглаживания и разреживания:

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

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

Разреживание по «чёрным спискам»:

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

Сглаживание по «белым спискам» (seed words):

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

Есть ли проблема $\ln 0$ при разреживании распределений?

В регуляризаторе сглаживания/разреживания

$$R(\Phi) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} \to \max$$

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

Подправим регуляризатор, при сколь угодно малом ε:

$$R(\Phi) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln(\phi_{wt} + \varepsilon) \to \max .$$

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

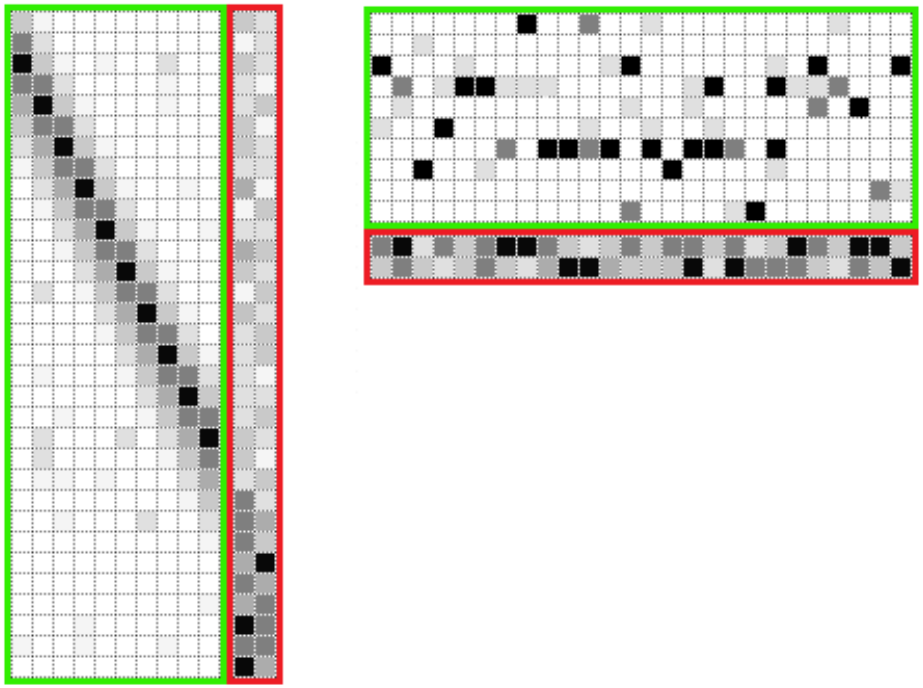
$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}} \left(n_{wt} + \beta_0 \beta_{wt} \frac{\phi_{wt}}{\phi_{wt} + \varepsilon} \right) \xrightarrow{\varepsilon \to 0} \underset{w \in W}{\operatorname{norm}} \left(n_{wt} + \beta_0 \beta_{wt} [\phi_{wt} \neq 0] \right),$$

Если $\phi_{wt} = 0$, то и на последующих итерациях $n_{wt} = \phi_{wt} = 0$.

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

$\Phi_{W \times T}$ $\Theta_{T \times D}$



Регуляризатор декоррелирования тем

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

Минимизируем ковариации между вектор-столбцами ϕ_t:

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

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

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

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation. 2010.

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:

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

Подставляем, получаем:

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

Эффект: обнуляются строки матрицы Θ 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.

Модификация M-шага, улучшающая сходимость

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

$$\phi_{wt} = \underset{w \in W}{\text{norm}} \left(n_{wt} + \hat{\phi}_{wt} \frac{\partial R(\hat{\Phi}, \hat{\Theta})}{\partial \phi_{wt}} \right)$$

$$\theta_{td} = \underset{t \in T}{\text{norm}} \left(n_{td} + \hat{\theta}_{td} \frac{\partial R(\hat{\Phi}, \hat{\Theta})}{\partial \theta_{td}} \right)$$

Доказано, что в результате такой модификации

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

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

Замена логарифма в функции потерь и модификация E-шага

Задача максимизации (при тех же ограничениях):

$$\sum_{d\in D}\sum_{w\in d} n_{dw}\,\ell\Big(\sum_{t\in T}\phi_{wt}\theta_{td}\Big) + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:}\quad p_{tdw} = \phi_{wt}\theta_{td}\,\ell'\Big(\sum_s \phi_{ws}\theta_{sd}\Big)$$

$$\text{M-шаг:}\quad \begin{cases} \phi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\Big), \quad n_{wt} = \sum\limits_{d\in D} n_{dw}p_{tdw} \\ \theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\Big), \quad n_{td} = \sum\limits_{w\in d} n_{dw}p_{tdw} \end{cases}$$

Только при $\ell(z) = \ln z$ на E-шаге возникает формула Байеса
При $\ell(z) = z$ уходит нормировка на E-шаге,
на начальных итерациях нормировку можно не делать

M.A.Apishev, K.V.Vorontsov. Learning Topic Models with Arbitrary Loss. 2020

Матричная реализация EM-алгоритма

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

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

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

Матричная запись (norm — нормировка по столбцам):

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

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

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

Включение и отключение регуляризаторов

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

$$\phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt}\right)$$
$$\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td}\right)$$

2. Коэффициенты регуляризации можно менять в итерациях.

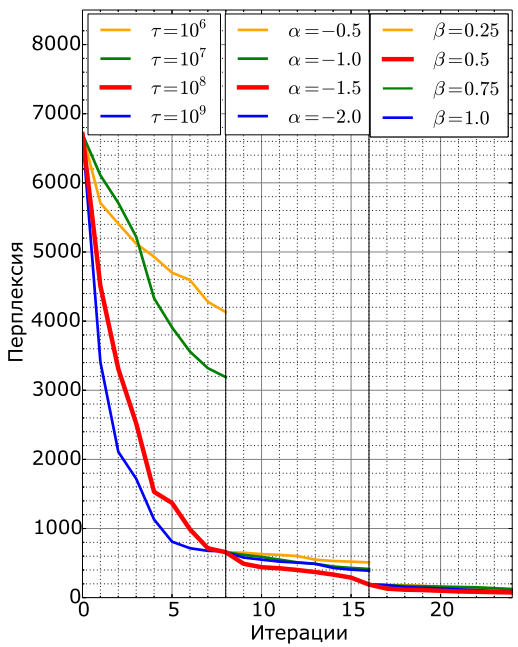
3. Регуляризаторы можно включать не сразу или по очереди.

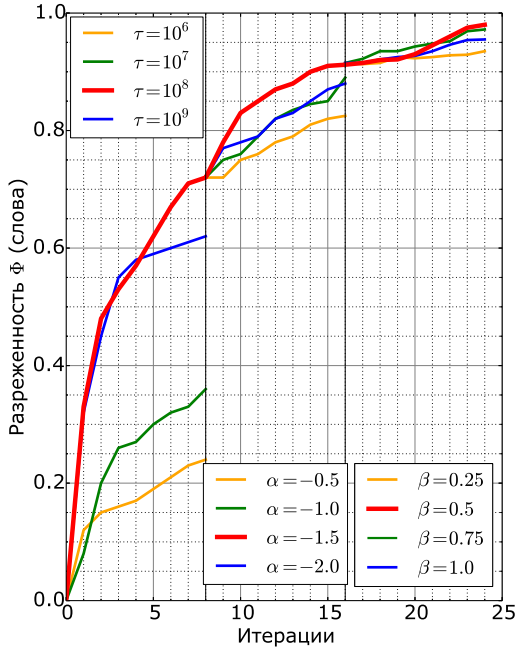
4. Регуляризаторы можно отключать по достижению эффекта.

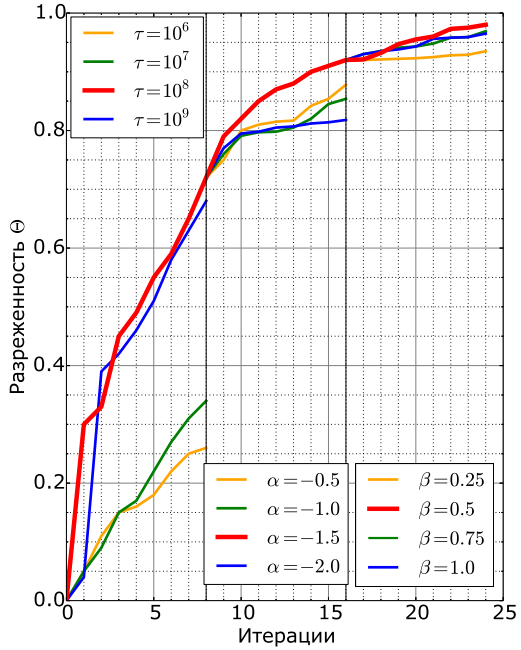
5. Одни регуляризаторы могут выполнять подготовительную работу для применения следующих регуляризаторов.

Управление траекторией регуляризации

1. задать диапазон и сетку значений каждого τ_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:

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

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

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

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 γ_i — *индивидуализация* воздействия R_i на темы.

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Пакетный онлайновый 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

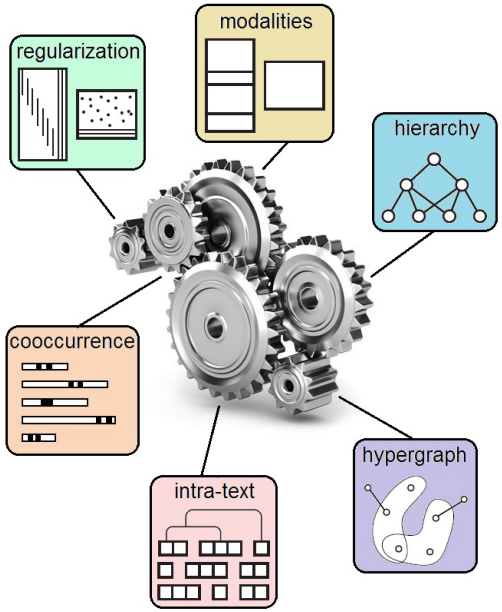
Лицензия и среда разработки:

- 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

3.7M статей Википедии, 100K 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):

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

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

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

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	0.69
	MIW	0.68	0.70
	DOCSIM	0.59	0.60
	PMI	**0.74**	**0.77**
Google	TITLES	0.51	
	LOGHITS	−0.19	
Gold-standard	IAA	0.82	0.78

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 ϕ_{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 $p(w|t)$
- $W_t = \{w\colon p(w|t) > p(w)\}$
- $W_t = \{w\colon p(w|t) > \frac{1}{|W|}\}$ [Кольцов и др., 2014]
- $W_t = \{w\colon p(t|w) > 0.25\}$ [Воронцов, Потапенко, 2014]

Характеристики лексического ядра темы:

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

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:

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

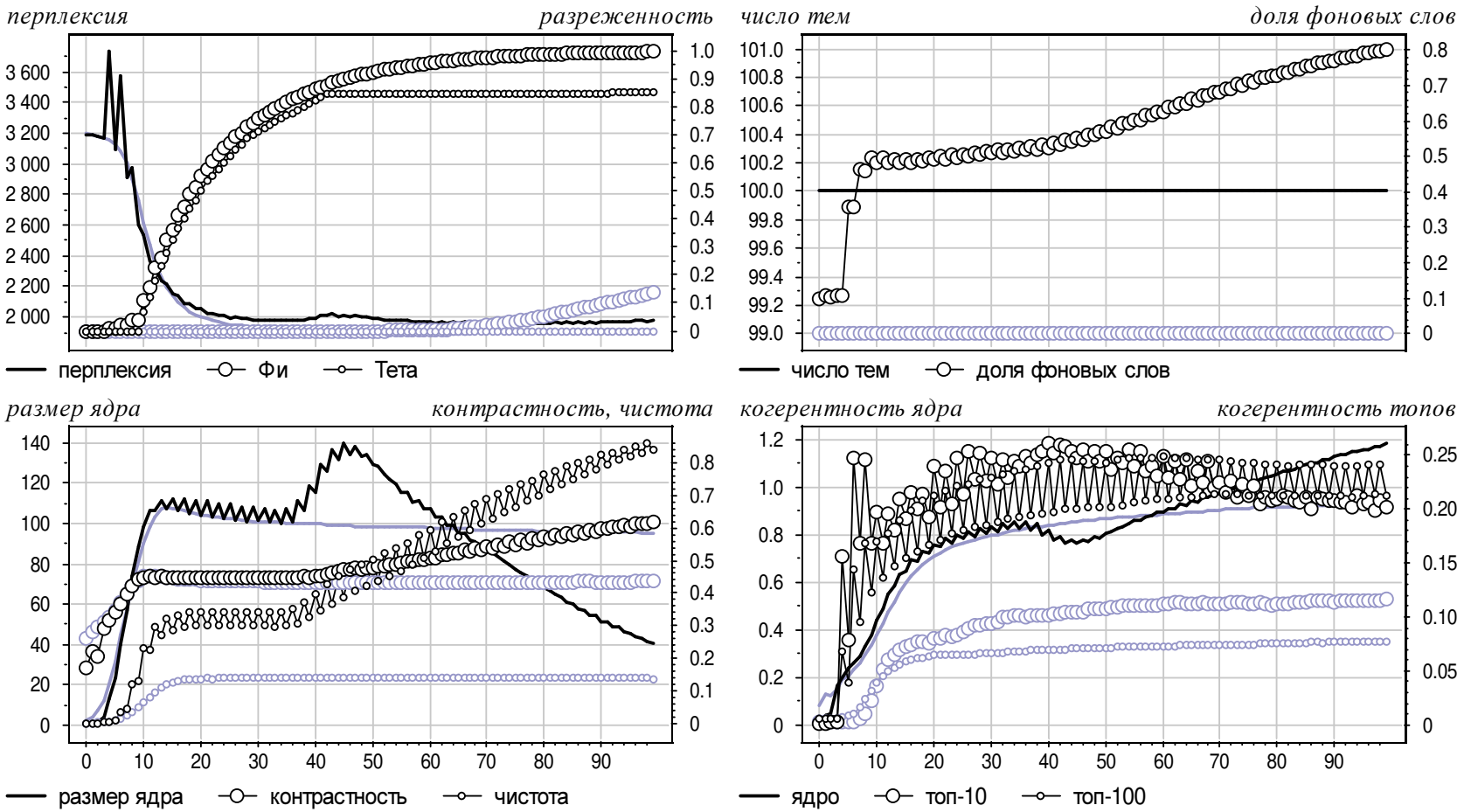
$$\theta_{td} = \underset{t}{\text{norm}}\Big(n_{td} + \tau_4 \underbrace{\alpha_t[t \in B]}_{\substack{\text{сглаживание}\\ \text{фоновых}\\ \text{тем}}} - \tau_5 \underbrace{\alpha_t[t \in S]}_{\substack{\text{разреживание}\\ \text{предметных}\\ \text{тем}}} - \tau_6 \underbrace{\frac{n_d}{n_t}\theta_{td}}_{\substack{\text{удаление}\\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

Данные: статьи NIPS (Neural Information Processing System)
$|D| = 1566$ статей, $n = 2.3\,\text{M}$, $|W| = 13\,\text{K}$,
контрольная коллекция: $|D'| = 174$.

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

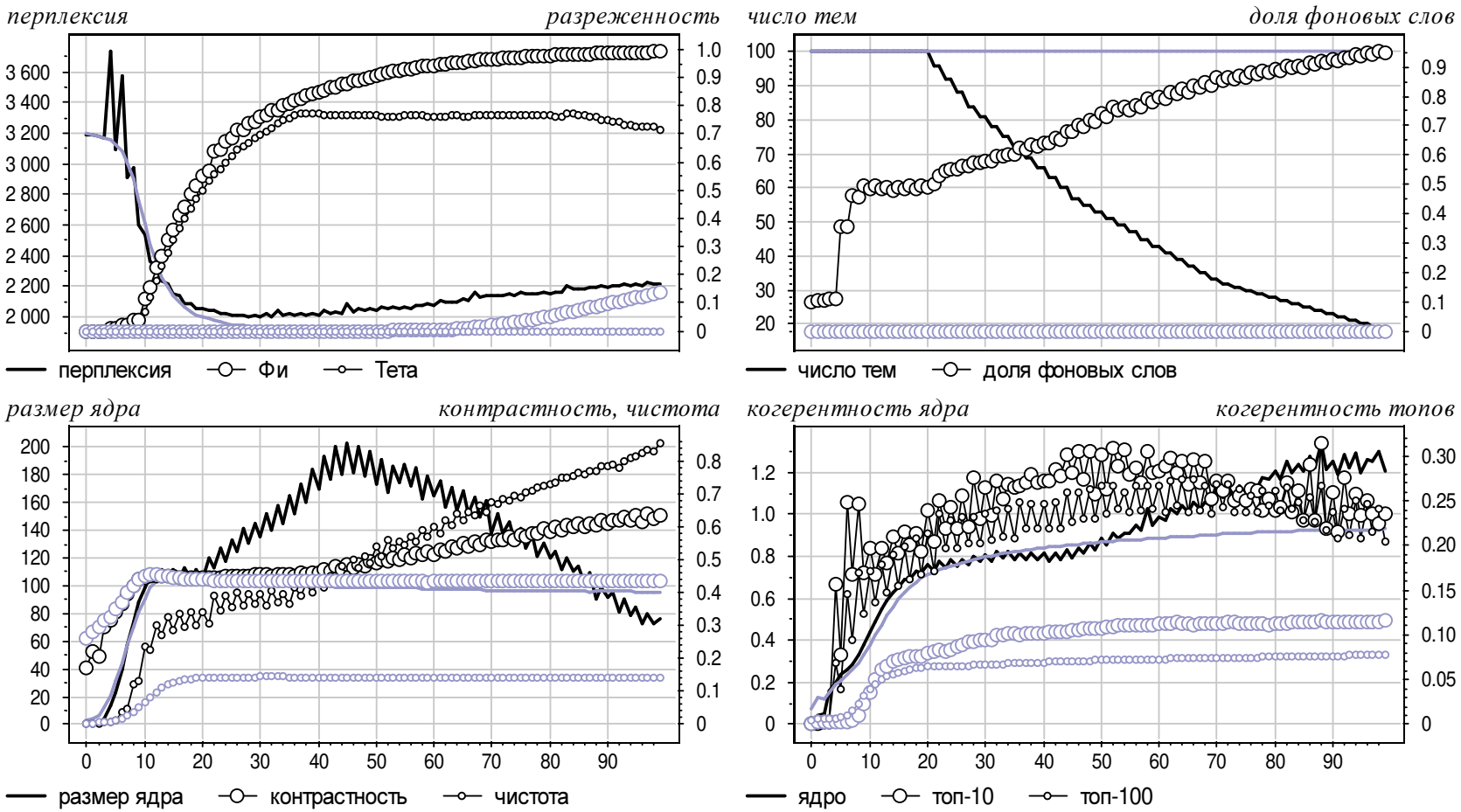
Разреживание, сглаживание, декоррелирование

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

Эксперименты с отбором тем на синтетических данных

Коллекция статей NIPS (Neural Information Processing System)

- $|D| = 1566$ обучающих документов; $|D'| = 174$ тестовых
- $|W| = 13\,\text{K}$ — мощность словаря

Синтетическая коллекция:

- строим PLSA за 500 итераций, $|T_0| = 50$ тем на NIPS
- генерируем коллекцию (n^0_{dw}) из полученных Φ и Θ:

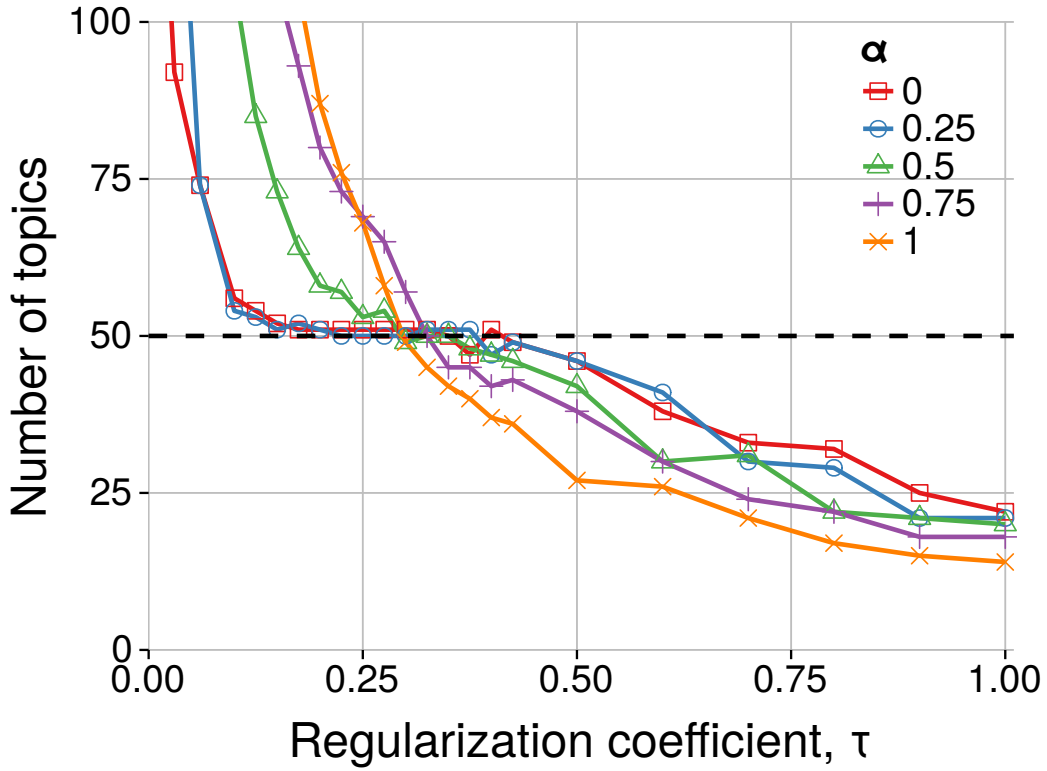
$$n^0_{dw} = n_d \sum_{t \in T_0} \phi_{wt}\theta_{td}$$

Параметрическое семейство полусинтетических данных:

- n^α_{dw} — смесь синтетических данных n^0_{dw} и реальных n_{dw}:

$$n^\alpha_{dw} = \alpha n_{dw} + (1-\alpha) n^0_{dw}$$

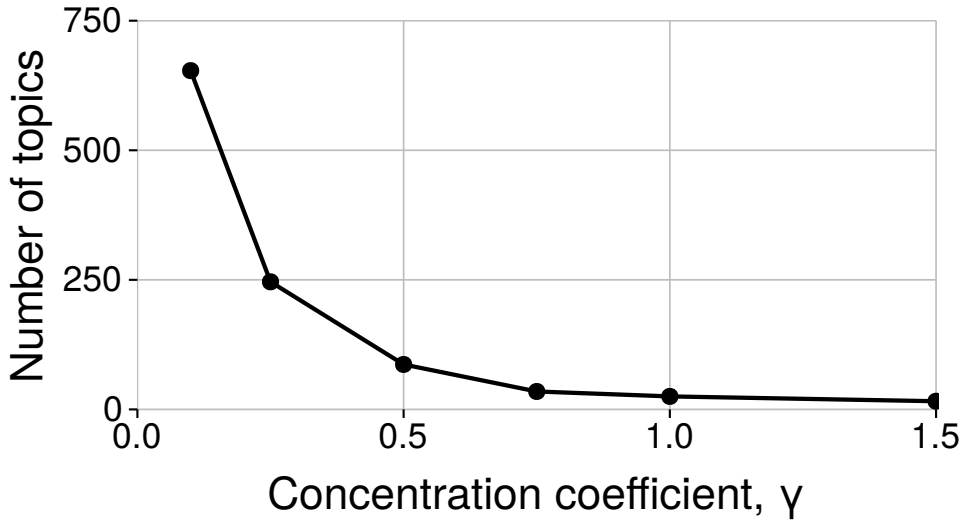
Попытка определения числа тем



- на синтетических данных надёжно находим $|T| = 50$
- причём в широком интервале значений коэффициента τ
- однако на реальных данных чёткого интервала нет

Сравнение с байесовской тематической моделью HDP

HDP, Hierarchical Dirichlet Process [Tech et.al, 2006] — «state-of-the-art» байесовский подход к определению числа тем

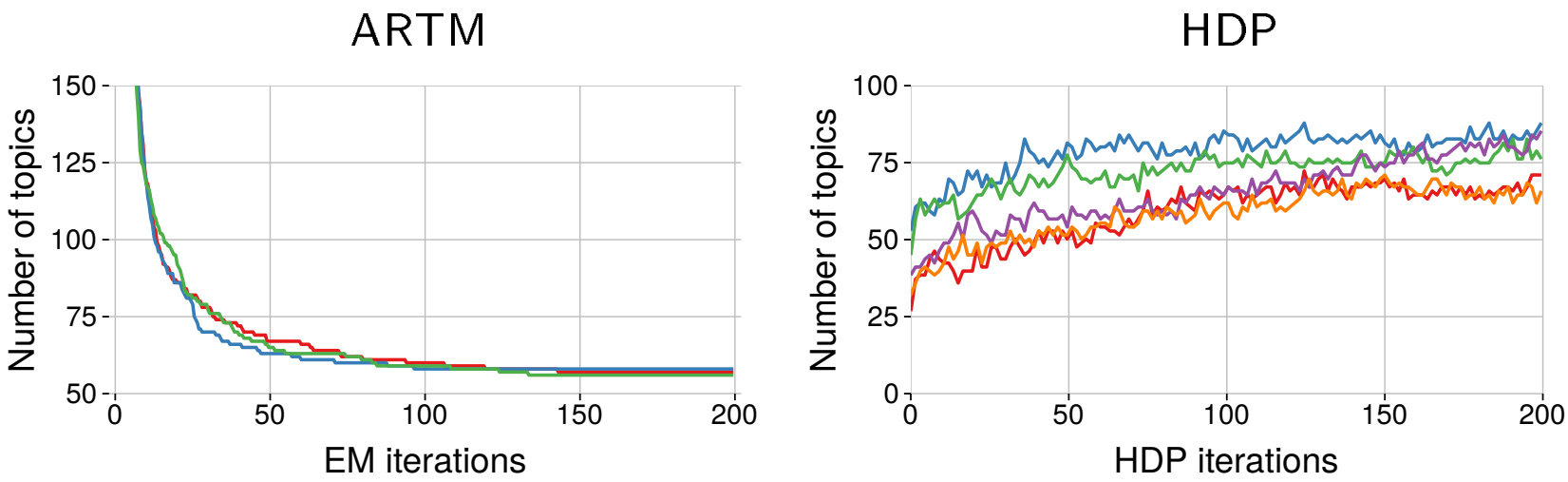


- *Коэффициент концентрации* γ в HDP влияет на $|T|$ так же сильно, как выбор коэффициента τ в ARTM.

Y.W.Teh, M.Jordan, M.Beal, D.Blei. Hierarchical Dirichlet Processes. 2005.

Сравнение ARTM и HDP по устойчивости

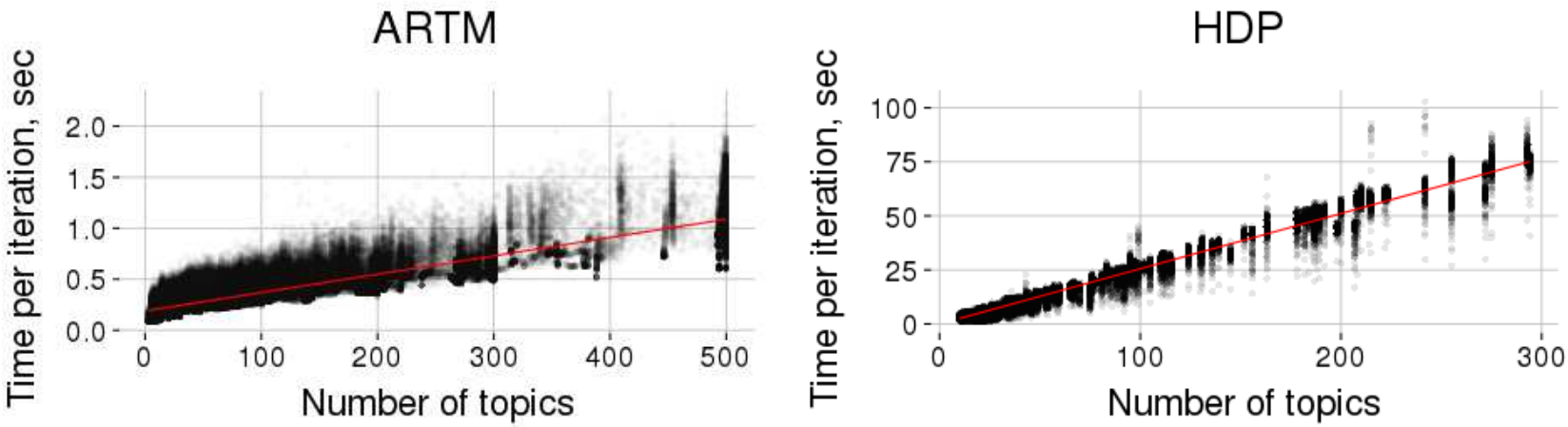
Запуск ARTM и HDP много раз из случайных инициализаций:



- HDP менее устойчив, причём в двух смыслах:
  - число тем сильнее флуктуирует от итерации к итерации;
  - результаты нескольких запусков различаются сильнее.
- «Рекомендуемые» значения параметров γ в HDP и τ в ARTM дают примерно равное число тем $|T| \approx 60$

Сравнение ARTM и HDP по времени вычислений

Сравнение времени одного прохода коллекции (sec)



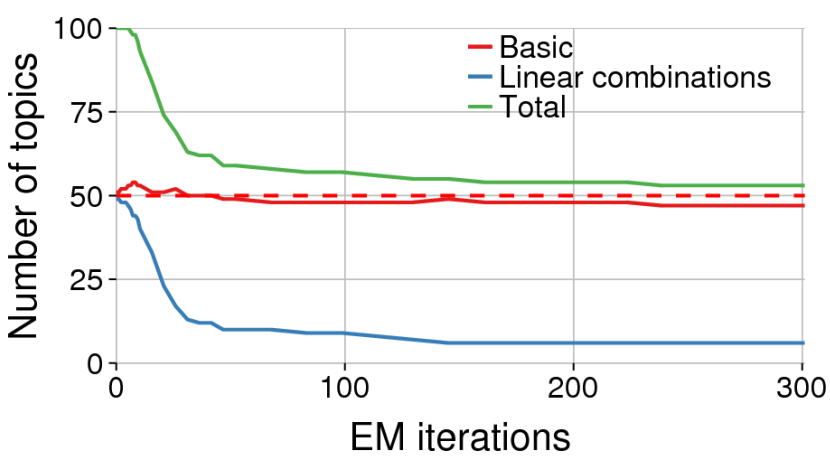
- ARTM в 100 раз быстрее!

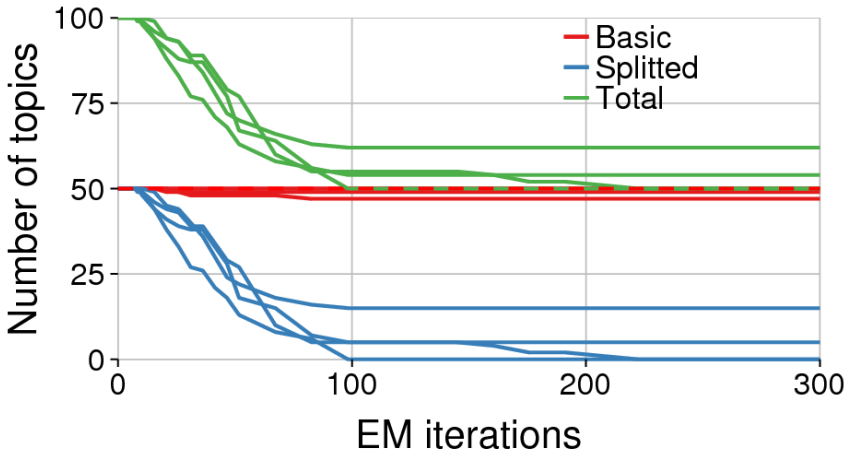
Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.

Александр Плавин. Отбор тем в задачах тематического моделирования. ВКР бакалавра, МФТИ. 2015.

Удаление линейно зависимых и расщеплённых тем

Добавили 50 линейных комбинаций тем в модельную Ф.
Расщепили 50 тем, каждую на две подтемы в модельной Ф.





- Удаляются линейно зависимые и расщеплённые темы
- Остаются наиболее различные темы исходной модели.

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.

Выводы по результатам экспериментов

- Регуляризатор отбора тем удаляет незначимые темы и определяет оптимальное число тем, если оно существует
- Увы, в реальных данных его не существует!
  Оно задаётся исходя из целей моделирования.
- Значит, надо иерархически дробить темы на подтемы, и пусть пользователь выбирает нужную ему детализацию
- Есть простой метод для удаления лишних тем, но как обнаруживать новые темы в потоке или в батчах и добавлять их в ARTM — пока **открытая проблема**
- Регуляризатор отбора тем имеет полезный побочный эффект, удаляя линейно зависимые и расщеплённые темы
- Почему это происходит — **открытая проблема**

Резюме

- Регуляризация — стандартный приём для решения некорректно поставленных задач
- ARTM позволяет комбинировать регуляризаторы и строить тематические модели с требуемыми свойствами
- BigARTM — эффективная реализация ARTM
- TopicNet — обёртка над BigARTM для экспериментов
- Декоррелирование — наиболее полезный регуляризатор
- Сглаживание + разреживание + декоррелирование — часто используемая комбинация регуляризаторов
- Оптимального числа тем, похоже, не существует
- Другие регуляризаторы — в следующих лекциях

Задания по курсу

Задача-минимум: научиться решать задачи NLP
с использованием тематического моделирования в BigARTM

Задача-максимум: сделать полезное мини-исследование

виды деятельности	оценка
теоретические задания	$\sum_i X_i$
решение прикладной задачи	$5X$
обзор по последним NeuralTM	$5X$
интеграция ARTM в pyTorch	$5X$
участие в одном из проектов	$10X$
работа над открытой проблемой	$10X$

где X — оценка за вид деятельности по 5-балльной шкале.
score — суммарная оценка по всем видам деятельности.

Итоговая оценка: $\min\big(10, \lfloor \text{score}/5 \rfloor\big)$ по 10-балльной шкале.

Теоретическое задание к лекции 1

Упражнения на принцип максимума правдоподобия:

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

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

Подсказка: применить условия KKT или основную лемму.

3. Творческое задание (возможны разные решения)
Предложите модель, определяющую роли слов в текстах:
— тематические слова
— специфичные слова документа (шум)
— слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов $p(r|w)$, $r \in \{\text{т}, \text{ш}, \text{ф}\}$.
Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.
Подсказка 3: можно использовать документную частоту слов.

Теоретическое задание к лекции 2

4. Запишите критерий логарифма правдоподобия с регуляризацией для тематической модели $p(w|d) = \sum_t \phi_{wt}\theta_{td}$, используя исходные данные $(d_i, w_i)_{i=1}^n$ вместо счётчиков n_{dw}. Выведете из него EM-алгоритм, докажите его эквивалентность обычному EM-алгоритму для ARTM.

5. Запишите критерий логарифма правдоподобия для локализованной тематической модели $p(w|C_i) = \sum_t \phi_{wt} p(t|C_i)$. Выведете из него EM-алгоритм с локализованным E-шагом.

Какие приближения пришлось сделать в процессе вывода? Какие переменные удобнее оставить в модели, ϕ_{wt} или ϕ'_{tw}?

6. Творческое задание (возможны разные решения) Предложите «какую-нибудь разумную» параметризацию для тематической модели внимания. Используя «основную лемму», получите уравнения для новых параметров модели.

Исследовательское задание к лекции 2

Открытая проблема. Продолжить исследование Ильи Ирхина:

- Освоить код: https://github.com/ilirhin/python_artm
- Реализовать локализованный E-шаг

Исследовать зависимость метрик качества от параметров (перплексия, разреженность, различность, когерентность):

- L — число проходов
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — асимметричность левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, глав
- β — баланса левого и правого контекста
- α, δ — параметры онлайнового EM-алгоритма
- опция «подставлять p_{ti}/n_t вместо $\phi_{w_i t}$ на E-шаге»
- опция «исключать p_{ti} позиции i из контекстов $\vec{\theta}_{ti}$, $\overleftarrow{\theta}_{ti}$»

Теоретическое задание к лекции 3

7. Выведете формулы EM-алгоритма в случае, когда логарифм в функции потерь заменяется гладкой монотонно возрастающей функцией ℓ:

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

Подумайте, какие замены логарифма полезны, и почему.

8. Замените ln гладкой монотонно возрастающей функцией μ 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?

9. Какому регуляризатору соответствует формула M-шага

$$\phi_{wt} = \operatorname*{norm}_{w}\big(n_{wt}[n_{wt} > \gamma n_t]\big)$$

Теоретическое задание к лекции 3

Аналитик построил тематическую модель Φ^0, Θ^0
и отметил среди столбцов матрицы Φ^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 Φ;
- остальные темы построились по-другому и были не похожи на каждую из неудачных тем $t \in T_-$.

10. Предложите регуляризаторы для этого.

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

12. Предложите способ инициализации Φ для новой модели.

Примеры датасетов для практических заданий по курсу

- Открытые датасеты (английский): 20 newsgroups, NIPS, KOS
- Научные статьи: eLibrary, Semantic Scholar, arXiv, PubMed
- Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
- TechCrunch (английский)
- Данные социальных сетей: VK, Twitter, Telegram,...
- Википедия
- Новостной поток (20 источников на русском языке)
- Данные кадровых агентств: резюме + вакансии
- Транзакции клиентов Sberbank DSD 2016
- Акты арбитражных судов РФ

---

http://bigartm.org
http://drive.google.com/drive/folders/1PPnw6aZOJAJoLRYuwdGm437RssV-XQx0

Проекты

- «Мастерская знаний» для научного поиска
  — пользователь строит тематические подборки статей,
  — поисковая выдача формируется моделью SciRus.
  — задача: показать пользователю тематику подборки
  — понадобится автоматическое выделение терминов,
  — выделение тематических фраз из документов,
  — автоматическое именование и суммаризация тем
  — конечная цель: ускорить понимание предметной области
- «Тематизатор» для социо-гуманитарных исследований
  — пользователь задаёт грубый фильтр текстового потока
  — задача: «классифицировать иголки в стоге сена»,
  — разделив темы на информативные и мусорные,
  — выделив аспекты и тональности в каждой теме
  — конечная цель: q&q аналитика проблемной среды

Открытые проблемы тематического моделирования

1. Проблема несбалансированности тем в коллекции
2. Обеспечение 100%-й интерпретируемости тем
3. Тематические модели внимания последовательного текста
4. Обнаружение новых тем или трендов в потоке текстов
5. Автоматическое именование и аннотирование тем
6. Обзор подходов в нейросетевых тематических моделях
7. Обеспечение полноты и устойчивости множества тем
8. Автоматический подбор гиперпараметров, AutoML
9. Оптимизация гиперпараметров в потоковом режиме
10. Проблема несбалансированности текстов по длине
11. Бережное слияние моделей нескольких коллекций
12. Гиперграфовые тематические модели в RecSys