Вероятностные тематические модели
Лекция 3. Онлайновый EM-алгоритм и часто используемые регуляризаторы

К. В. Воронцов
vokov@forecsys.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 • 21 сентября 2023

Содержание

1. **Онлайновый EM-алгоритм**
   - Тематические модели и регуляризация (напоминания)
   - Рациональный и онлайновый EM-алгоритм
   - Библиотеки BigARTM и TopicNet
2. **Часто используемые регуляризаторы**
   - Сглаживание и разреживание
   - Частичное обучение
   - Декоррелирование
3. **Проблема оптимизации числа тем**
   - Разреживающий регуляризатор для отбора тем
   - Сравнение с моделью HDP
   - Проблема несбалансированности тем

Напоминание. Задача тематического моделирования

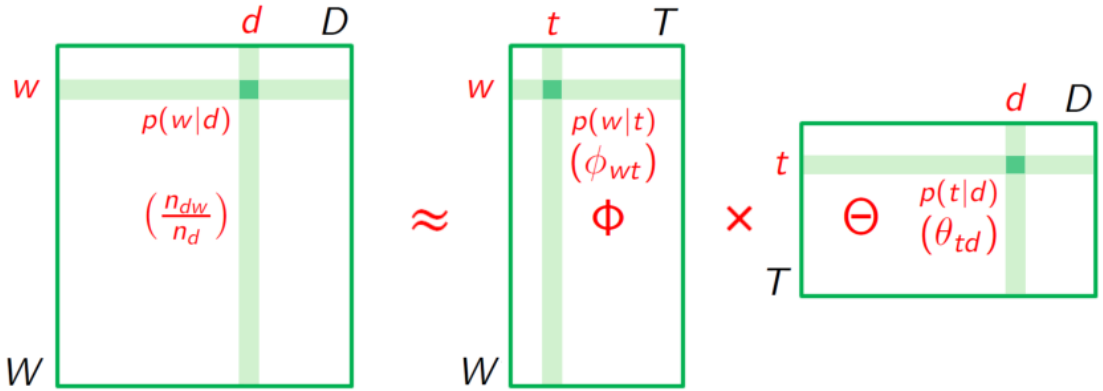
Дано: коллекция текстовых документов, $p(w|d) = \frac{n_{dw}}{n_d}$

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. JMLR, 2003.

Напоминание. ARTM — аддитивная регуляризация

Максимизация log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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. Комбинирование регуляризаторов в ARTM

Максимизация log правдоподобия с k регуляризаторами R_i:

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

где τ_i — коэффициенты регуляризации.

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Vorontsov K., Potapenko A. Additive regularization of topic models. Machine Learning, 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 Φ и тем документов Θ;

инициализация ϕ_{wt}, θ_{td} для всех $d \in D$, $w \in W$, $t \in T$;
для всех итераций $i = 1, \ldots, i_{\max}$
  $n_{wt}, n_{td} := 0$ для всех $d \in D$, $w \in W$, $t \in T$;
  для всех документов $d \in D$ и всех термов $w \in d$
    $n_{tdw} := n_{dw} \operatorname*{norm}_{t \in T}\left(\phi_{wt}\theta_{td}\right)$ для всех $t \in T$;
    $n_{wt} \mathrel{+}= n_{tdw}$; $n_{td} \mathrel{+}= n_{tdw}$ 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 $j_{\max}$, γ;
Выход: матрицы термов тем Φ и тем документов Θ;

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 $j = 1, \ldots, j_{\max}$ (итерации по документу)
    $n_{tdw} := n_{dw} \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$ для всех $w \in d$;
    $\theta_{td} := \underset{t \in T}{\text{norm}}\left(\sum_w n_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$;
  $\tilde{n}_{wt} := \tilde{n}_{wt} + n_{tdw}$ для всех $w \in d$;
  если пора обновить матрицу Φ **то**
    $n_{wt} := \gamma n_{wt} + \tilde{n}_{wt}$; $\tilde{n}_{wt} := 0$;
    $\phi_{wt} := \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)$;

Пакетный онлайновый EM-алгоритм в BigARTM

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

Вход: коллекция документов D, число тем $|T|$,
параметры $\delta \equiv$ `decay_weight`, $\alpha \equiv$ `apply_weight`;
Выход: матрица Φ;

инициализировать $n_{wt} := 0$, $\tilde{n}_{wt} := 0$, $\phi_{wt} :=$ random;
для всех пакетов D_b, $b = 1, \ldots, B$
  $(\tilde{n}_{wt}) := (\tilde{n}_{wt}) +$ **ProcessBatch**(D_b, Φ);
  если пора обновить матрицу Φ **то**
    $n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}$, $\tilde{n}_{wt} := 0$;
    $\phi_{wt} := \underset{w \in W}{\text{norm}} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$;

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

Пакетный онлайновый EM-алгоритм: функция ProcessBatch

Функция **ProcessBatch** обрабатывает пакет документов D_b, не меняя матрицу Φ, 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 $j = 1, \ldots, j_{\max}$ (итерации по документу)
    $p_{tdw} := \underset{t \in T}{\operatorname{norm}}\left(\phi_{wt}\theta_{td}\right)$;
    $\theta_{td} := \underset{t \in T}{\operatorname{norm}}\left(\sum_{w \in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$;
  $\tilde{n}_{wt} := \tilde{n}_{wt} + n_{dw} p_{tdw}$;

Сравнение оффлайнового и онлайнового алгоритмов

Оффлайн EM-алгоритм:

1. многократное итерирование по коллекции
2. однократный проход по документу
3. хранение матрицы Θ
4. обновление Φ 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 Θ
4. обновление Φ через заданное число пакетов
5. применяется при потоковой обработке больших коллекций

Матричная запись 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):

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

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

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.

Улучшение сходимости

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.

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.

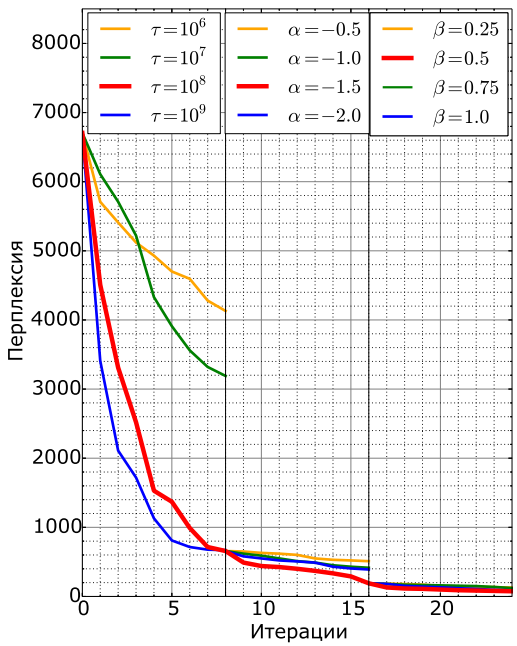
3. Регуляризаторы можно включать не сразу или по очереди.

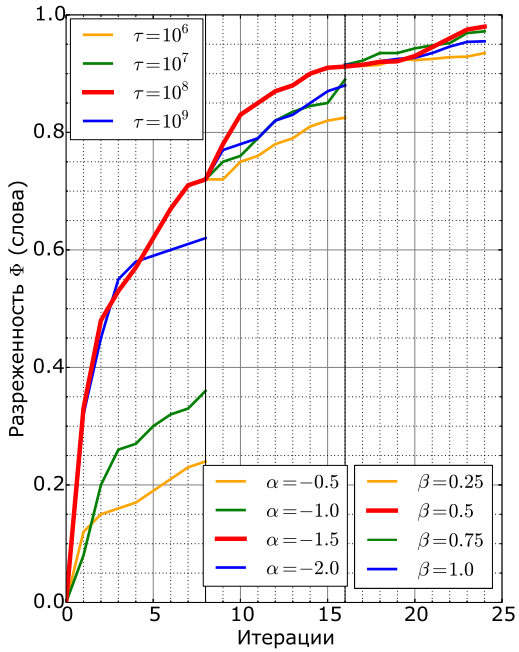
4. Регуляризаторы можно отключать по достижению эффекта.

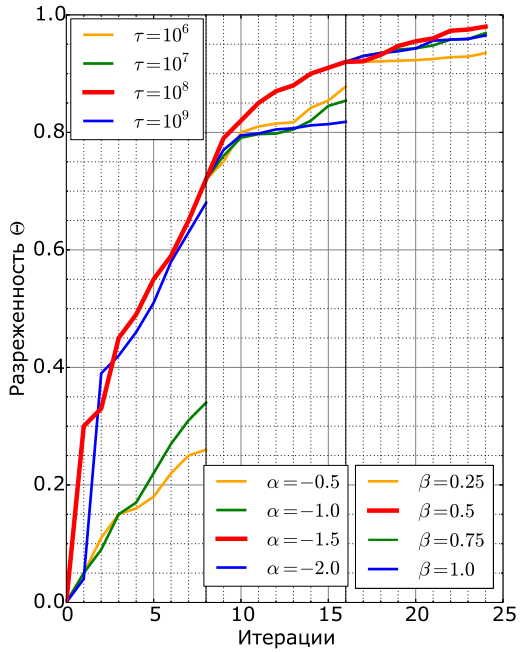
5. Одни регуляризаторы могут выполнять подготовительную работу для применения следующих регуляризаторов.

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

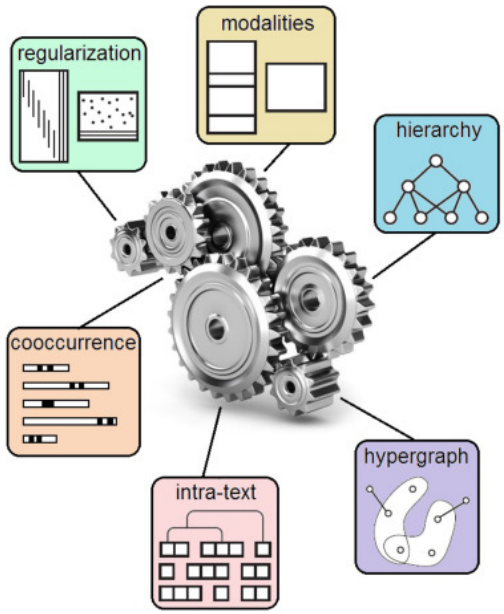
Лицензия и среда разработки:

- 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.

Дивергенция Кульбака–Лейблера и её свойства

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

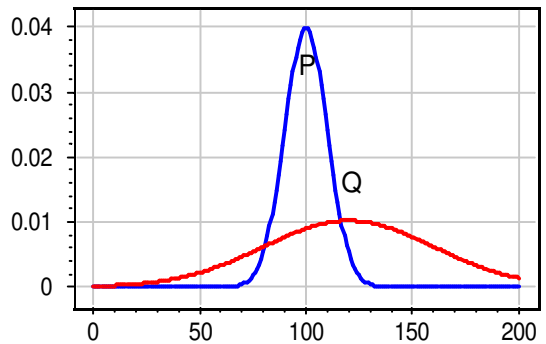
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

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

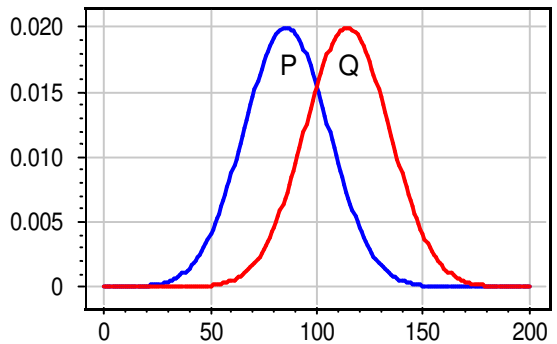
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

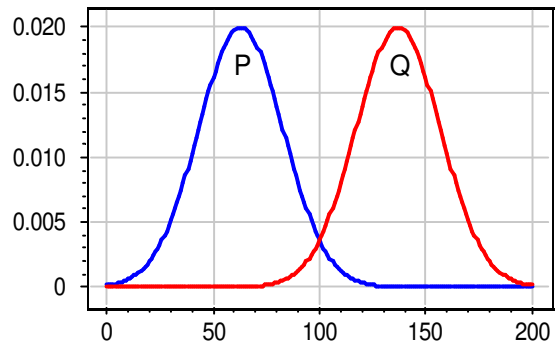
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



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



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



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

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

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

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. 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 псевдо-документ с ключевыми термами темы
- задать фоновые темы с общей лексикой языка
- задать шумовую тему для нетематичных термов

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»:

- $\beta_{wt} = \frac{1}{|W_t|}\big[w \in W_t\big]$ — 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

Проблема $\ln 0$ в дивергенции Кульбака–Лейблера

В регуляризаторе сглаживания/разреживания

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

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

Подправим регуляризатор, при сколь угодно малом ε:

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

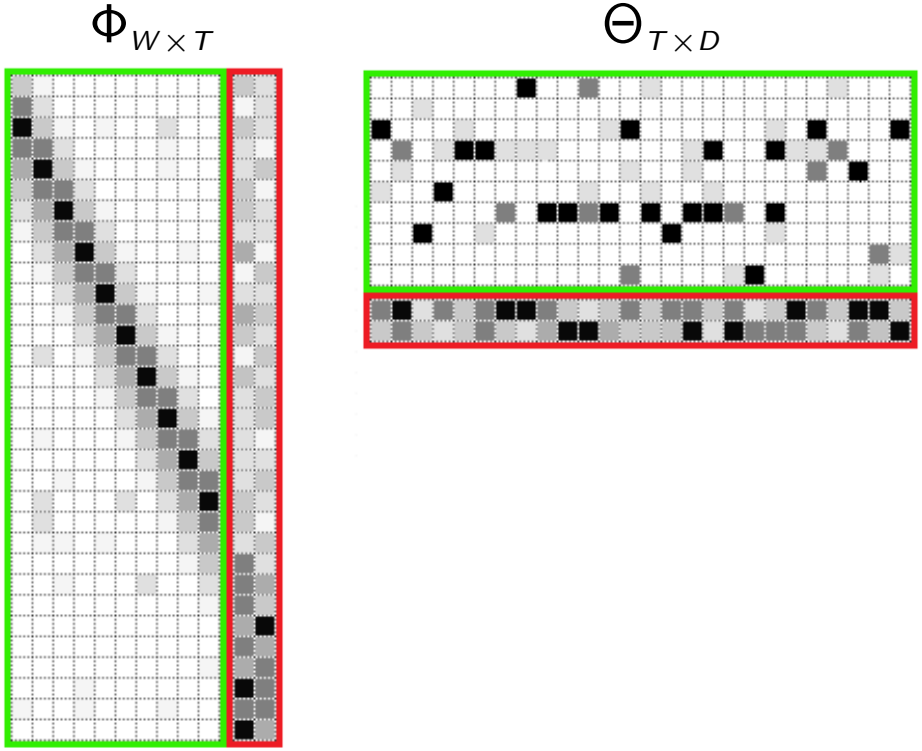
Подставив в формулу M-шага, получим для всех $t \in S$:

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

Если $\phi_{wt} = 0$, то разреживания не будет, но оно и не нужно.

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 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):

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

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.

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:

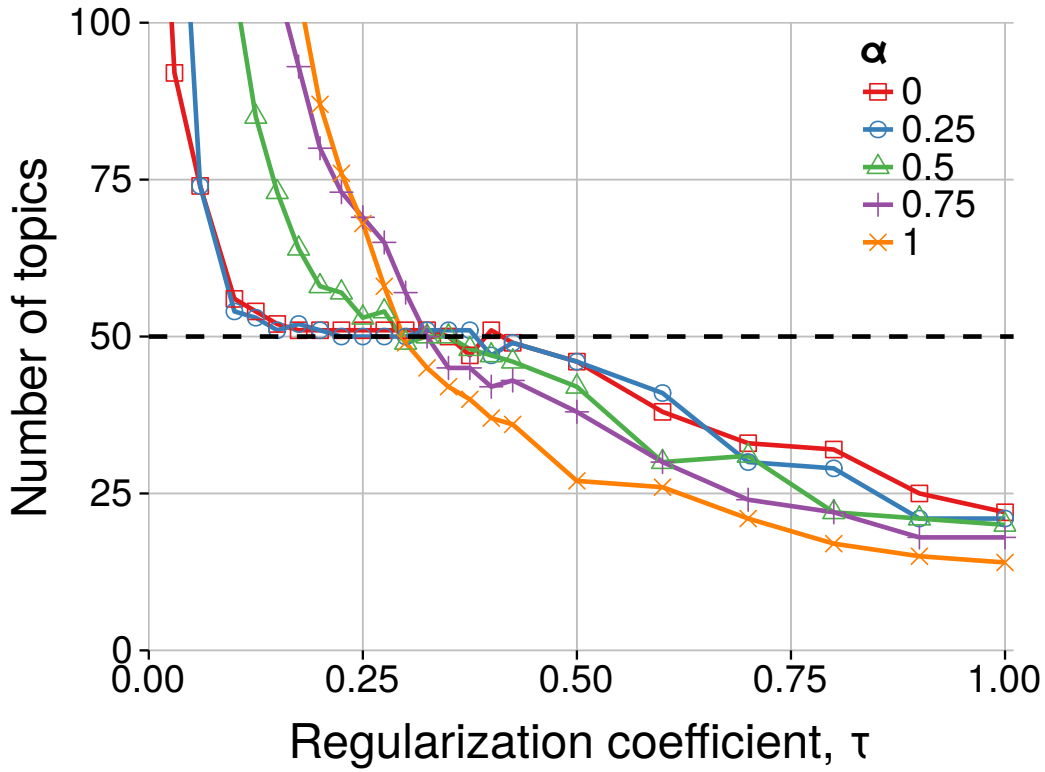
$$n^0_{dw} = n_d \sum_{t \in T_0} \phi_{wt}\theta_{td}$$

Параметрическое семейство полусинтетических данных:

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

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

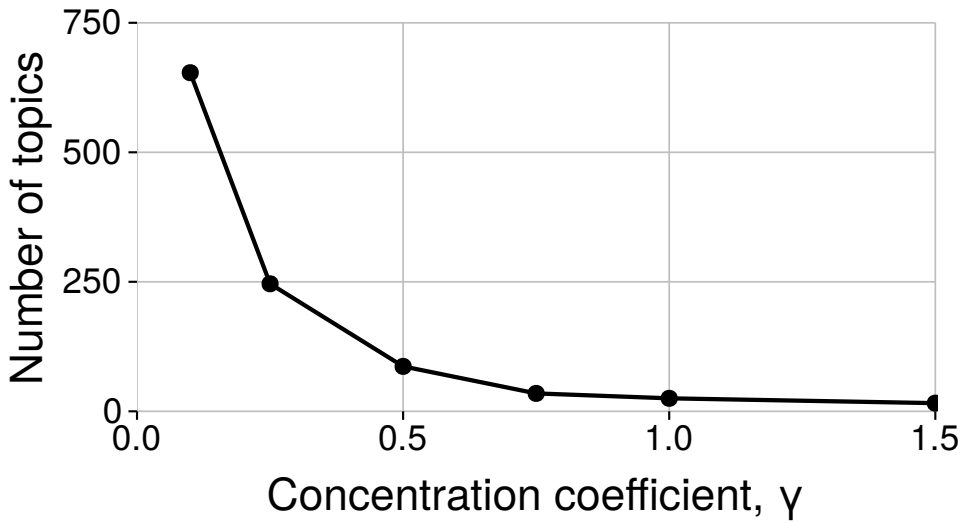
Попытка определения числа тем



- 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

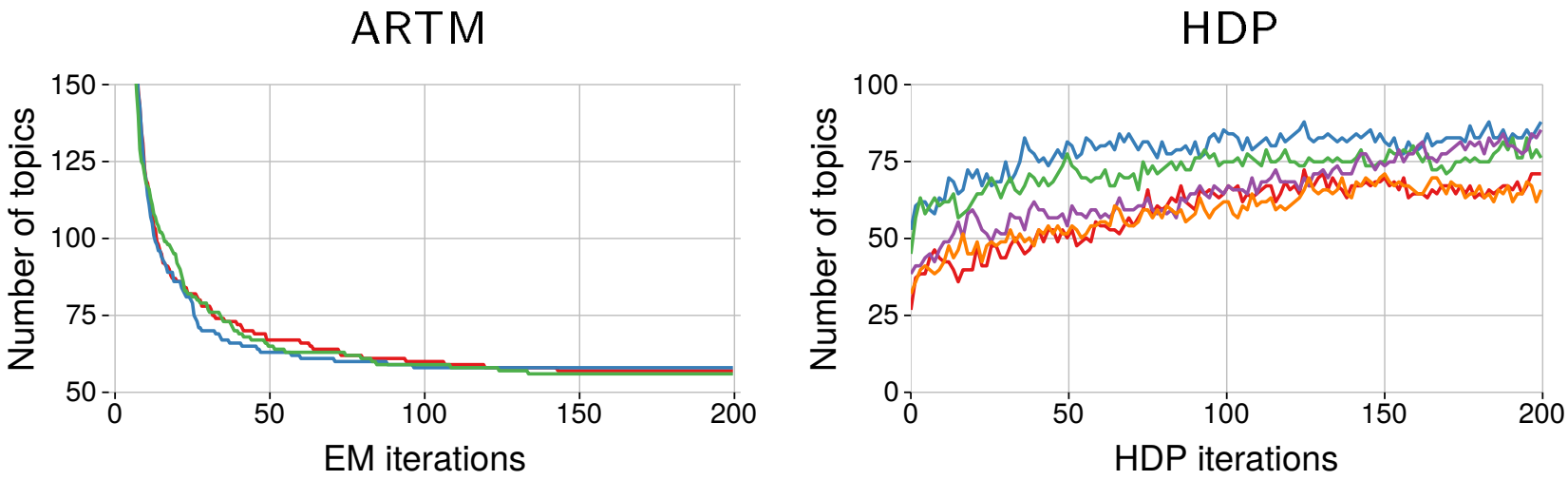


- *Коэффициент концентрации* γ в HDP влияет на $|T|$ так же сильно, как выбор коэффициента τ в ARTM.

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

Сравнение ARTM и HDP по устойчивости

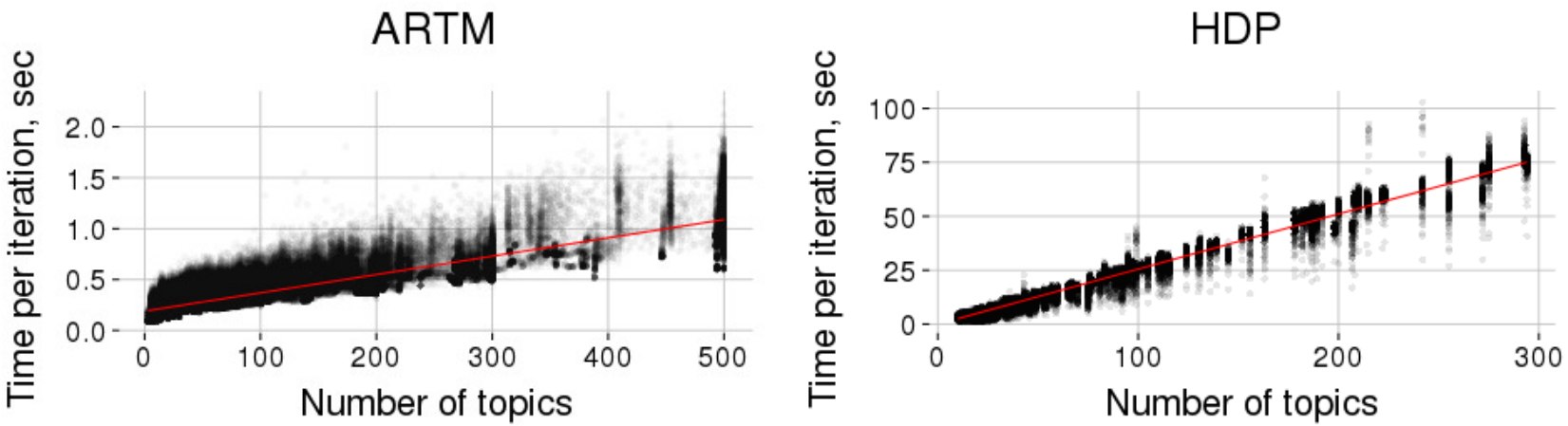
Запуск ARTM и HDP много раз из случайных инициализаций:



- 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)

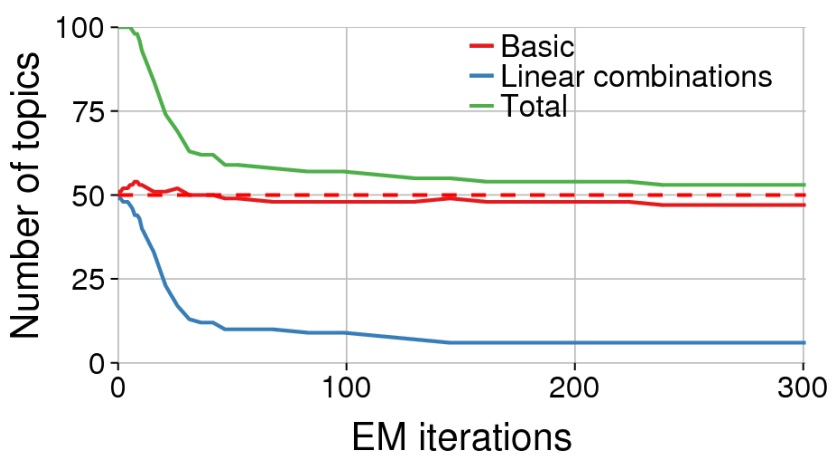


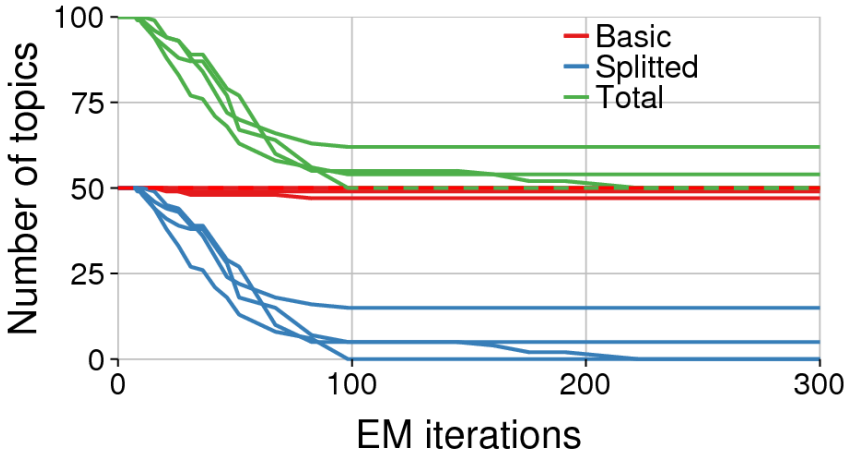
- ARTM в 100 раз быстрее!

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

Удаление линейно зависимых и расщеплённых тем

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





- 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.

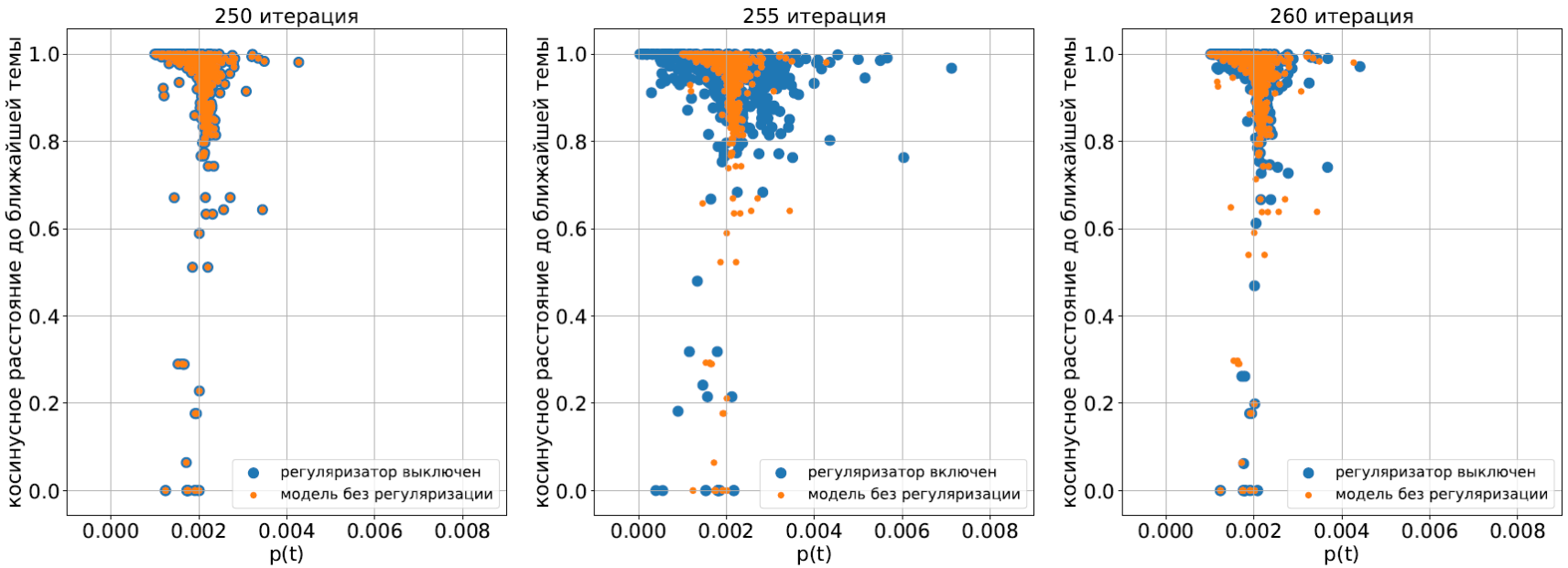
Выводы по результатам экспериментов

- 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 добавлять темы в ARTM — пока **открытая проблема**
- Регуляризатор отбора тем имеет полезный побочный эффект, удаляя линейно зависимые и расщеплённые темы
- Почему это происходит — **открытая проблема**

Проблема малых тем и тем-дубликатов

Эксперимент на коллекции postnauka.ru, $|T| = 500$

- Регуляризатор отбора тем плохо устраняет дубликаты,
- усиливает разброс тем по их мощности $p(t)$,
- который исчезает после отключения регуляризатора.
- Матричное разложение само не производит малые темы.

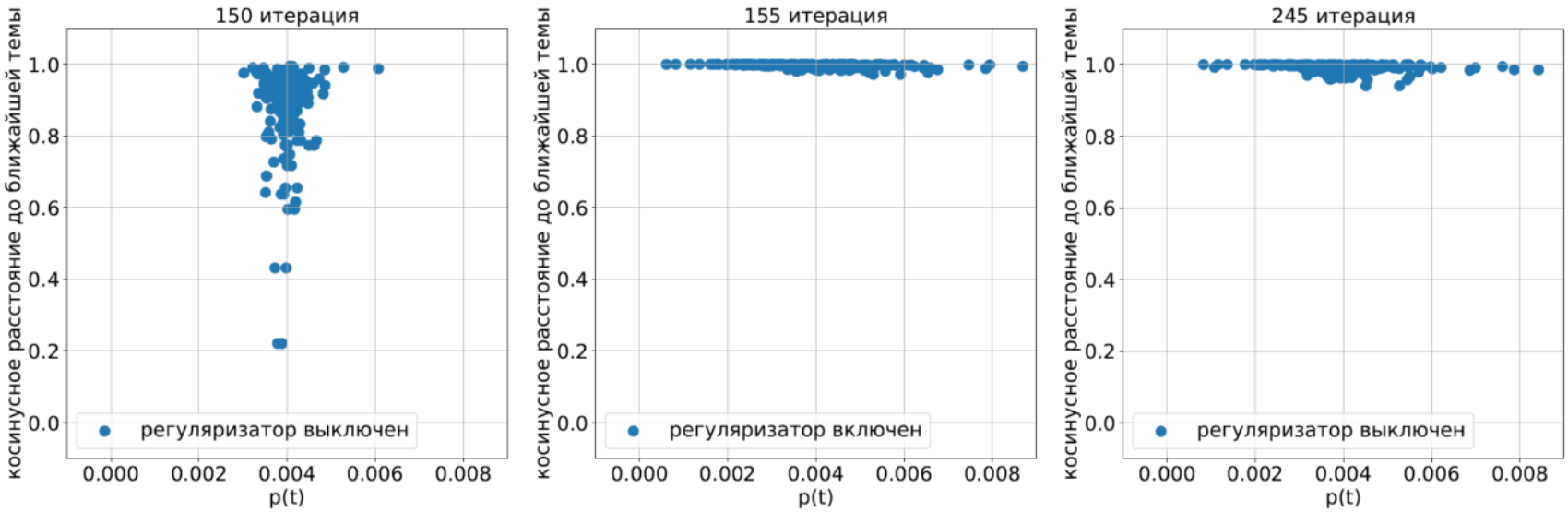


Г.Фоминская. Выявление тем-дубликатов в тематических моделях. Курсовая работа, ВМК МГУ, 2018.

Проблема малых тем и тем-дубликатов

Эксперимент на коллекции postnauka.ru, $|T| = 250$

- Регуляризатор декоррелирования удаляет дубликаты,
- усиливает разброс тем по их мощности $p(t)$;
- после отключения регуляризатора эти эффекты остаются.

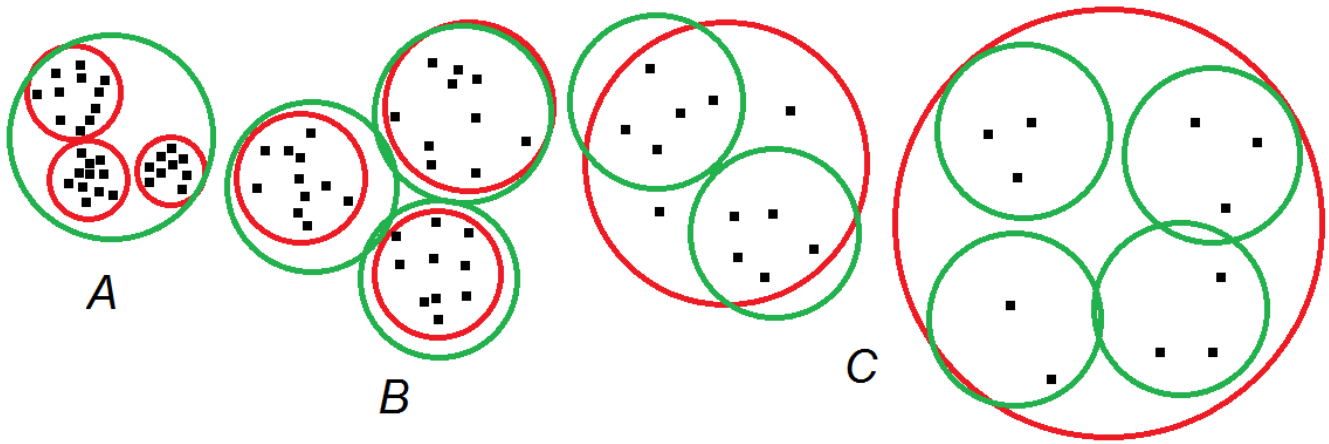


Г.Фоминская. Выявление тем-дубликатов в тематических моделях. Курсовая работа, ВМК МГУ, 2018.

Проблема расщепления и слияния тем

Тема — кластер на единичном симплексе размерности $|W| - 1$ с центром $p(w|t)$ и точками $p(w|t,d)$, $d \in D: \theta_{td} > 0$

- Тематические модели стремятся выравнивать темы по их мощности (красные кластеры).
- Это приводит к появлению тем-дубликатов (A) и семантически разнородных тем (C).
- Выравнивание тем по *радиусу семантической однородности* (зелёные кластеры) должно решать обе проблемы.



Резюме

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

Теоретическое задание к лекции №1

Два упражнения на принцип максимума правдоподобия:

- Униграммная модель документов: $p(w|d) = \xi_{dw}$
  Найти параметры модели ξ_{dw}.
- Униграммная модель коллекции: $p(w|d) = \xi_w$ для всех d
  Найти параметры модели ξ_w.

Подсказка: применить условия ККТ.

Третье упражнение в продолжение — более творческое:

- 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

1. Заменим log другой монотонно возрастающей функцией μ:

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

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

2. Заменим log 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?

3*. Какому регуляризатору соответствует формула M-шага

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

Подсказка: см. слайд 12 лекции №3.

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 $t \in T_-$.

1. Предложите регуляризаторы для этого.

2. 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?

3. Почему это плохо? Как этого избежать?

4. Предложите способ инициализации Φ для новой модели.