Вероятностные тематические модели коллекций текстовых документов

К. В. Воронцов
vokov@forecsys.ru

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

ноябрь 2012

Содержание

Задача определения тематики коллекции документов

Дано:

W — словарь, множество слов (терминов)

D — множество (коллекция, корпус) текстовых документов

n_{dw} — сколько раз термин $w \in W$ встретился в документе $d \in D$

Найти:

— к каким темам относится каждый документ

— какими терминами определяется каждая тема

Дополнительно найти (не в этой лекции!):

— сколько тем содержится в коллекции

— как темы выстраиваются в иерархию

— как темы развиваются во времени

— тематику объектов, связанных с документами: рисунков, авторов, журналов, конференций, организаций, стран и т. д.

Цели тематического моделирования (topic modeling)

- Тематический поиск документов и объектов по тексту любой длины или по любому объекту
- Категоризация, классификация, аннотирование, суммаризация текстовых документов

Типичные приложения:

- Поиск научной информации
- Поиск экспертов (expert search), рецензентов, проектов
- Выявление трендов и фронта исследований
- Анализ и агрегирование новостных потоков
- Рекомендательные сервисы (коллаборативная фильтрация)
- Рубрикация коллекций изображений, видео, музыки
- Аннотация генома и другие задачи биоинформатики

Стандартные гипотезы тематического моделирования

1. Порядок документов в коллекции не важен
2. Порядок слов в документе не важен (bag of words)
3. Слова, встречающиеся «почти во всех» документах, не важны
4. Слово в разных формах — это одно и то же слово
5. Документ обычно относится к небольшому числу тем
6. Тема обычно определяется небольшим числом терминов

Предварительная обработка текстов:

- Приведение всех слов к нормальной форме (лемматизация или стемминг)
- Выделение терминов (term extraction) и/или выделение словосочетаний (key phrase extraction) (сводятся к задачам классификации или ранжирования)
- Удаление стоп-слов $w \in W$: $\#\{d : w \in d\} \geqslant \alpha|D|$, $\alpha \sim 0.05 \ldots 0.5$

Вероятностная формализация постановки задачи

Вероятностные предположения:

- каждое слово в документе связано с некоторой темой $t \in T$;
- коллекция D — это выборка независимых наблюдений (d, w) из дискретного распределения $p(d, w, t)$ на $D \times W \times T$;
- гипотеза условной независимости: $p(w|d, t) = p(w|t)$;

Вероятностная модель порождения документа d:

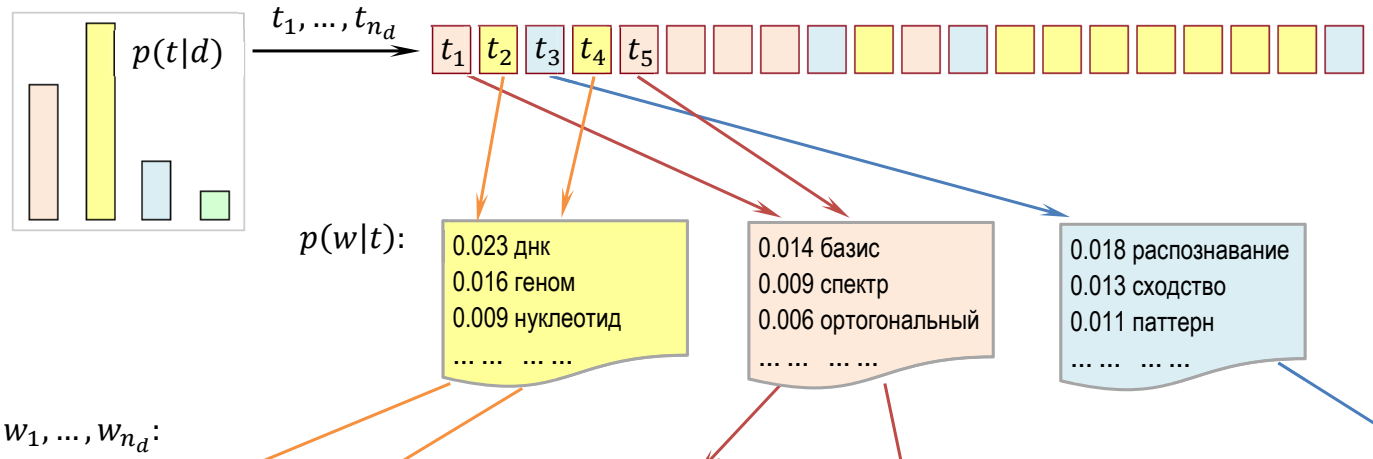
$$p(w|d) = \sum_{t \in T} p(w|d, t)\, p(t|d) = \boxed{\sum_{t \in T} p(w|t)\, p(t|d)}$$

Найти:

- $p(w|t)$ — распределение терминов в каждой теме $t \in T$;
- $p(t|d)$ — распределение тем в каждом документе $d \in D$.

Вероятностная модель порождения документа d

Вероятностная тематическая модель: $p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Процесс порождения документа d

Вероятностная тематическая модель: $p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$

Вход:

распределение $p(w|t)$ для каждой темы $t \in T$;

распределение $p(t|d)$ для каждого документа $d \in D$;

Выход:

коллекция документов;

1: **для всех** документов $d \in D$
2: **для всех** слов w в документе d
3: выбрать тему t из $p(t|d)$;
4: выбрать слово w из $p(w|t)$;

Униграммная модель порождения текста

Два упражнения на принцип максимума правдоподобия:

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

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \xi_{dw} \to \max_{\xi}, \quad \sum_{w\in W} \xi_{dw} = 1, \quad \xi_{dw} \geqslant 0.$$

$$\mathscr{L} = \sum_{d\in D}\Big(\sum_{w\in d} n_{dw} \ln \xi_{dw} - \lambda_d\Big(\sum_{w\in W} \xi_{dw} - 1\Big)\Big);$$

$$\frac{\partial \mathscr{L}}{\partial \xi_{dw}} = n_{dw}\frac{1}{\xi_{dw}} - \lambda_d = 0 \quad \Rightarrow \quad \lambda_d = n_d, \quad \xi_{dw} = \frac{n_{dw}}{n_d} \equiv \hat{p}(w|d).$$

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

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \xi_{w} \to \max_{\xi}, \quad \sum_{w\in W} \xi_{w} = 1, \quad \xi_{w} \geqslant 0.$$

$$\mathscr{L} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \xi_{w} - \lambda\Big(\sum_{w\in W} \xi_{w} - 1\Big);$$

$$\frac{\partial \mathscr{L}}{\partial \xi_{w}} = n_{dw}\frac{1}{\xi_{w}} - \lambda = 0 \quad \Rightarrow \quad \lambda = n, \quad \xi_{w} = \frac{n_{w}}{n} \equiv \hat{p}(w).$$

Недостатки униграммной модели. Модель смеси униграмм

- тематика не выявляется
- число $|W| \cdot |D|$ оцениваемых параметров $p(w|d)$ линейно зависит от $|D|$ — числа документов в коллекции
- зависимости между документами не учитываются

Эти недостатки устраняются в *модели смеси униграмм*:

$$\sum_{d\in D} \ln \underbrace{\sum_{t\in T} p(t) \overbrace{\prod_{w\in d} p(w|t)^{n_{dw}}}^{p(w_1,\ldots,w_{n_d}|d,t)}}_{p(w_1,\ldots,w_{n_d}|d)} \to \max_{\{p(t),p(w|t)\}}$$

НО модель смеси униграмм имеет свой недостаток:

- каждый документ порождается только одной темой.

Nigam, McCallum, Thrun, Mitchell. Text classification from labeled and unlabeled documents using EM // Machine Learning, 2000, 39(2–3): 103–134

Вероятностная латентная семантическая модель PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Максимизация правдоподобия по $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w|d) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w\in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{d\in D} \theta_{td} = 1.$$

Интерпретация №1: минимизация суммарной (по $d \in D$) дивергенции Кульбака–Лейблера между тематическими моделями $p(w|d)$ и униграммными $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$:

$$\mathrm{KL}(\hat{p}\|p) = \sum_{d\in D}\sum_{w\in d} \hat{p}(w|d) \ln \frac{\hat{p}(w|d)}{p(w|d)} \to \min.$$

Вероятностная латентная семантическая модель PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Максимизация правдоподобия по $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w|d) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w\in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{d\in D} \theta_{td} = 1.$$

Интерпретация №2: неотрицательное матричное разложение

$$\|F - \Phi\Theta\|_{\mathrm{KL}} \to \min_{\Phi,\Theta}$$

$F = \big(\hat{p}(w|d)\big)_{W\times D}$ — известная матрица исходных данных;
$\Phi = \big(\phi_{wt}\big)_{W\times T}$ — искомая матрица терминов тем $\phi_{wt} = p(w|t)$;
$\Theta = \big(\theta_{td}\big)_{T\times D}$ — искомая матрица тем документов $\theta_{td} = p(t|d)$.

EM-алгоритм

E-шаг: условные вероятности тем $p(t|d,w)$ для всех t, d, w вычисляются через ϕ_{wt}, θ_{td} по формуле Байеса:

$$H_{dwt} \equiv p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}.$$

M-шаг: решение задачи максимизации правдоподобия выражается аналитически через частотные оценки условных вероятностей, если положить $\hat{n}_{dwt} = n_{dw} H_{dwt}$:

$$\phi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \qquad \hat{n}_{wt} = \sum_{d \in D} \hat{n}_{dwt}, \qquad \hat{n}_t = \sum_{w \in W} \hat{n}_{wt};$$

$$\theta_{td} = \frac{\hat{n}_{dt}}{\hat{n}_d}, \qquad \hat{n}_{dt} = \sum_{w \in d} \hat{n}_{dwt}, \qquad \hat{n}_d = \sum_{t \in T} \hat{n}_{dt}.$$

EM-алгоритм — это чередование E и M шагов до сходимости.

Частотные оценки условных вероятностей

Вероятностная тематическая модель: $p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$

Если рассматривать коллекцию как выборку троек (d, w, t), то

$$\hat{p}(w|d) = \frac{n_{dw}}{n_d}, \qquad \hat{p}(w|t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(t|d) = \frac{n_{dt}}{n_d};$$

n_{dwt} — число троек (d, w, t) во всей коллекции;

$n_{dw} = \sum_{t \in T} n_{dwt}$ — число вхождений термина w в документ d;

$n_{dt} = \sum_{w \in d} n_{dwt}$; $\qquad n_d = \sum_{w \in d} \sum_{t \in T} n_{dwt}$ — длина документа d;

$n_{wt} = \sum_{d \in D} n_{dwt}$; $\qquad n_t = \sum_{d \in D} \sum_{w \in d} n_{dwt}$ — «длина темы» t;

$n = \sum_{d \in D} \sum_{w \in d} \sum_{t \in T} n_{dwt}$ — длина всей коллекции;

Вывод формулы M-шага для ϕ_{wt}

Лагранжиан задачи максимизации правдоподобия при ограничениях нормировки но без ограничений неотрицательности:

$$\mathscr{L} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} - \sum_{t\in T}\lambda_t\left(\sum_{w\in W}\phi_{wt} - 1\right) - \sum_{d\in D}\mu_d\left(\sum_{t\in T}\theta_{td} - 1\right)$$

$$\frac{\partial \mathscr{L}}{\partial \phi_{wt}} = \sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)} - \lambda_t = 0;$$

$$\sum_{d\in D} n_{dw}\frac{\theta_{td}\phi_{wt}}{p(w|d)} = \lambda_t\phi_{wt} \quad \Rightarrow \quad \lambda_t = \sum_{d\in D}\sum_{w\in W} n_{dw}H_{dwt};$$

$$\phi_{wt} = \frac{\sum\limits_{d\in D} n_{dw}H_{dwt}}{\sum\limits_{d\in D}\sum\limits_{w'\in d} n_{dw'}H_{dw't}} \equiv \frac{\hat{n}_{wt}}{\hat{n}_t} \quad \text{для всех } w\in W,\ t\in T.$$

Вывод формулы M-шага для θ_{td}

Лагранжиан задачи максимизации правдоподобия при ограничениях нормировки но без ограничений неотрицательности:

$$\mathscr{L} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} - \sum_{t\in T}\lambda_t\left(\sum_{w\in W}\phi_{wt} - 1\right) - \sum_{d\in D}\mu_d\left(\sum_{t\in T}\theta_{td} - 1\right)$$

$$\frac{\partial \mathscr{L}}{\partial \theta_{td}} = \sum_{w\in d} n_{dw}\frac{\phi_{wt}}{p(w|d)} - \mu_d = 0;$$

$$\sum_{w\in d} n_{dw}\frac{\phi_{wt}\theta_{td}}{p(w|d)} = \mu_d\theta_{td} \quad \Rightarrow \quad \mu_d = \sum_{t\in T}\sum_{w\in d} n_{dw}H_{dwt};$$

$$\theta_{td} = \frac{\sum\limits_{w\in d} n_{dw}H_{dwt}}{\sum\limits_{w\in d} n_{dw}\sum\limits_{t'\in T} H_{dwt'}} \equiv \frac{\hat{n}_{dt}}{\hat{n}_d} \quad \text{для всех } d\in D,\ t\in T.$$

Недостатки классического PLSA

- PLSA медленно сходится на больших коллекциях, т.к. Φ и Θ обновляются после каждого прохода коллекции
- PLSA не разреживает распределение $H_{dwt} = p(t|d, w)$
- PLSA вынужден хранить 3D-матрицу $H = (H_{dwt})_{D \times W \times T}$
- PLSA переобучается, т.к. параметров ϕ_{wt} и θ_{td} слишком много $(|D| \cdot |T| + |W| \cdot |T|)$, возможно переобучение
- PLSA неверно оценивает вероятность новых слов: если $n_w = 0$, то $\hat{p}(w|t) = 0$ для всех $t \in T$
- PLSA не позволяет управлять *разреженностью* Φ и Θ, т.к.
  (в начале $\phi_{wt} = 0$) $\Leftrightarrow$ (в финале $\phi_{wt} = 0$)
  (в начале $\theta_{td} = 0$) $\Leftrightarrow$ (в финале $\theta_{td} = 0$)

Рациональный EM-алгоритм: E-шаг встроен внутрь M-шага

Идея: не хранить H_{dwt}, а вычислять по мере необходимости.
Сложность алгоритма $O(|D| \cdot |W| \cdot |T|)$.

Вход: коллекция D, число тем $|T|$, начальные Φ и Θ;
Выход: распределения Φ и Θ;

повторять
  обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ для всех $d \in D$, $w \in W$, $t \in T$;
  для всех $d \in D$, $w \in d$
    $Z := \sum_t \phi_{wt}\theta_{td}$;
    для всех $t \in T$ таких, что $\phi_{wt}\theta_{td} > 0$
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на $n_{dw}\frac{1}{Z}\phi_{wt}\theta_{td}$;
  $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W$, $t \in T$;
  $\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D$, $t \in T$;
пока Φ и Θ не стабилизируются.

Обобщённый EM-алгоритм (GEM, generalized EM-algorithm)

Идея: не обязательно точно решать задачу M-шага, достаточно сместиться в направлении максимума и снова сделать E-шаг.
В PLSA это приводит к *частым обновлениям параметров* Φ, Θ:

Вход: коллекция D, число тем $|T|$, начальные Φ и Θ;
Выход: распределения Φ и Θ;

обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$, n_{dwt} для всех $d \in D$, $w \in W$, $t \in T$;
повторять
  для всех $d \in D$, $w \in d$
    для всех $t \in T$ таких, что $n_{dwt} > 0$ или $\phi_{wt}\theta_{td} > 0$
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$ на $(n_{dw}H_{dwt} - n_{dwt})$;
      $n_{dwt} := n_{dw}H_{dwt}$;
    если не первая итерация и пора обновить Φ, Θ **то**
      $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t$; $\theta_{td} := \hat{n}_{dt}/\hat{n}_d$;
пока Φ и Θ не стабилизируются.

Стохастический EM-алгоритм (S-GEM, Stochastic GEM)

В PLSA-GEM приходится хранить 3D-массив $n_{dwt} = n_{dw} H_{dwt}$.

Гипотеза разреженности: «употребление слова w в документе d связано с не более чем s темами».

Сэмплирование: для каждой пары (d, w) генерируется s случайных тем t_{dwi}, $i = 1, \ldots, s$, из распределения $p(t|d, w)$.

Это эквивалентно замене $H_{dwt} \equiv p(t|d, w)$ эмпирической оценкой по сгенерированной случайной выборке длины s:

$$\hat{H}_{dwt} \equiv \hat{p}(t|d, w) = \frac{1}{s} \sum_{i=1}^{s} \big[t_{dwi} = t\big].$$

Усредняя $\hat{H}_{dwt}$ вместо H_{dwt},
— получаем несмещённые оценки :)
— добиваемся разреженности матрицы H :)

Латентное размещение Дирихле
LDA — Latent Dirichlet Allocation [David Blei, 2003]

Вероятностная тематическая модель: $p(w|d) = \sum_{t \in T} \underbrace{p(w|t)}_{\phi_{wt}} \underbrace{p(t|d)}_{\theta_{td}}$

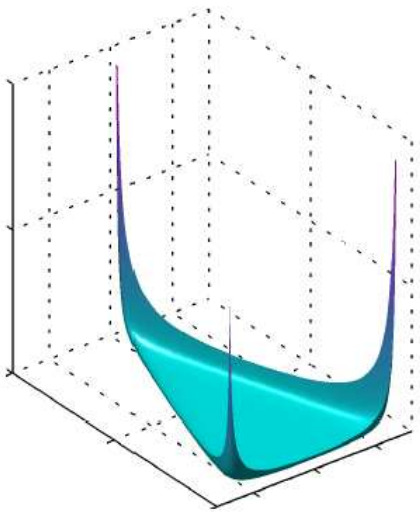
Гипотеза об априорных распределениях Дирихле:

- $\theta_d = (\theta_{td})_{t \in T} \in \mathbb{R}^{|T|}$ — случайные векторы из распределения Дирихле с параметром $\alpha \in \mathbb{R}^{|T|}$:
$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \alpha_0 = \sum_t \alpha_t, \quad \sum_t \theta_t = 1;$$
- $\phi_t = (\phi_{wt})_{w \in W} \in \mathbb{R}^{|W|}$ — случайные векторы из распределения Дирихле с параметром $\beta \in \mathbb{R}^{|W|}$:
$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \beta_0 = \sum_w \beta_w, \quad \sum_w \phi_{wt} = 1;$$

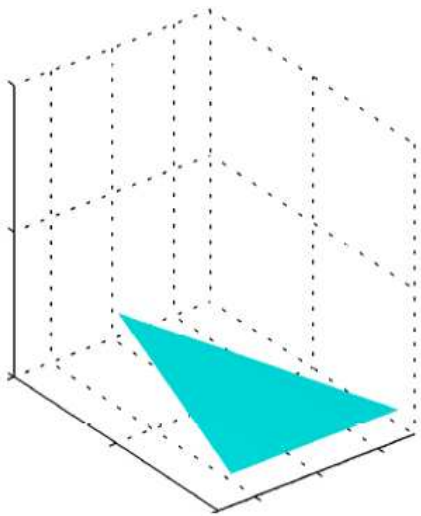
Почему именно Дирихле?

- Распределение Дирихле позволяет описывать кластерную структуру множества мультиномиальных распределений,
- в том числе разреженных мультиномиальных распределений.

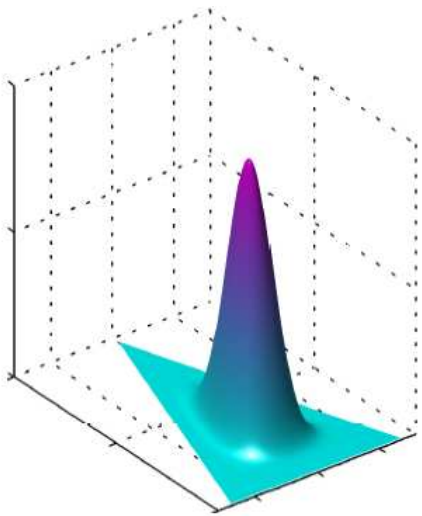
Пример. $\mathrm{Dir}(\theta|\alpha)$, $\theta = (\theta_1, \theta_2, \theta_3)$, $|T| = 3$:



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$



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



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

Почему именно Дирихле?

Пусть темы слов в документах $d \in D$ выбираются из θ_d:

$$X_d = \{t_1, \ldots, t_{n_d}\} \sim \theta_d.$$

Тогда вероятность встретить каждую из тем t ровно n_{td} раз подчиняется мультиномиальному распределению:

$$p(X_d|\theta_d) = \text{Mult}(n_{1d}, \ldots, n_{Td}|\theta_d) = \frac{n_d!}{\prod_t n_{td}!} \prod_t \theta_{td}^{n_{td}}.$$

Если предположить, что $\theta_d \sim \text{Dir}(\alpha)$, то по формуле Байеса апостериорное распределение также из $\text{Dir}(\alpha')$, $\alpha'_t = \alpha_t + n_{td}$:

$$p(\theta_d|X_d) = \frac{p(X_d|\theta_d)\,\text{Dir}(\theta_d|\alpha)}{\int p(X_d|\theta)\,\text{Dir}(\theta|\alpha)\,d\theta} \propto \prod_t \theta_{td}^{n_{td}} \theta_{td}^{\alpha_t - 1} = \text{Dir}(\theta_d; \alpha').$$

Распределение Дирихле — *сопряжённое* к мультиномиальному, что упрощает байесовское оценивание параметров ϕ_{wt} и θ_{td}.

Байесовская оценка параметров $\theta_{td} \equiv p(t|d)$

Оценка θ_{td} при априорном распределении:

$$\mathrm{E}p(t|d,\alpha) = \int \theta_{td}\mathrm{Dir}(\theta_d|\alpha)\,d\theta_d = \frac{\alpha_t}{\alpha_0}.$$

Пусть известна выборка тем $X_d = \{t_1, \ldots, t_{n_d}\} \sim \theta_d$.
Оценка θ_{td} при апостериорном распределении:

$$\mathrm{E}p(t|d,X_d,\alpha) = \int \theta_{td}\mathrm{Dir}(\theta_d|\alpha')\,d\theta_d = \frac{n_{td}+\alpha_t}{\sum\limits_{t'} n_{t'd}+\alpha_{t'}} = \frac{n_{td}+\alpha_t}{n_d+\alpha_0},$$

n_{td} — сколько раз слово документа d было отнесено к теме t,
n_d — длина документа в словах.

Замечание. Эта оценка переходит в МП-оценку при $\alpha_t \equiv 0$, хотя при $\alpha_t = 0$ распределение Дирихле не определено.

Байесовская оценка параметров $\phi_{wt} \equiv p(w|t)$

Оценка ϕ_{wt} при априорном распределении:

$$\mathrm{E}p(w|t,\beta) = \int \phi_{wt} \mathrm{Dir}(\phi_t|\beta)\, d\phi_t = \frac{\beta_w}{\beta_0}.$$

Коллекция порождается двумя распределениями $p(t|d)$, $p(w|t)$.
Часть коллекции, порождённая темой t:

$$X_t = \big\{(d,w,t)\colon d \in D,\ w \sim \phi_t\big\}.$$

Апостериорное распределение для ϕ_t по формуле Байеса:

$$p(\phi_t|X_t,\beta) = \frac{p(X_t|\phi_t)\,\mathrm{Dir}(\phi_t|\beta)}{\int p(X_t|\phi)\,\mathrm{Dir}(\phi|\beta)\,d\phi} = \mathrm{Dir}(\phi_t|\beta'), \quad \beta'_w = \beta_w + n_{wt}.$$

Оценка ϕ_{wt} через апостериорное распределение:

$$\mathrm{E}p(w|t,X_d,\beta) = \int \phi_{wt} \mathrm{Dir}(\phi_t|\beta')\, d\phi_t = \frac{n_{wt}+\beta_w}{n_t+\beta_0}.$$

Регуляризованный EM-алгоритм (R-GEM, Regularized GEM)

Чтобы преобразовать PLSA в LDA, достаточно
в EM-алгоритме частотные оценки максимума правдоподобия

$$\phi_{wt} \equiv p(w|t) = \frac{n_{wt}}{n_t}, \qquad \theta_{td} \equiv p(t|d) = \frac{n_{td}}{n_d}$$

заменить сглаженными (смещёнными) байесовскими оценками

$$\phi_{wt} = \frac{n_{wt} + \beta_w}{n_t + \beta_0}, \qquad \theta_{td} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}.$$

Более строгий вывод алгоритма сэмплирования Гиббса:

Yi Wang. Distributed Gibbs Sampling of Latent Dirichlet Allocation: The Gritty Details. 2011.

Алгоритм сэмплирования Гиббса [Griffiths, Steyvers, 2004]

Вход: коллекция D, число тем $|T|$, параметры α, β;
Выход: распределения Φ и Θ;

1: $n_{wt} := \beta_w$; $n_{td} := \alpha_t$ для всех $d \in D$, $w \in W$, $t \in T$;
2: $n_t := \beta_0$; $n_d := \alpha_0$ для всех $d \in D$, $t \in T$;
3: **для** $i := 1, \ldots, M$ (итерация = один проход коллекции)
4: **для** всех документов $d \in D$ и всех вхождений слова $w \in d$
5: **если** $i \geqslant 2$ **то** $t := t_{dw}$; $--n_{wt}$; $--n_{td}$; $--n_t$; $--n_d$;
6: $\tilde{p}(t|d,w) := (n_{wt}/n_t) \cdot (n_{td}/n_d)$ для всех $t \in T$;
7: выбрать t из ненормированного распределения $\tilde{p}(t|d,w)$;
8: $t_{dw} := t$; $++n_{wt}$; $++n_{td}$; $++n_t$; $++n_d$;
9: $\phi_{wt} := n_{wt}/n_t$ для всех $w \in W$, $t \in T$;
10: $\theta_{td} := n_{td}/n_d$ для всех $d \in D$, $t \in T$;

Алгоритмы обучения параметров модели LDA

- Сэмплирование Гиббса (GS — Gibbs Sampling)
  можно рассматривать как специальный случай S-GEM
  — с обновлением по каждой словопозиции (n_{dw} раз);
  — с сэмплированием 1 темы для каждой словопозиции;
  — с регуляризацией Дирихле и гиперпараметрами α, β.

  Griffiths, Steyvers. Finding scientific topics // Proceedings of the National Academy of Sciences. USA, 2004. — Vol. 101. — Pp. 5228–5235.
- VB, CVB — (Collapsed) Variational Bayesian inference
  можно рассматривать как специальный случай R-GEM:
  — с регуляризацией, но без сэмплирования.

  Teh, Newman, Wellingm. A collapsed variational Bayesian inference algorithm for Latent Dirichlet Allocation // Advances in Neural Information Processing Systems (NIPS). Cambridge, MA, MIT Press, 2006.

Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models // Int'l conf. on Uncertainty in Artificial Intelligence, 2009.

RS-GEM обобщает классический PLSA, LDA-GS и CVB0

- Частота обновления параметров Θ и Φ:
  - PLSA — раз в коллекцию;
  - RS-GEM — с произвольной периодичностью;
  - LDA-GS — для каждой словопозиции.
- Сэмплирование:
  - PLSA, LDA-VB — нет;
  - RS-GEM — произвольное число s тем для каждого (d, w);
  - LDA-GS — 1 тема для каждой словопозиции.
- Отбрасывание (d, w) при сэмплировании темы:
  - PLSA, LDA-VB — нет;
  - RS-GEM — нет, легко добавить, но это не даёт результата;
  - LDA-GS — необходимо, согласно теории.
- Регуляризация Дирихле с гиперпараметрами α, β:
  - PLSA — нет;
  - RS-GEM — легко включаемая опция;
  - LDA-GS, LDA-VB — необходимо, согласно теории.

Стандартная методика оценивания моделей языка

Перплексия тестовой коллекции D' (hold-out perplexity):

$$\mathcal{P}(D') = \exp\left(-\frac{\sum\limits_{d\in D'}\sum\limits_{w\in d''} n_{dw}\ln p(w|d)}{\sum\limits_{d\in D'}\sum\limits_{w\in d''} n_{dw}}\right),$$

$d = d' \sqcup d''$ — случайное разбиение контрольного документа на две половины равной длины;

параметры ϕ_{wt} оцениваются по обучающей коллекции D;
параметры θ_{td} оцениваются по первой половине d';
перплексия вычисляется по второй половине d''.

Интерпретации перплексии:
1) $\mathcal{P}(D') \to |W|$ при $n \to \infty$, если слова равновероятны;
2) насколько хорошо мы можем предсказывать появление слов (чем меньше перплексия, тем лучше).

Другие методики оценивания тематических моделей

- Число ошибок классификации размеченных текстов D'.
- Качество ранжирования при тематическом поиске.
- Отклонение от *гипотезы условной независимости* $p(w|d,t) = p(w|t)$ на обучающей коллекции D для темы t:

$$\mathrm{KL}\Big(\hat{p}(d,w|t) \,\Big\|\, \hat{p}(d|t)\cdot\hat{p}(w|t)\Big) = \sum_{d,w} \frac{n_{dwt}}{n_t} \log \frac{n_{dwt}\cdot n_t}{n_{td}\cdot n_{wt}}$$

D.Mimno, D.Blei. Bayesian checking for topic models // Empirical Methods in Natural Language Processing, 2011.

- Доля случаев, когда эксперт верно определяет:
  — лишнюю тему в списке главных тем документа;
  — лишний термин в списке главных терминов темы.

J.Chang, J.Boyd-Graber, S.Gerrish, C.Wang, D.Blei. Reading tea leaves: how humans interpret topic models // Advances in Neural Information Processing Systems 22, 2009, pp. 288–296.

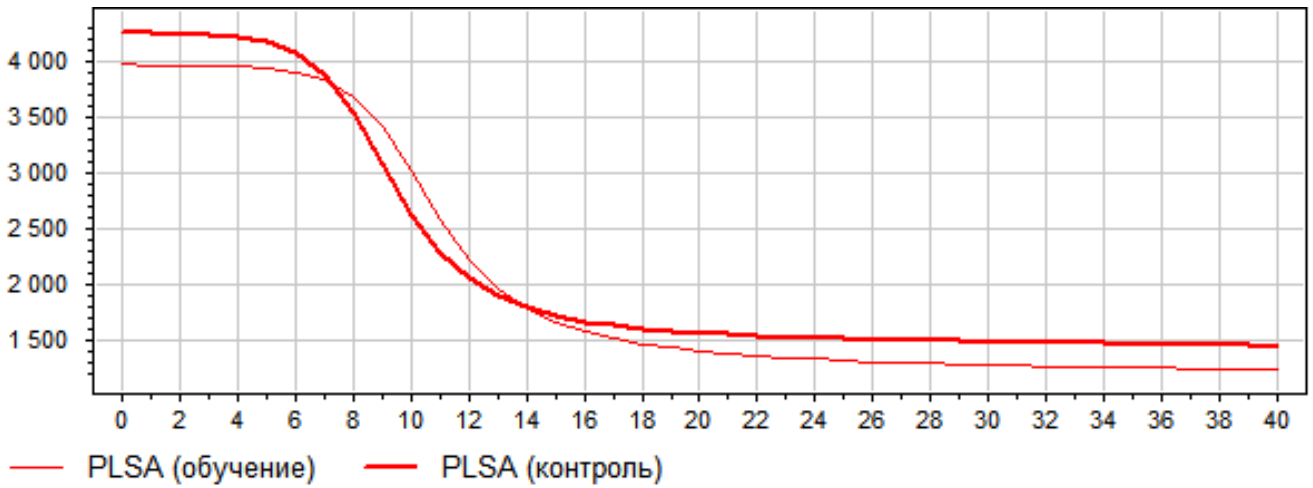
Методика эксперимента

D — коллекция 2000 авторефератов диссертаций на русском языке суммарной длины $n \approx 8.7 \cdot 10^6$, словарь $|W| \approx 3 \cdot 10^4$.

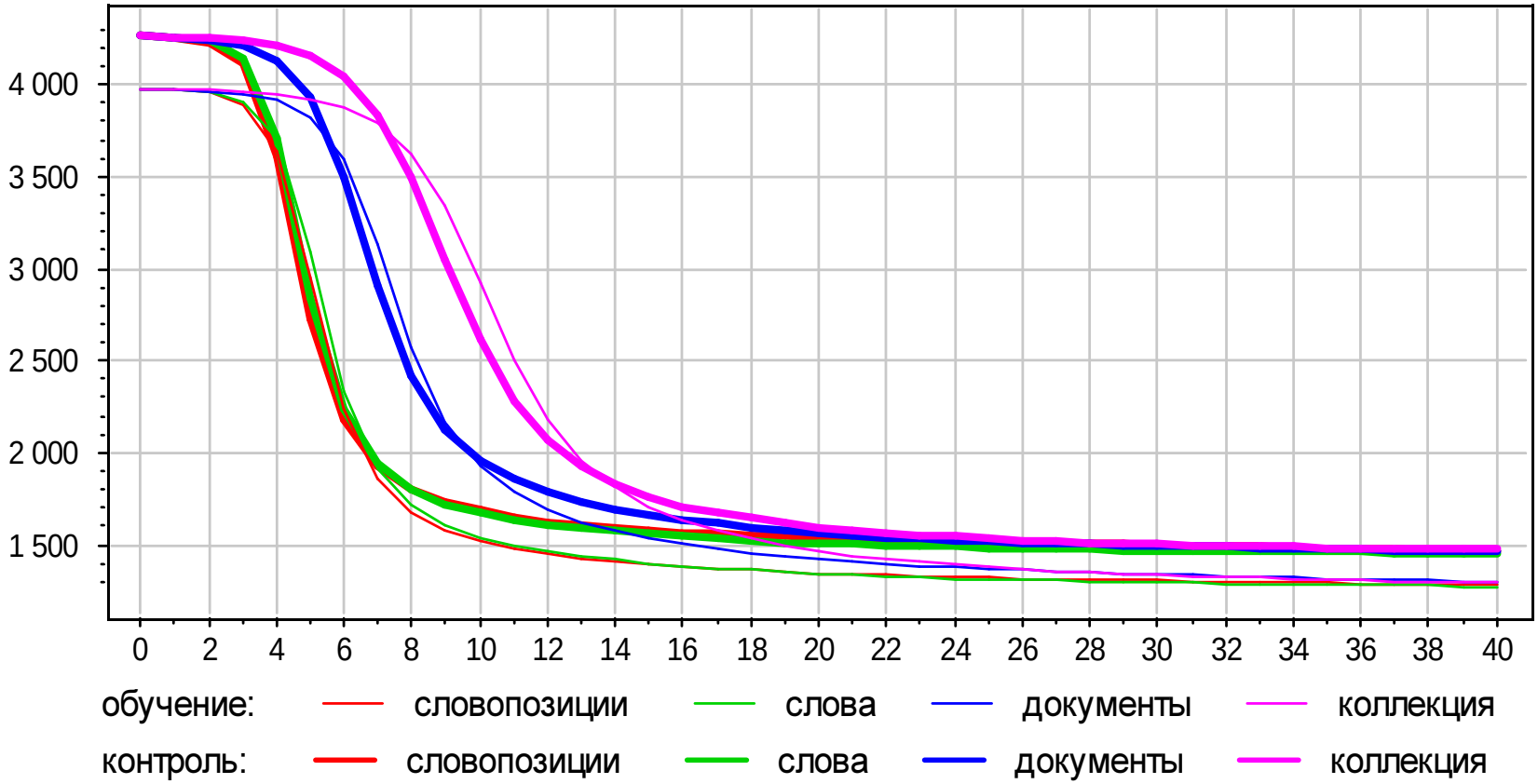
Предобработка: лемматизация, удаление стоп-слов.

D' — коллекция 200 авторефератов, не включённых в D.

Строятся графики зависимости перплексии от числа итераций (проходов коллекции); число итераций 40; число тем $|T| = 100$;



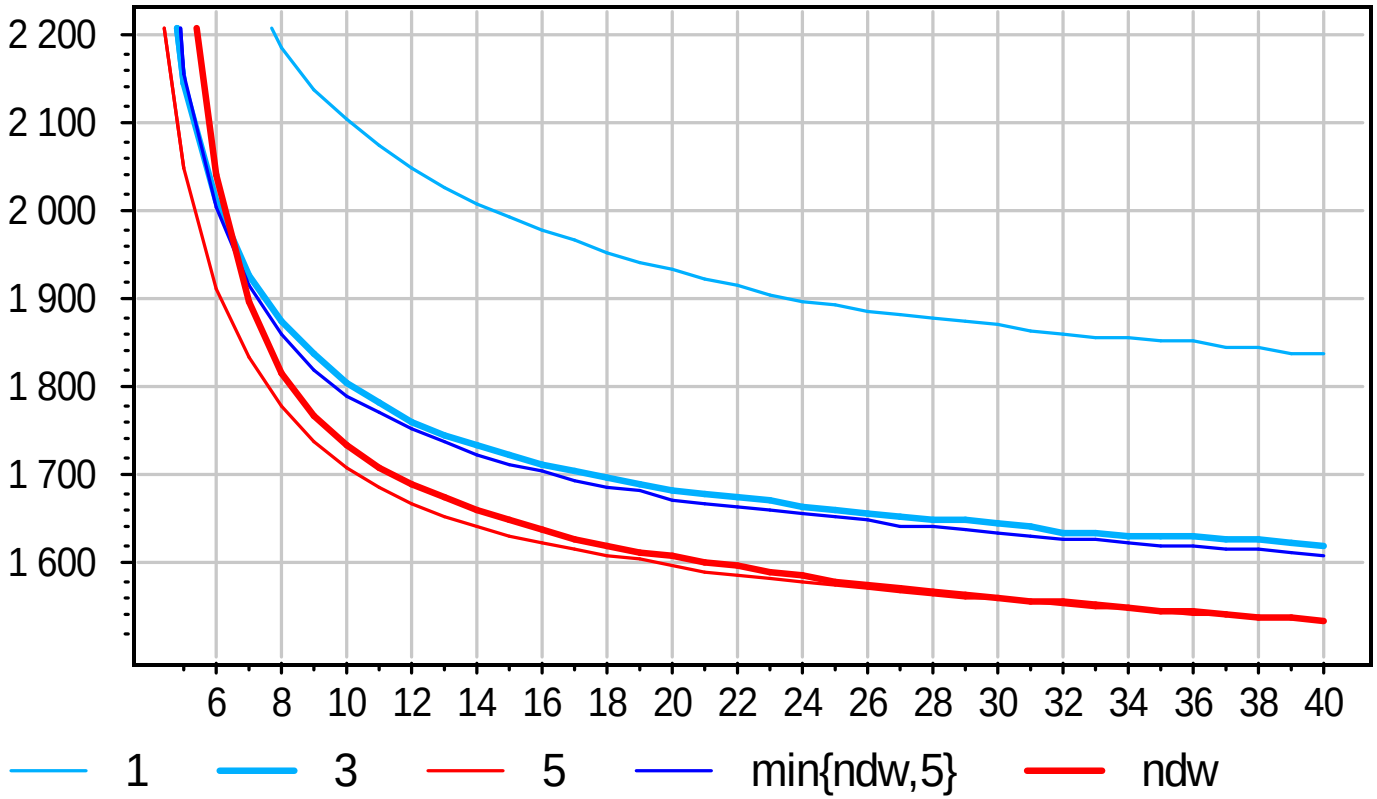
Частота обновления Φ и Θ не влияет на качество модели



Частота обновления параметров Φ и Θ не влияет на качество, а только на скорость сходимости.

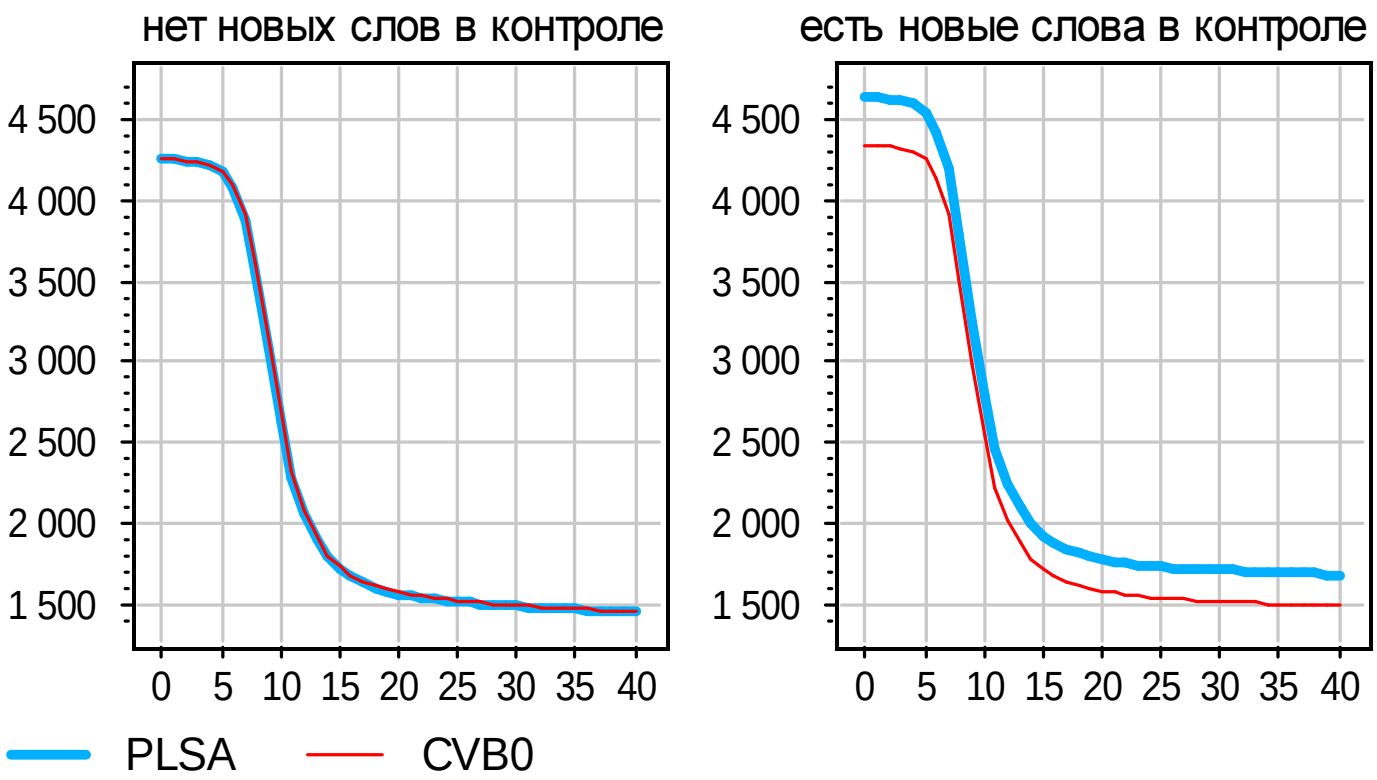
Вывод: лучше обновлять после каждого слова (d, w).

Сколько тем достаточно сэмплировать?



При сэмплировании пяти тем для каждой пары (d, w) перплексия не хуже, чем при сэмплировании n_{dw} тем.
Но одной или трёх тем не достаточно.

Регуляризация решает проблему новых слов, а не переобучения



PLSA без регуляризации, CVB0 с регуляризацией.

Вывод: регуляризация даёт преимущество только когда в контроле есть новые термины.

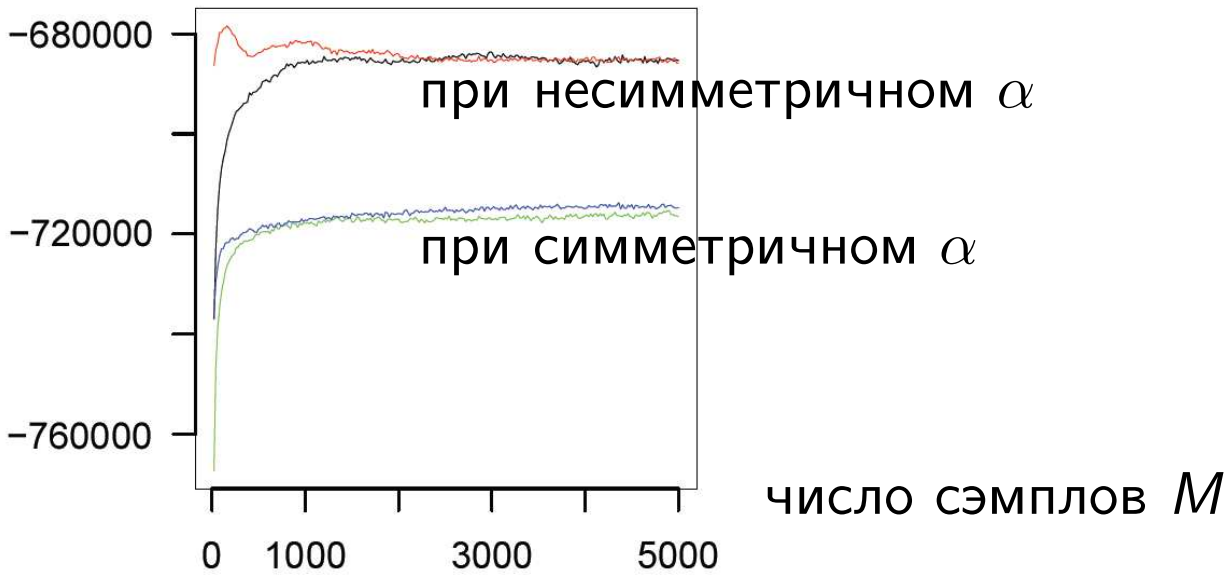
Проблема выбора гиперпараметров α и β

Стандартная рекомендация [2004]: $\alpha_t = 50/|T|$, $\beta_t = 0.01$.

Выводы по результатам более тонкого исследования [2009]:

- $p(t|d) \sim \mathrm{Dir}(\theta; \alpha)$, оптимизировать $\alpha = (\alpha_1, \ldots, \alpha_T)$.
- $p(w|t) \sim \mathrm{Dir}(\phi; \beta)$, взять симметричное $\beta_1 = \cdots = \beta_T \ll 1$.

правдоподобие



Hanna Wallach, David Mimno, Andrew McCallum.
Rethinking LDA: why priors matter. Neural Information Processing Systems, 2009.

Оптимизация гиперпараметра α

Обоснованность (evidence) модели на коллекции $X = (X_d)_{d \in D}$:

$$P(X|\alpha) = \int P(X|\theta) p(\theta|\alpha)\, d\theta =$$

$$= \prod_{d \in D} \frac{\Gamma(\alpha_0)}{\Gamma(n_d + \alpha_0)} \prod_{t \in T} \frac{\Gamma(n_{td} + \alpha_t)}{\Gamma(\alpha_t)} \to \max_{\alpha}$$

Метод неподвижной точки [Minka, 2003] — итерационный процесс, встраиваемый между проходами по всей коллекции:

$$\alpha_t := \alpha_t \frac{\sum_d \psi(n_{td} + \alpha_t) - \psi(\alpha_t)}{\sum_d \psi(n_d + \alpha_0) - \psi(\alpha_0)},$$

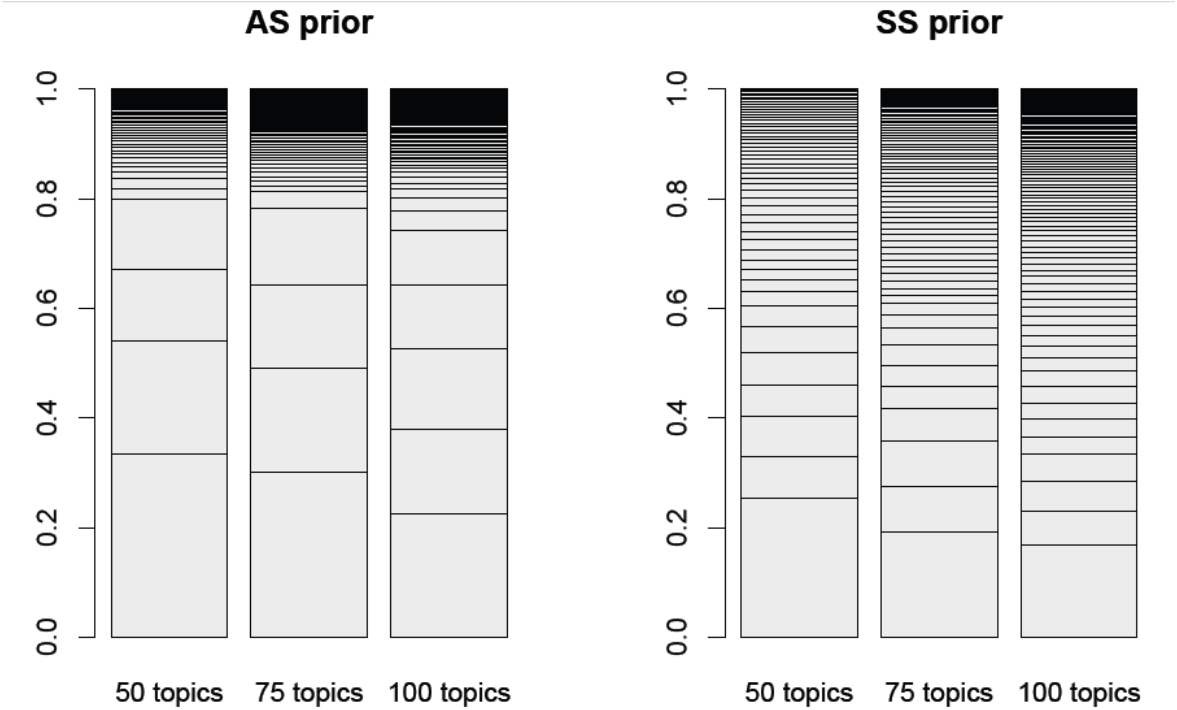
где $\psi(z) = \big(\ln \Gamma(z)\big)' = \Gamma'(z)/\Gamma(z)$ — дигамма-функция.

Более быстрые и точные методы оптимизации α:

Hanna Wallach. Structured Topic Models for Language.
PhD thesis, University of Cambridge, 2008.

Преимущество оптимизации гиперпараметра α

- Правдоподобие существенно выше.
- Сходимость быстрее, сэмплов нужно намного меньше.
- Меньшая чувствительность к избыточному $|T|$.
- Меньшее дробление тематики (это хорошо или плохо?):



(Мифы про) преимущества LDA перед PLSA

- Число параметров модели $|T| + |W|$ не зависит от $|D|$; при $p(w|t) \sim \mathrm{Dir}(\beta)$ число параметров $|T| + 1$.
- LDA строит более разреженную тематическую модель.
- LDA оценивает $p(t|d)$ нового документа тем же алгоритмом, что и документы обучающей коллекции.
- LDA меньше переобучается.

Реальное преимущество:

- LDA адекватнее оценивает вероятности новых и редких слов.

David Blei, Andrew Ng, Michael Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.

Эффективная реализация LDA — в проекте Vowpal Wabbit:
http://hunch.net/~vw/

Модель с фоновой и шумовой компонентами SWB — Special Words with Background [Steyvers et al. 2006]

Гипотеза: каждое употребление термина в документе объясняется либо темой, либо специфично для данного документа (шум), либо это общеупотребительный термин (фон).

Модель смеси тематической, шумовой и фоновой компонент:

$$p(w|d) = \frac{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}{1 + \gamma + \varepsilon}; \qquad Z_{dw} = \sum_{t \in T} \phi_{wt}\theta_{td},$$

$\pi_{dw} \equiv p_{\text{ш}}(w|d)$ — *шумовая* компонента, γ — параметр;
$\pi_w \equiv p_{\text{ф}}(w)$ — *фоновая* компонента, ε — параметр.

Требуется найти ϕ_{wt}, θ_{td}, π_{dw}, π_w для всех $d \in D$, $w \in W$, $t \in T$.

Chemudugunta C., Smyth P., Steyvers M. Modeling general and specific aspects of documents with a probabilistic topic model // Advances in Neural Information Processing Systems, MIT Press, 2006. — Vol. 19. — Pp. 241–248.

EM-алгоритм для робастной модели

E-шаг: вероятности тем, фона и шума для каждого (d, w):

$$H_{dwt} = \frac{\phi_{wt}\theta_{td}}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}, \quad t \in T;$$

$$H_{dw} = \frac{\gamma\pi_{dw}}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}; \quad H'_{dw} = \frac{\varepsilon\pi_w}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}.$$

M-шаг: решение задачи максимизации правдоподобия

ϕ_{wt}, θ_{td} — вычисляются по формулам PLSA;

$$\pi_w = \frac{\nu'_w}{\nu'}; \quad \nu'_w = \sum_{d \in D} n_{dw} H'_{dw}; \quad \nu' = \sum_{w \in W} \nu'_w;$$

$$\pi_{dw} = \frac{n_{dw} H_{dw}}{\nu_d}; \quad \nu_d = \sum_{w \in d} n_{dw} H_{dw};$$

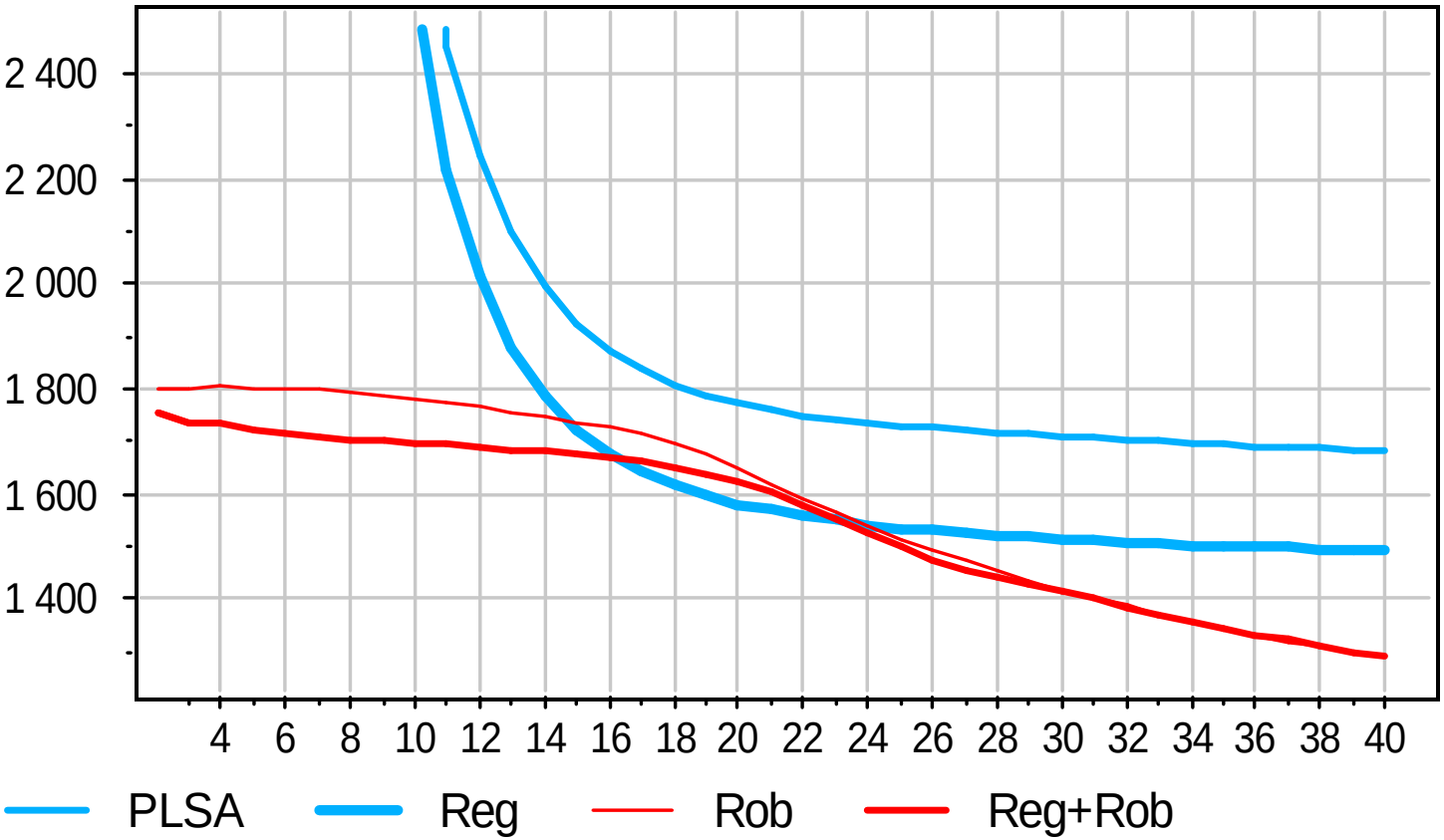
Альтернативный способ оценивания π_{dw} на M-шаге

В робастной модели возможно аналитическое выражение π_{dw} через остальные переменные без вычисления H_{dw},
назовём его *аддитивным M-шагом для шумовой компоненты*:

$$\pi_{dw} = \left(\frac{n_{dw}}{\nu_d} - \frac{Z_{dw} + \varepsilon\pi_w}{\gamma} \right)_+,$$

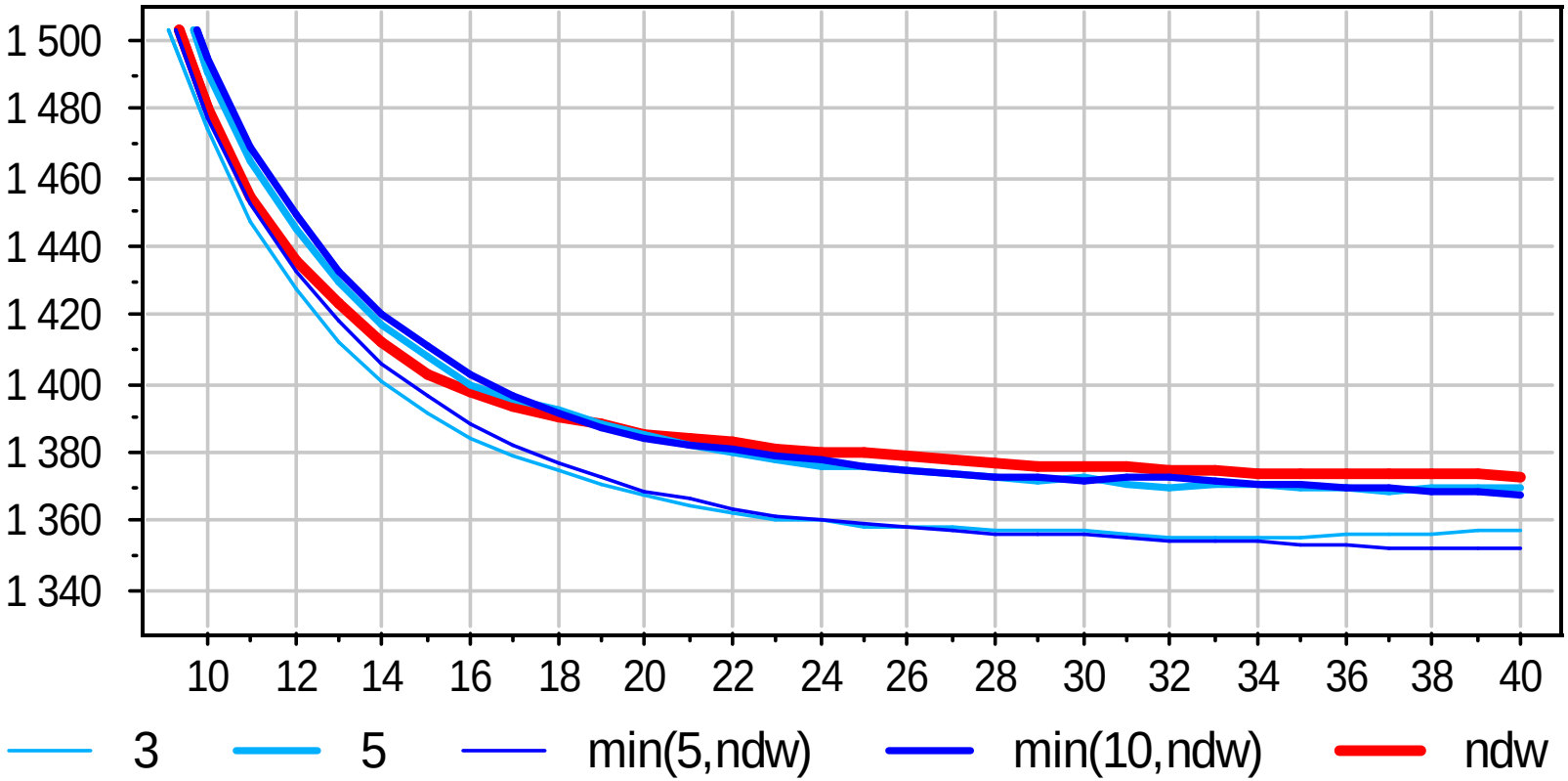
Таким образом, если термин w в документе d встречается существенно чаще, чем предсказывают тематическая и фоновая компоненты модели, то его появление объясняется особенностями данного документа, и тогда $\pi_{dw} > 0$.

Робастная модель не нуждается в регуляризации



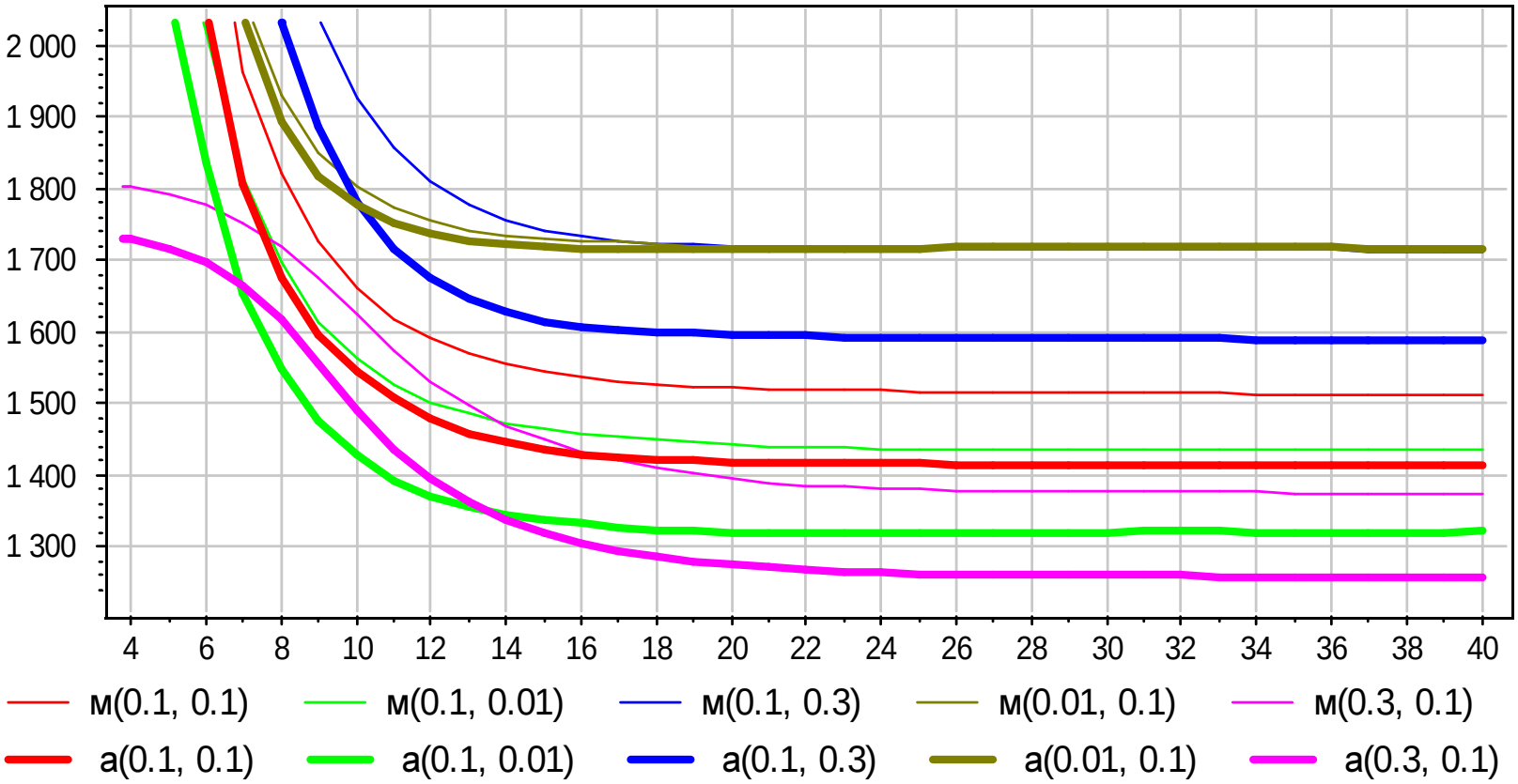
Робастность сильнее уменьшает перплексию PLSA, чем регуляризация. Регуляризация не улучает робастную модель.

Экономное сэмплирование для робастной модели



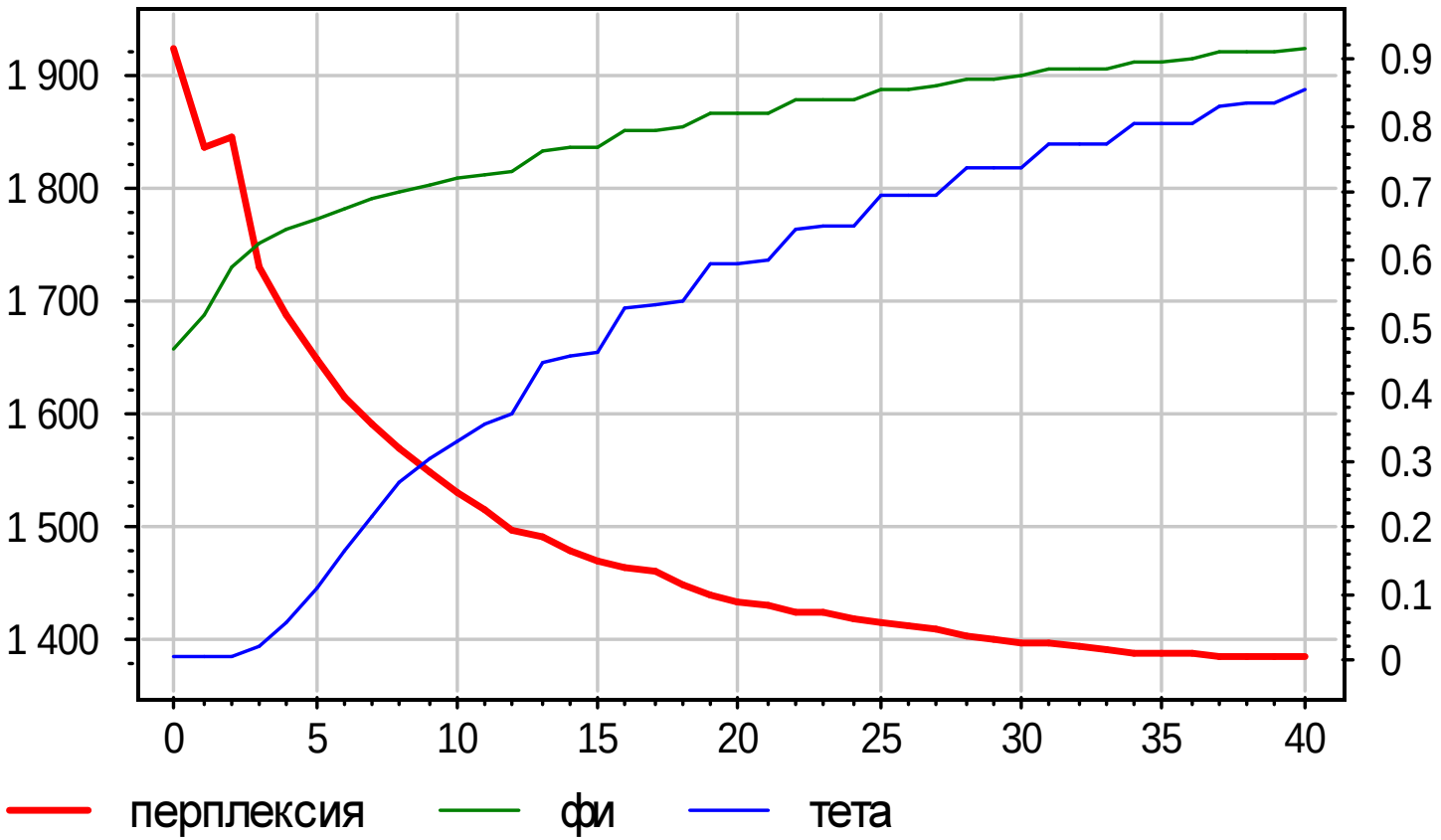
Робастная модель менее чувствительна к выбору числа сэмплируемых тем. Оптимум при $\min\{5, n_{dw}\}$.

Аддитивный и мультипликативный M-шаг



Аддитивный M-шаг a(ε, γ) лучше мультипликативного м(ε, γ).
Перплексия чувствительна к выбору параметров (ε, γ).
Для данной коллекции оптимум при $(\varepsilon, \gamma) = (0.3, 0.1)$.

Робастная модель допускает принудительное разреживание



В процессе разреживания доля нулевых ϕ_{wt} и θ_{td} (отложена по правой оси) увеличивается при монотонном уменьшении перплексии.

Преимущества робастной модели

- Не требуется регуляризация, следовательно
  — используются только несмещённыне оценки,
  — не надо настраивать гиперпараметры.
- Перплексия существенно лучше, чем у LDA.
- Новые слова естественно воспринимаются как шум,
  пока по ним не наберётся выборка,
  достаточная для определения тематики.
- Параметр γ и эвристика принудительного разреживания
  вместе позволяют управлять разреженностью.

Модели с оптимизацией числа тем

Два основных подхода:

- Оценивание качества кластеризации тем
- Использование иерархических процессов Дирихле HDP

Основная идея в обоих случаях схожа:
в процессе итераций темы могут создаваться и удаляться.

- **Если** $\phi_{wt}\theta_{td}$ малы для всех t, **то** создаётся новая тема.
- **Если** $n_t \leqslant \varepsilon$, **то** тема t удаляется.

Оценивание качества кластеризации тем

Проверка нулевой гипотезы *условной независимости* $p(w|d,t) = p(w|t)$ для темы t в документе d:

$$\hat{p}(w|t) = \frac{\hat{n}_{wt}}{\hat{n}_t}, \quad \hat{p}(w|d,t) = \frac{n_{dwt}}{\hat{n}_{dt}}, \quad t \in T,\ d \in D.$$

Критерий χ^2 Пирсона:

$$X^2_{dt} = \sum_{w \in W_d t} \frac{(E_{dwt} - n_{dwt})^2}{E_{dwt}},$$

$E_{dwt} = \hat{n}_{dt}\hat{p}(w|t)$ — ожидаемое число вхождений термина w;
$W_{dt} = \{w \in W : E_{dwt} > 0\}$.

Если $X^2_{dt} > \chi^2_{k,1-\alpha}$, то нулевая гипотеза отвергается ($k = |W_{dt}| - 1$).

Проблема разреженности: если $E_{dwt} < 5$ для большинства терминов w, то критерий χ^2 применять нельзя!

Перестановочный тест

1. Распределить n_t слов из распределения $\hat{p}(w|t)$ по документам n_{dt} случайным образом, N раз.
2. Вычислить N раз значения статистики X^2_{dt}.
3. Построить эмпирическое распределение статистики X^2_{dt} и найти её $(1-\alpha)$-квантиль.

Найденную квантиль использовать вместо $X^2_{dt} > \chi^2_{k,1-\alpha}$ для проверки гипотезы условной независимости, т.е. что «*тема t согласована в документе d* при уровне значимости α».

Далее можно оценивать *среднюю несогласованность* отдельных тем и документов.

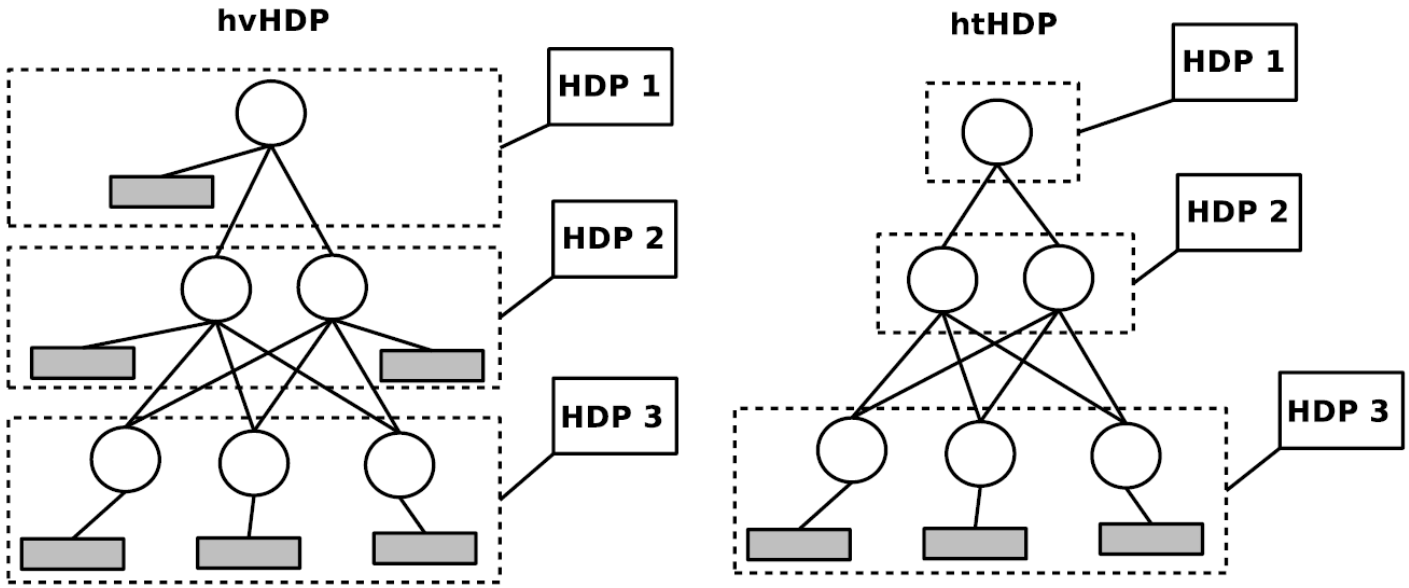
Проблема построения иерархических тематических моделей

- "Despite recent activity in the field of HPTMs, determining the hierarchical model that best fits a given data set, in terms of the structure and size of the learned hierarchy, still remains a challenging task and an open issue."
- "The evaluation of topic models is also an open issue."

E. Zavitsanos, G. Paliouras, G. A. Vouros. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes // Journal of Machine Learning Research 12 (2011) 2749-2775.

Две восходящие стратегии построения иерархии

- hvHDP: внутренние вершины — темы, имеющие $p(w|t)$
- htHDP: внутренние вершины — кластеры тем



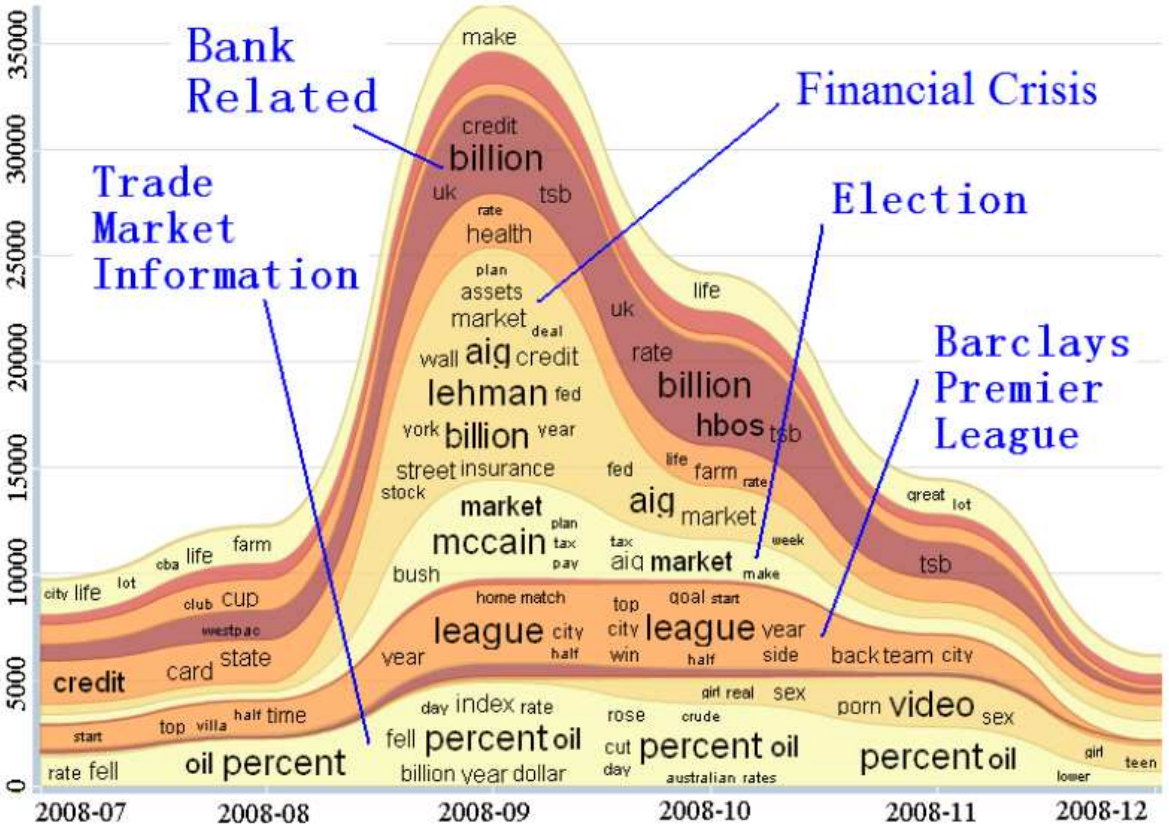
E. Zavitsanos, G. Paliouras, G. A. Vouros. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes // Journal of Machine Learning Research 12 (2011) 2749-2775.

Темпоральные модели (temporal topic models)

Основные предположения:

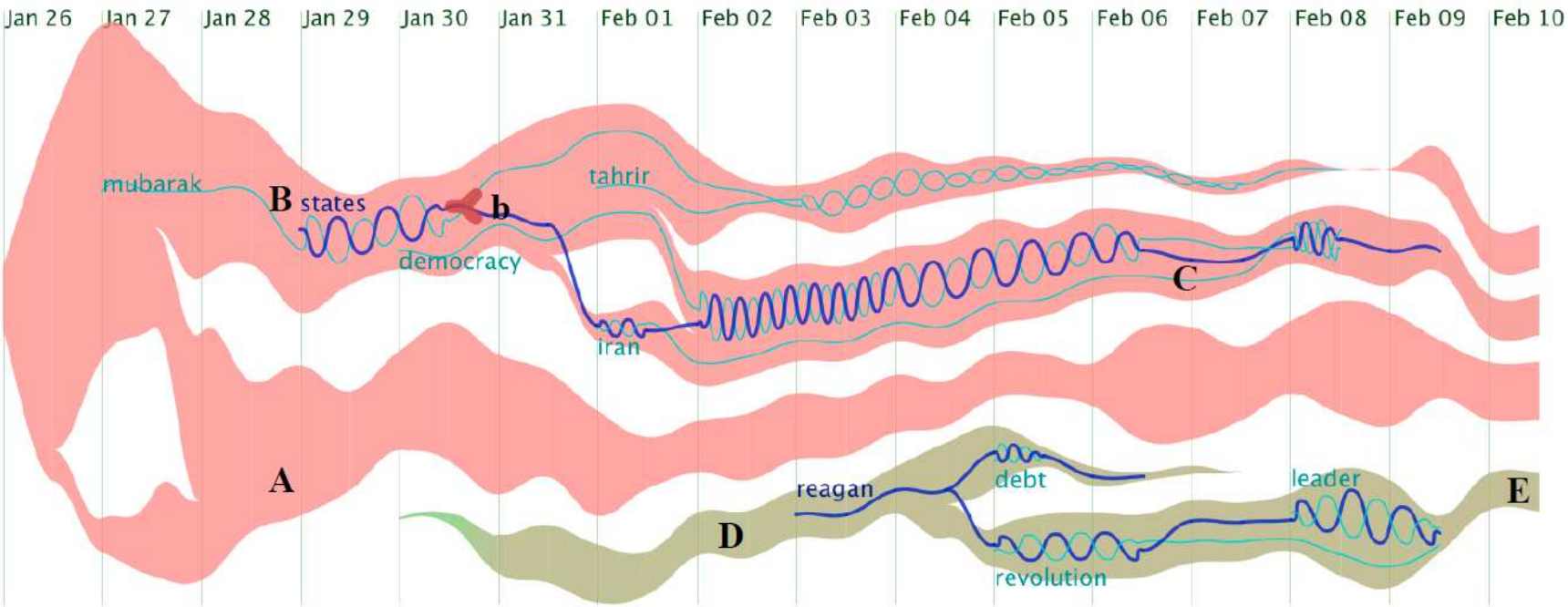
- Документы имеют привязку к моменту времени.
- Все распределения $p(w|t)$, $p(t|d)$ зависят от времени.
- Резкие изменения происходят редко.
- Темы могут
  — появляться;
  — исчезать;
  — сливаться;
  — расщепляться.
- Для обнаружения этих событий используются статистические критерии.

Пример темпоральной модели



Jianwen Zhang, Yangqiu Song, Changshui Zhang, Shixia Liu Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora // KDD'10, July 25–28, 2010.

Пример темпоральной модели



Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai J. Gao, Xin Tong, Huamin Qu TextFlow: Towards Better Understanding of Evolving Topics in Text // IEEE Transactions On Visualization And Computer Graphics, Vol. 17, No. 12, December 2011.

Другие обобщения и модификации тематических моделей

- Онлайновые (динамические) модели
- Author-topic models — пытаются приписать распределение авторов $p(a|w)$ каждому слову документа
- Entity-topic models — оценивают тематику связанных с текстами сущностей: людей, мест, организаций
- Модели связей между документами (ссылки, цитирование)
- Модели, учитывающие связи слов внутри документа
- Многоязыковые модели

Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad. Knowledge discovery through directed probabilistic topic models: a survey. Frontiers of Computer Science in China, Vol. 4, No. 2., 2010, Pp. 280–301. Русский перевод: http://www.machinelearning.ru/wiki/images/9/90/Daud2009survey-rus.pdf

Topic Modeling Bibliography:
http://www.cs.princeton.edu/∼mimno/topics.html