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)»

4 декабря 2013

Содержание

Задача определения тематики коллекции документов

Тема — это набор терминов, неслучайно часто совместно встречающихся в относительно узком подмножестве 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

Возможные дополнительные задачи:

— определить число тем

— восстановить иерархию тем

— построить динамику развития тем во времени

— найти тематику связанных с документами объектов...

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 (key phrase extraction) (сводится к задаче классификации или ранжирования)
- Удаление стоп-слов и слишком редких слов

Вероятностная формализация постановки задачи

Базовые предположения:

- каждое слово в документе связано с некоторой темой $t \in T$
- $D \times W \times T$ — дискретное вероятностное пространство
- коллекция D — выборка троек $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- гипотеза условной независимости: $p(w|d,t) = p(w|t)$

Вероятностная модель порождения документа d:

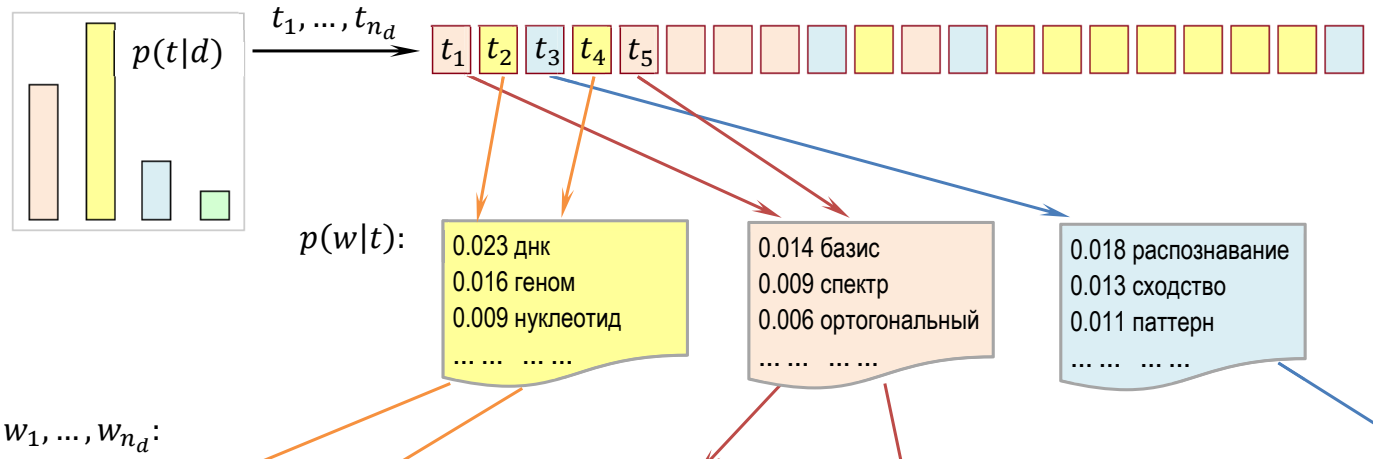
$$p(w|d) = \sum_{t \in T} p(w|d,t)\, p(t|d) = \boxed{\sum_{t \in T} p(w|t)\, p(t|d)}$$

Дано $\hat{p}(w|d) \equiv n_{dw}/n_d$, **найти:**

- $\phi_{wt} \equiv p(w|t)$ — распределение терминов в темах $t \in T$;
- $\theta_{td} \equiv p(t|d)$ — распределение тем в документах $d \in D$.

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).

Вероятностная модель порождения текстовых документов

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 $p(w|t)$;

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;

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

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

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

Принцип максимума правдоподобия

Правдоподобие — это плотность распределения выборки D:

$$p(D) = \prod_{i=1}^{n} p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}},$$

где n_{dw} — число вхождений термина w в документ d.

Пусть $p(w|d, \alpha)$ — параметрическая вероятностная модель документа d, зависящая от вектора параметров α.

Логарифм правдоподобия выборки D:

$$\log p(D, \alpha) = \sum_{d \in D} \sum_{w \in d} n_{dw} \log p(w|d, \alpha) p(d) \to \max_{\alpha}.$$

Избавимся от $p(d)$, не влияющего на точку максимума:

$$\mathscr{L}(D, \alpha) = \sum_{d \in D} \sum_{w \in d} n_{dw} \log p(w|d, \alpha) \to \max_{\alpha}.$$

Униграммные модели порождения текстовых документов

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: $p(w|d) = \xi_w$ для всех d

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

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

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

Недостатки униграммной модели. Модель смеси униграмм

- темы не выявляются, это *не тематическая* модель языка
- параметров в $\{\xi_w\}$ слишком мало, в $\{\xi_{dw}\}$ слишком много
- 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]

Задача: найти максимум правдоподобия

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

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 $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$:

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

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

Задача: найти максимум правдоподобия

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

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: стохастическое 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:

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

M-шаг: решение задачи максимизации правдоподобия выражается аналитически через частотные оценки условных вероятностей, если положить $\hat{n}_{dwt} = n_{dw}p(t|d,w)$:

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

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

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

Вывод формулы M-шага для ϕ_{wt}

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

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

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

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

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

Вывод формулы M-шага для θ_{td}

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

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

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

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

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

Недостатки PLSA

1. необходимость хранить 3D-матрицу $p(t|d,w)$,
   медленная сходимость на больших коллекциях
   — рациональный, стохастический, онлайновый алгоритмы
2. неединственность и неустойчивость решения,
   на малых коллекциях возможно переобучение
   — регуляризации: сглаживание, разреживание,
   учёт дополнительной внешней информации
3. нет выделения нетематических слов
   — робастные модели с шумом и фоном
4. нет управления *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 $\theta_{td} = 0$)
   — регуляризации, постепенное разреживание

Рациональный EM-алгоритм

Идея: E-шаг встраивается внутрь M-шага

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

инициализация ϕ_{wt}, θ_{td} для всех $d \in D,\ w \in W,\ t \in T$;
для всех итераций $i = 1, \ldots,$ MaxIter
  $n_{wt}, n_{dt}, n_t, n_d := 0$ для всех $d \in D,\ w \in W,\ t \in T$;
  для всех $d \in D,\ w \in d$
    $p(t|d,w) \propto \phi_{wt}\theta_{td}$ для всех $t \in T$;
    возможно, применить разреживание к $p(t|d,w)$;
    $n_{wt}, n_{dt}, n_t, n_d \mathrel{+}= n_{dw}p(t|d,w)$ для всех $t \in T$;
  $\phi_{wt} := n_{wt}/n_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := n_{dt}/n_d$ для всех $d \in D,\ t \in T$;

Онлайновый EM-алгоритм для модели PLSA

инициализировать ϕ_{wt} для всех $w \in W,\ t \in T$;
$n_{wt} := 0,\ n_t := 0$ для всех $w \in W,\ t \in T$;
для всех пакетов $D_j,\ j = 1, \ldots, J$
  $\tilde{n}_{wt} := 0,\ \tilde{n}_t := 0$ для всех $w \in W,\ t \in T$;
  для всех $d \in D_j$
    инициализировать θ_{td} для всех $t \in T$;
    повторять
      $p(t|d,w) \propto \phi_{wt}\theta_{td}$ для всех $t \in T$;
      $\theta_{td} := \frac{1}{n_d} \sum_{w \in d} n_{dw} p(t|d,w)$ для всех $t \in T$;
    пока θ_d не сойдётся;
    $\tilde{n}_{wt}, \tilde{n}_t \mathrel{+}= n_{dw} p(t|d,w)$ для всех $w \in d,\ t \in T$;
  $n_{wt} := \rho_j n_{wt} + \tilde{n}_{wt};\ \ n_t := \rho_j n_t + \tilde{n}_t$ для всех $w \in W,\ t \in T$;
  $\phi_{wt} := n_{wt}/n_t$ для всех $w \in W,\ t \in T$;

Стохастический EM-алгоритм SEM

Гипотеза разреженности $p(t|d,w)$: «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 $p(t|d,w)$.

Вместо $p(t|d,w)$ используется его несмещённая эмпирическая оценка по сгенерированной выборке длины s:

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

Варианты обновления параметров ϕ_{wt}, θ_{td}: после каждого прохода коллекции / документа / слова / вхождения слова

Преимущества:
разреженность $p(t|d,w)$, ускорение сходимости

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

Методика эксперимента

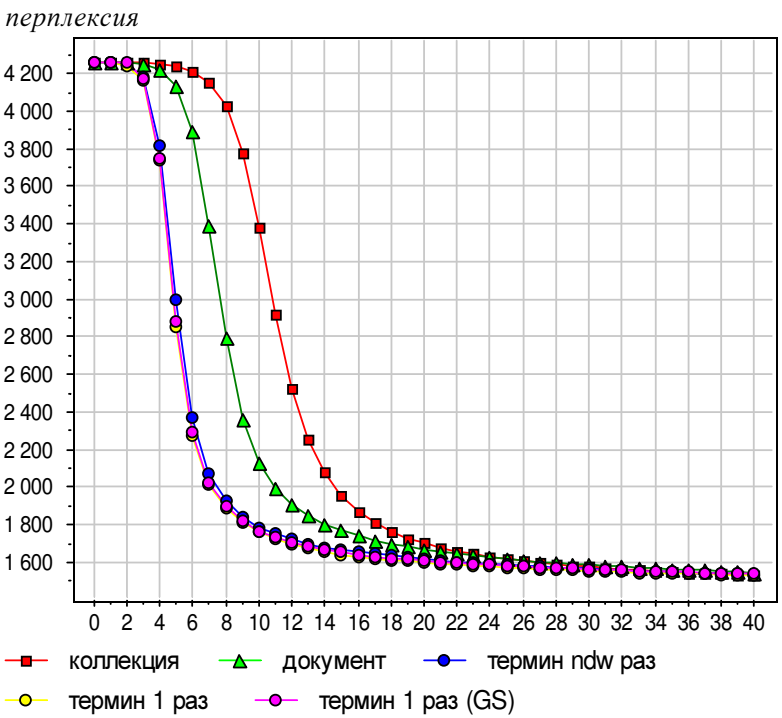
Использовались две коллекции:

- NIPS:
  - $|D| = 1566$ статей конференции NIPS на английском языке;
  - суммарной длины $n \approx 2.3 \cdot 10^6$,
  - словарь $|W| \approx 1.3 \cdot 10^4$.
  - Контрольная коллекция: $|D'| = 174$.
- RuDis:
  - $|D| = 2000$ авторефератов диссертаций на русском языке;
  - суммарной длины $n \approx 8.7 \cdot 10^6$,
  - словарь $|W| \approx 3 \cdot 10^4$.
  - Контрольная коллекция: $|D'| = 200$.

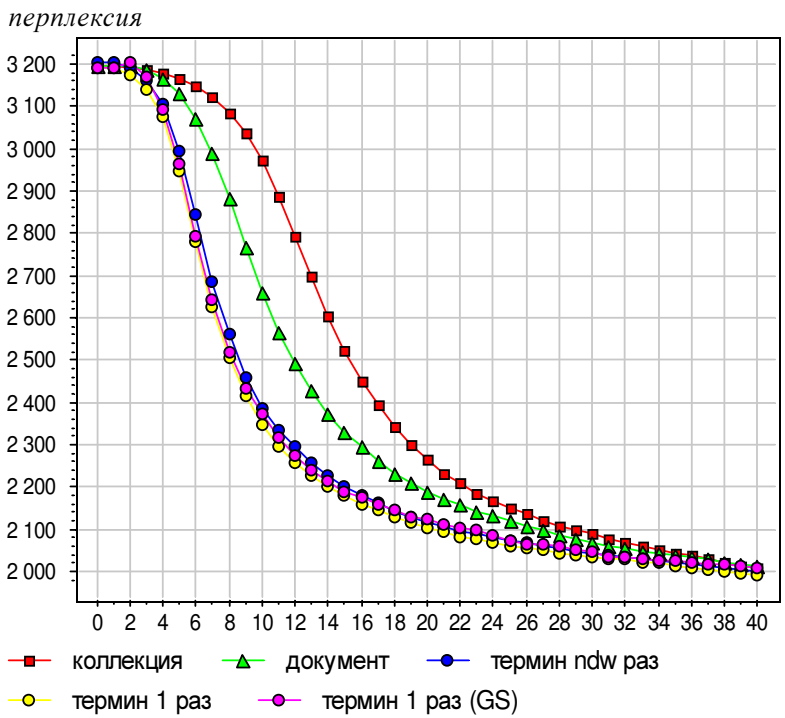
Предобработка: лемматизация, удаление стоп-слов.

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



RuDis

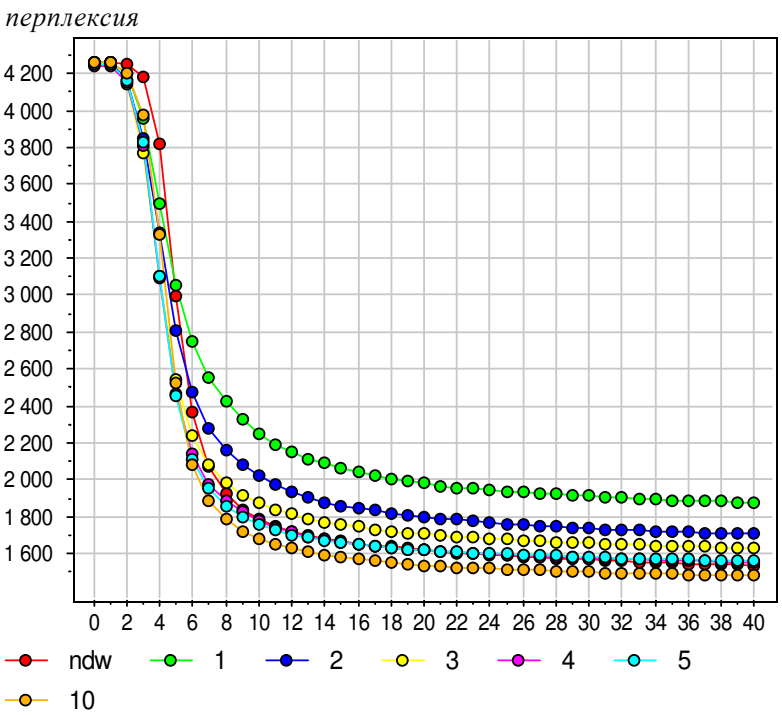


NIPS

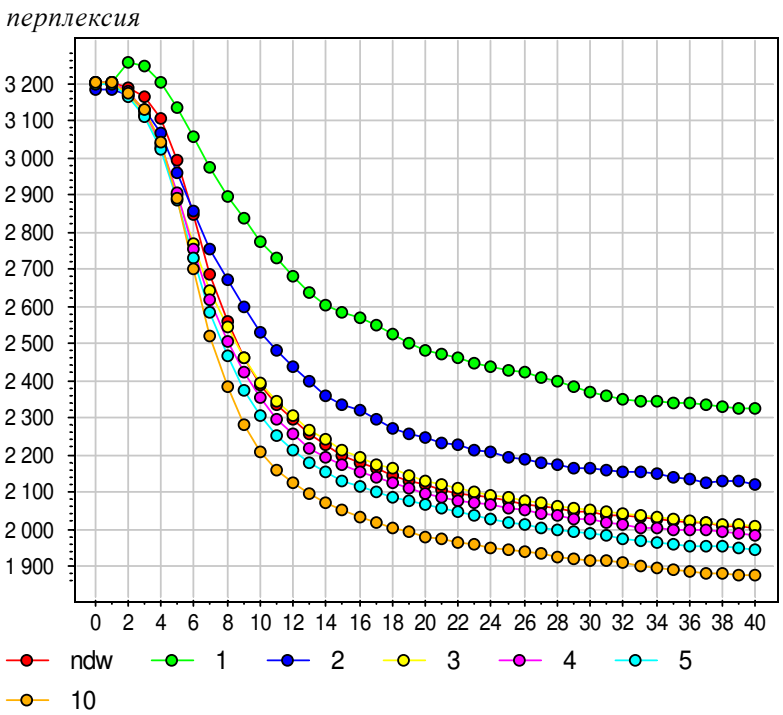
Частота обновления параметров Φ и Θ не влияет на качество, но влияет на скорость сходимости.

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

Сколько тем достаточно сэмплировать для $\hat{p}(t|d,w)$?



RuDis



NIPS

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

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

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

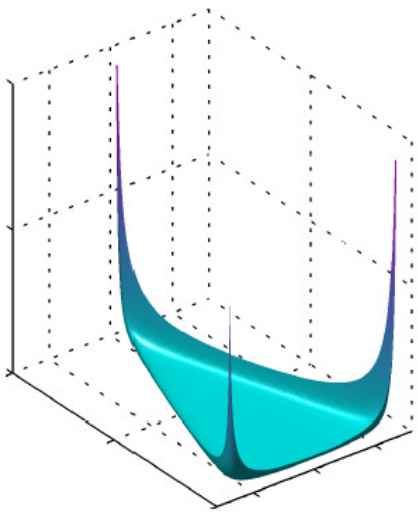
Гипотеза об априорных распределениях Дирихле:

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

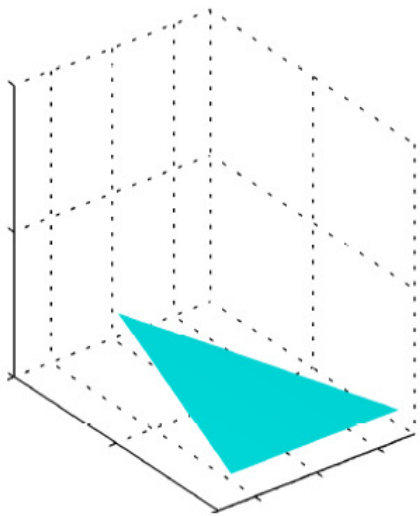
Почему именно Дирихле?

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

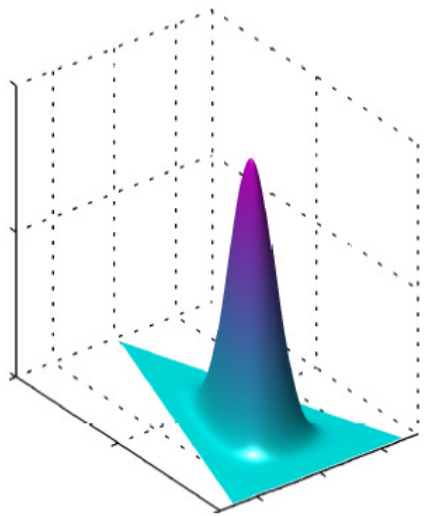
Пример. $\mathrm{Dir}(\theta|\alpha)$, $\theta = (\theta_1, \theta_2, \theta_3)$, $|T| = 3$:



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



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



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

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

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

апостериорное распределение также из $\text{Dir}(\alpha')$, $\alpha'_t = \alpha_t + n_{td}$.
Распределение Дирихле — *сопряжённое* к мультиномиальному.
Это 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:

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

Алгоритм сэмплирования Гиббса [Griffiths, Steyvers, 2004]

Итак, 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

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

Gibbs Sampling (GS) — регуляризованный стохастический EM-алгоритм с максимально частым обновлением параметров.

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: $\tilde{p}(t|d,w) := \frac{n_{wt}+\beta_w}{n_t+\beta_0} \cdot \frac{n_{td}+\alpha_t}{n_d+\alpha_0}$ 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 $d \in D$, $t \in T$;

Основные алгоритмы обучения параметров модели LDA

- Сэмплирование Гиббса (GS — Gibbs Sampling)
  можно рассматривать как специальный случай EM
  — 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 EM:
  — 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.

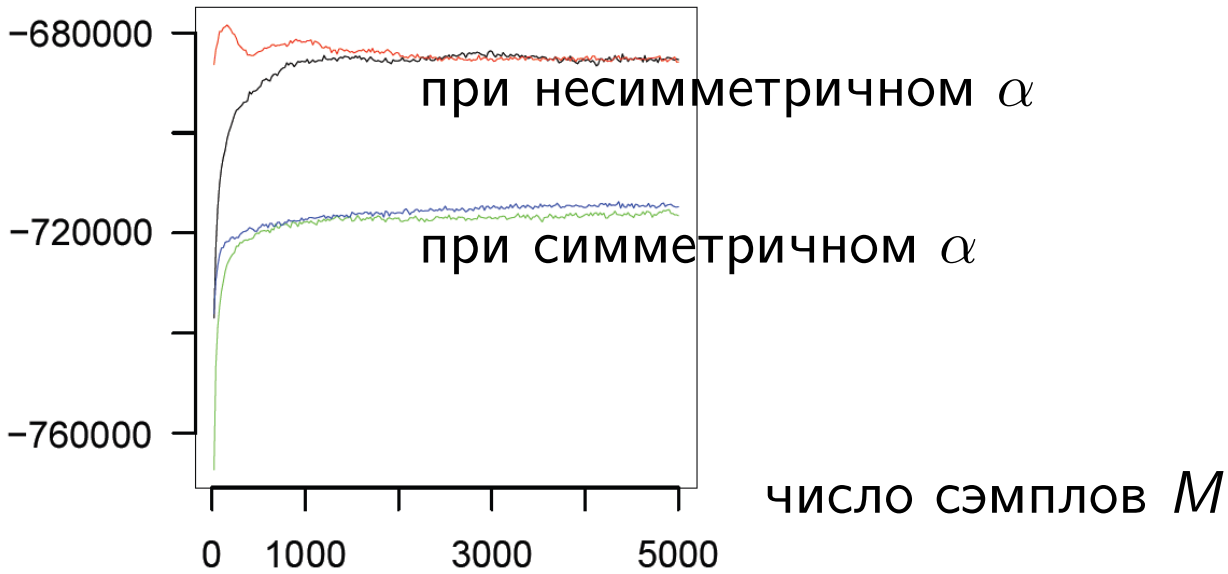
Проблема выбора гиперпараметров α и β

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

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:

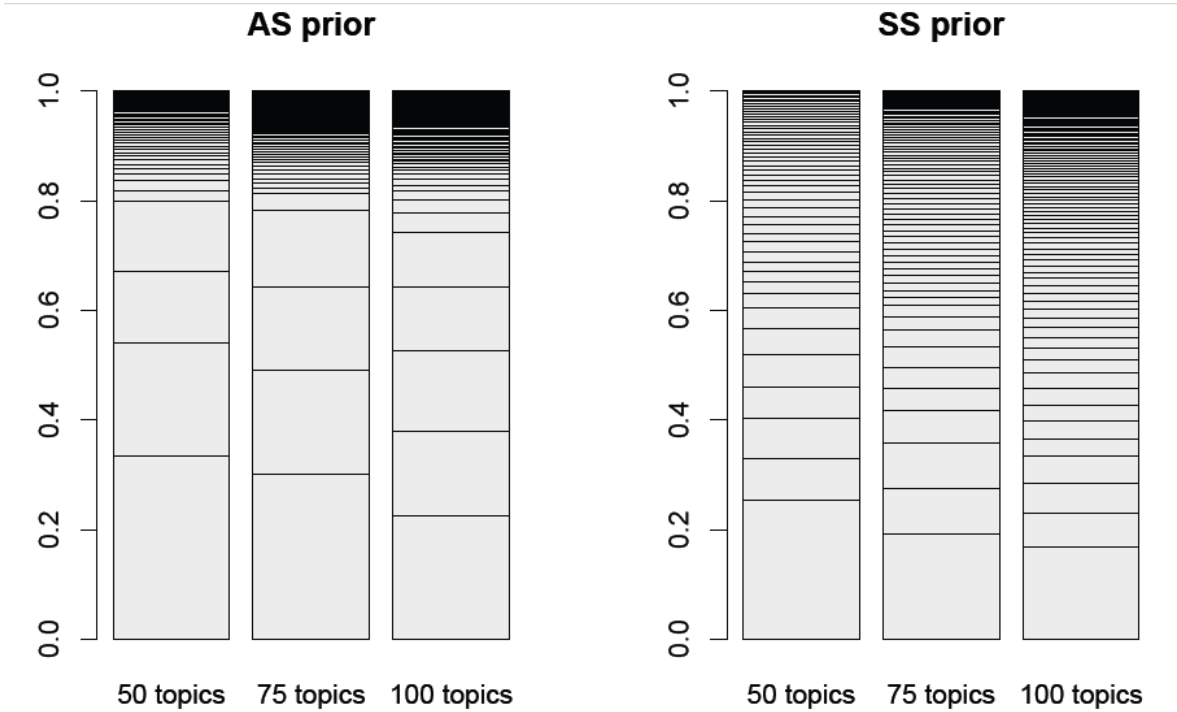
$$\alpha_t := \alpha_t \frac{\sum_d \psi(n_{td} + \alpha_t) - \psi(\alpha_t)}{\sum_d \psi(n_d + \alpha_0) - \psi(\alpha_0)},$$

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

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?):



Недостатки LDA

1. распределение Дирихле — слишком слабый регуляризатор, оно не имеет убедительных лингвистических обоснований
2. ограниченность байесовского вывода: он требует
   — либо громоздкого интегрирования по параметрам,
   — либо применения сопряжённых распределений
3. необходимость оптимизация гиперпараметров α, β
4. сложность совмещения многих требований в одной модели
5. неединственность и неустойчивость решения
6. нет выделения нетематических слов
7. нет существенного превосходства над PLSA по качеству модели на больших коллекциях

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

Гипотеза: каждый термин в документе (d, w)
— либо связан с какой-то темой t,
— либо специфичен для данного документа (шум),
— либо является общеупотребительным (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.

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 PLSA

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

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

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

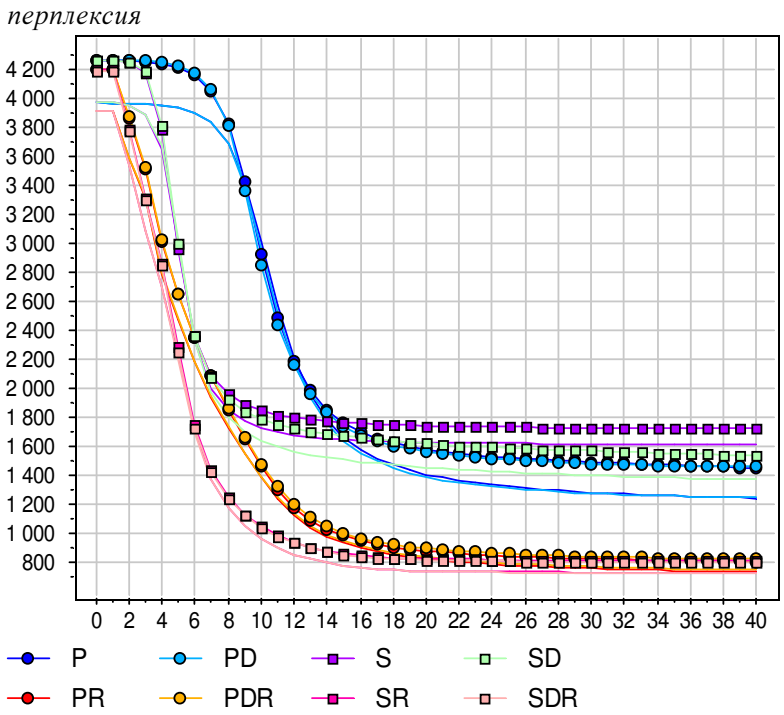
M-шаг: решение задачи максимизации правдоподобия

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

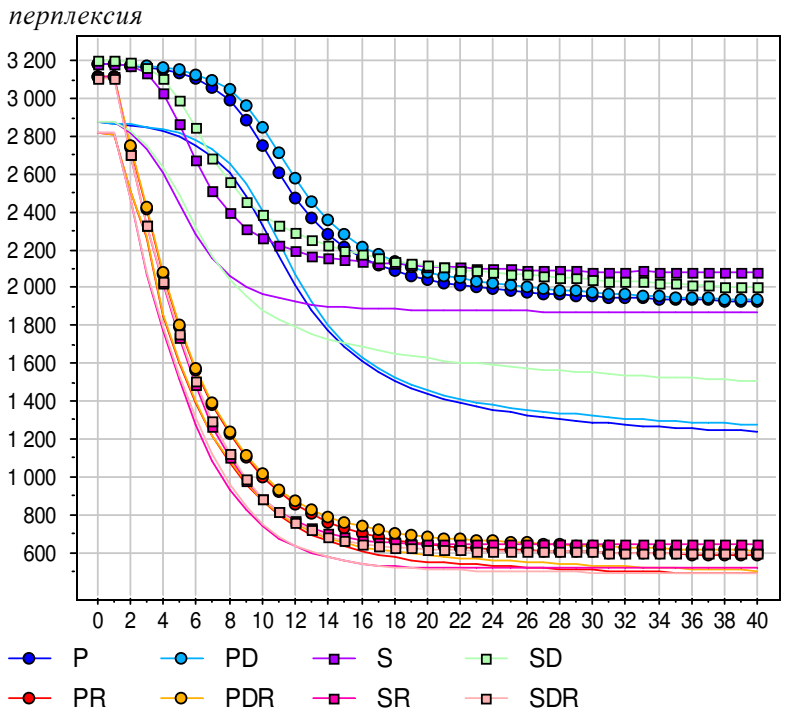
$$\pi_{dw} = \frac{n_{dw}H_{dw}}{\nu_d}; \quad \nu_d = \sum_{w \in d} n_{dw}H_{dw};$$

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

Эксперименты с робастными моделями



RuDis



NIPS

Обозначения: P – PLSA
D – регуляризация Дирихле ($\alpha_t = 0.5,\ \beta_w = 0.01$)
S – сэмплирование ($s = n_{dw}$)
R – робастность (шум $\gamma = 0.3$, фон $\varepsilon = 0.01$)

Выводы

Робастный PLSA лучше стандартного LDA

- перплексия ниже, переобучение меньше
- робастным алгоритмам не нужна регуляризация Дирихле

Недостатки робастного алгоритма с фоном и шумом

- невозможно оптимизировать параметры γ, ε
- приходится хранить матрицу шума $(\pi_{dw})_{D\times W}$

Potapenko A. A., Vorontsov K. V. Robust PLSA performs better than LDA // European Conf. on Inform. Retrieval ECIR-2013, Springer LNCS. Pp. 784–787.

Воронцов К. В., Потапенко А. А. Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование, 2012. — Т. 4, № 12. — С. 693–706.

Упрощённая робастная модель

Робастная тематическая модель с шумом и фоном:

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

Упрощённая робастная модель с шумом без фона:

$$p(w|d) = \nu_d Z_{dw} + \big[Z_{dw}=0\big]\frac{n_{dw}}{n_d}, \qquad \nu_d = \sum_{w\in d}\big[Z_{dw}>0\big]\frac{n_{dw}}{n_d}.$$

Упрощённая робастная модель с фоном без шума:

$$p(w|d) = \nu'_d Z_{dw} + \big[Z_{dw}=0\big]\frac{n_w}{n}, \qquad \nu'_d = \sum_{w\in W}\big[Z_{dw}>0\big]\frac{n_w}{n}.$$

Воронцов К. В., Потапенко А. А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных (www.jmlda.org), 2013. — Т. 1, № 6. — С. 657–686.

Резюме в конце лекции

Общий EM-алгоритм для PLSA и LDA с сочетанием эвристик:

- частота обновления параметров
- сэмплирование: нет / s раз / n_{dw} раз (Gibbs Sampling)
- сглаживание: нет-PLSA / есть-LDA
- робастность: с фоном и/или с шумом / упрощённая

«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»
http://www.MachineLearning.ru

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

Способны ли PLSA и LDA правильно восстанавливать темы?

Модельные коллекции порождаются заданными матрицами Φ_0 и Θ_0 при $|D| = 500$, $|W| = 1000$, $|T| = 30$, $n_d \in [100, 600]$.

Отклонение восстановленных распределений $p(i|j)$ от исходных модельных распределений $p_0(i|j)$ измеряются средним расстоянием Хеллингера:

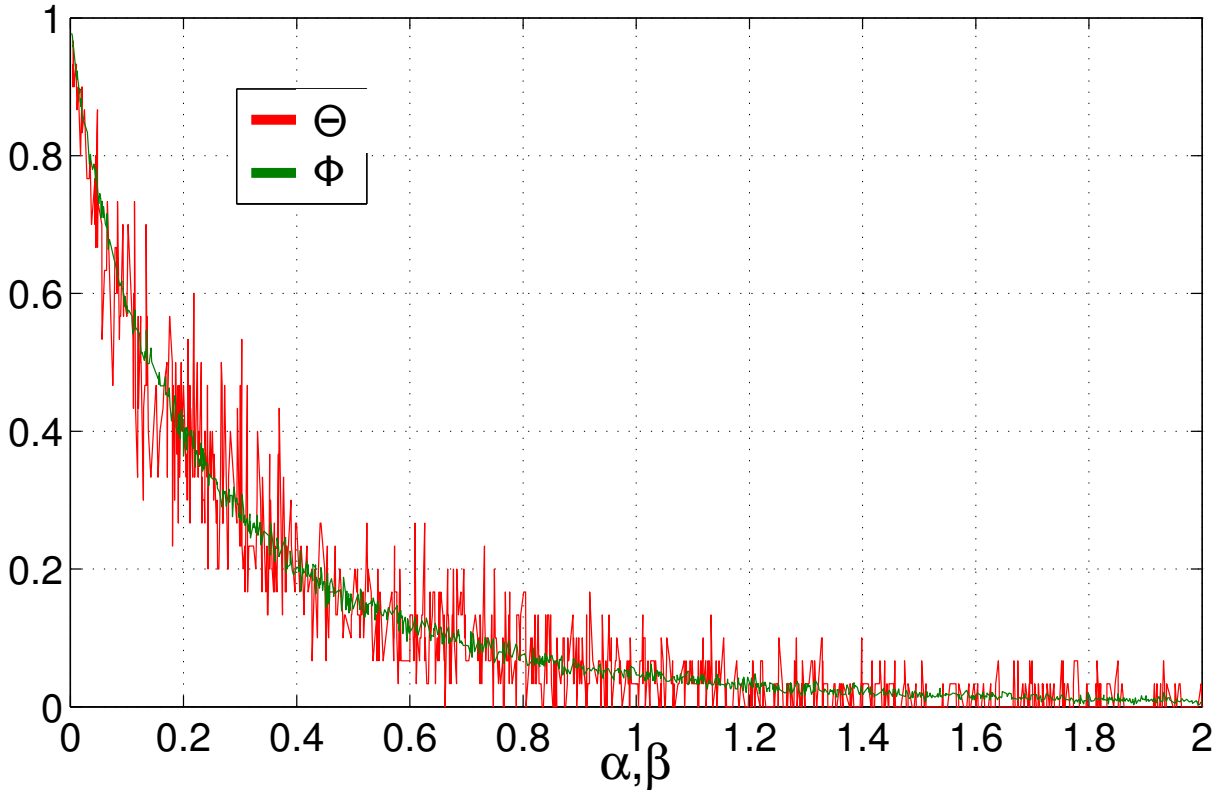
$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{p(i|j)} - \sqrt{p_0(i|j)} \right)^2},$$

как для самих матриц Φ и Θ, так и для их произведения:

$$D_\Phi(\Phi, \Phi_0) = H(\Phi, \Phi_0);$$
$$D_\Theta(\Theta, \Theta_0) = H(\Theta, \Theta_0);$$
$$D_{\Phi\Theta}(\Phi\Theta, \Phi_0\Theta_0) = H(\Phi\Theta, \Phi_0\Theta_0).$$

Генерация модельных данных различной степени разреженности

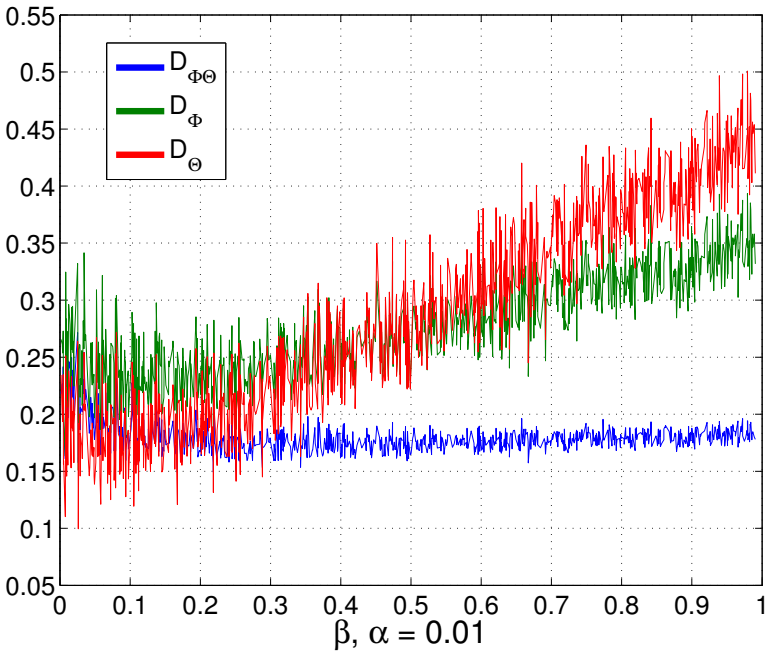
Зависимость разреженности (доли почти нулевых элементов) распределений $\theta_d^0 \sim \mathrm{Dir}(\alpha)$ и $\phi_t^0 \sim \mathrm{Dir}(\beta)$ от параметров α и β симметричного распределения Дирихле:



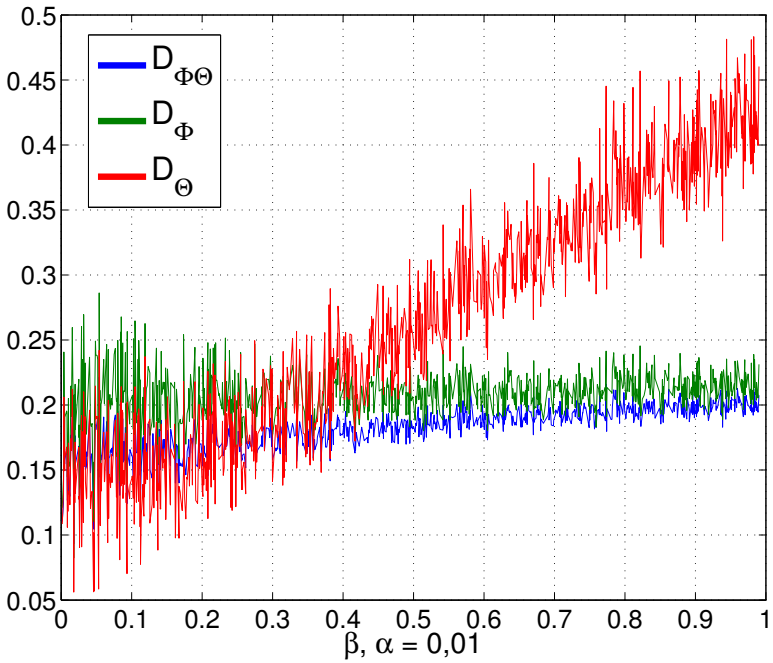
Эксперимент: неустойчивость восстановления Φ, Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Φ_0
(по 1000 случайных коллекций с $\alpha = 0.01$ и разными β)

PLSA



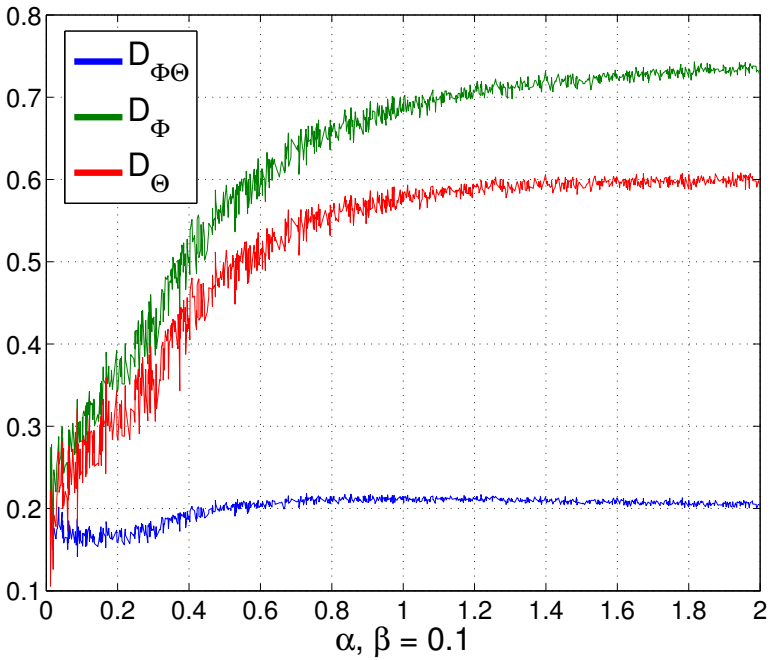
LDA



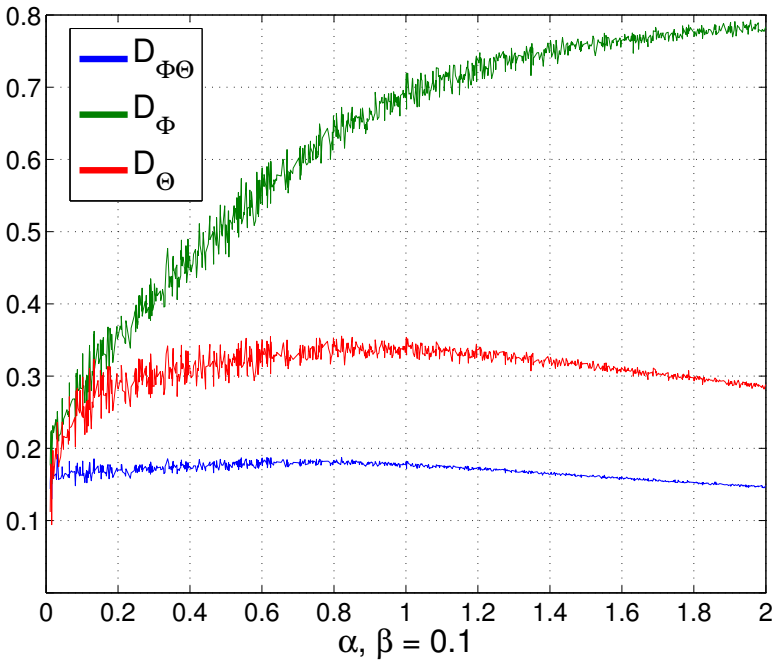
Эксперимент: неустойчивость восстановления Φ, Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Θ_0
(по 1000 случайных коллекций с $\beta = 0.1$ и разными α)

PLSA



LDA



Выводы

1. Произведение $\Phi\Theta$ восстанавливается устойчиво, точность восстановления не зависит от разреженности исходных модельных данных Φ_0, Θ_0
2. Матрицы Φ, Θ восстанавливаются неустойчиво, результат зависит от случайной инициализации
3. Методы PLSA и LDA одинаково неустойчивы (сглаживание не спасает от неединственности)
4. Устойчивое восстановление матриц Φ, Θ возможно только при сильной разреженности (более 80% нулей)

Реализация экспериментов:
Виталий Глушаченков. Магистерская диссертация. МФТИ. — 2013.
Михаил Колупаев. Курсовая работа. ВШЭ, ШАД Яндекс. — 2013.

Причина неустойчивости тематических моделей

Задача стохастического матричного разложения:

$$\hat{F} \approx F = \Phi\Theta$$

$\hat{F} = \left(n_{dw}/n_d\right)_{W\times D}$ — известная матрица исходных данных;
$F = \left(p(w|d)\right)_{W\times D}$ — матрица тематической модели;
$\Phi = \left(\phi_{wt}\right)_{W\times T}$ — искомая матрица терминов тем $\phi_{wt} = p(w|t)$;
$\Theta = \left(\theta_{td}\right)_{T\times D}$ — искомая матрица тем документов $\theta_{td} = p(t|d)$.
Все матрицы неотрицательные, с нормированными столбцами.

Проблема неединственности матричного разложения:

$$F = \Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

для любых невырожденных $S_{T\times T}$ таких, что $\Phi', \Theta' > 0$.

Регуляризация — это выбор лучшего из множества разложений

Обоснование EM-алгоритма PLSA

Теорема

Максимум правдоподобия

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

при ограничениях неотрицательности и нормировки

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

и фиксированных значениях $p(t|d,w)$ достигается при

$$\phi_{wt} \propto n_{wt} \equiv \sum_{d\in D} n_{dw} p(t|d,w) \qquad \theta_{td} \propto n_{dt} \equiv \sum_{w\in W} n_{dw} p(t|d,w)$$

Обоснование регуляризованного EM-алгоритма PLSA

Теорема

Максимум регуляризованного правдоподобия

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

при ограничениях неотрицательности и нормировки

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

и фиксированных значениях $p(t|d,w)$ достигается, когда

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+ \qquad \theta_{td} \propto \left(n_{dt} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+$$

Дивергенция Кульбака–Лейблера

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

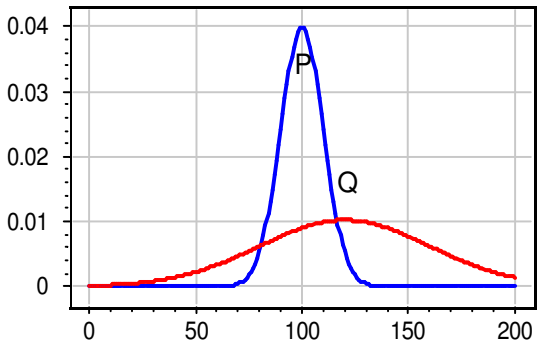
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

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

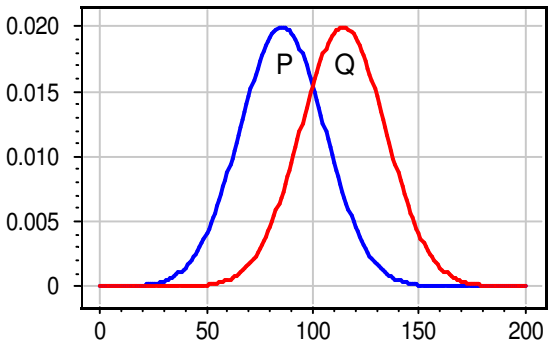
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

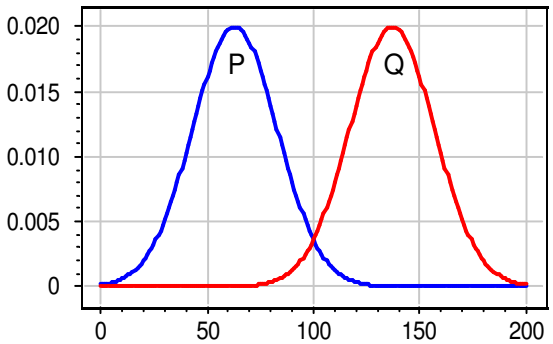
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



$\mathrm{KL}(P\|Q) = 0.442$
$\mathrm{KL}(Q\|P) = 2.966$



$\mathrm{KL}(P\|Q) = 0.444$
$\mathrm{KL}(Q\|P) = 0.444$



$\mathrm{KL}(P\|Q) = 2.969$
$\mathrm{KL}(Q\|P) = 2.969$

Регуляризатор №1: Сглаживание LDA

Гипотеза сглаженности:

распределения ϕ_{wt} близки к заданным распределениям β_w

распределения θ_{td} близки к заданным распределениям α_t

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

Минимизируем взвешенную сумму этих KL-дивергенций:

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

Подставляем, получаем формулы M-шага LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0\beta_w, \qquad \theta_{td} \propto n_{dt} + \alpha_0\alpha_t.$$

Этого вы не найдёте в *D.Blei, A.Ng, M.Jordan.* Latent Dirichlet allocation // Journal of Machine Learning Research, 2003. — Vol. 3. — Pp. 993–1022.

Регуляризатор №1: Сглаживание LDA

Выводы:

- Найдено альтернативное обоснование LDA:
  оказывается, это всего лишь притягивание столбцов Φ, Θ
  к заданным распределениям
- Формулы M-шага LDA получены без байесовского вывода:
  — без предположения об априорном распределении
  — без интегрирования по пространству параметров модели
  — без требования сопряжённости
- Распределение Дирихле утрачивает «особую роль»,
  это один из многих регуляризаторов, и не самый лучший

Регуляризатор №2: Частичное обучение

Пусть известно, что
1) документы $d \in D_0$ относятся к темам $T_d \subset T$,
2) к темам $t \in T_0$ относятся термины $W_t \subset W$.

ϕ^0_{wt} — распределение, равномерное на W_t
θ^0_{td} — распределение, равномерное на T_d

Минимизируем сумму дивергенций $\mathrm{KL}_w(\phi^0_{wt} \| \phi_{wt})$, $\mathrm{KL}_t(\theta^0_{td} \| \theta_{td})$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W_t} \phi^0_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d \in D_0} \sum_{t \in T_d} \theta^0_{td} \ln \theta_{td} \to \max$$

Подставляем, получаем обобщение LDA:

$$\theta_{td} \propto n_{dt} + \beta_0 \theta^0_{td} \qquad \phi_{wt} \propto n_{wt} + \alpha_0 \phi^0_{wt}$$

Nigam K., McCallum A., Thrun S., Mitchell T. Text classification from labeled and unlabeled documents using EM // Machine Learning, 2000, no. 2–3.

Регуляризатор №2: Частичное обучение (новое обобщение)

Идея: вместо логарифма можно взять любую другую монотонно возрастающую функцию μ

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in T_0}\sum_{w\in W_t} \phi^0_{wt}\mu(\phi_{wt}) + \alpha_0 \sum_{d\in D_0}\sum_{t\in T_d} \theta^0_{td}\mu(\theta_{td}) \to \max$$

Подставляем, получаем ещё одно обобщение LDA:

$$\theta_{td} \propto n_{dt} + \beta_0\theta^0_{td}\theta_{td}\mu'(\theta_{td}) \qquad \phi_{wt} \propto n_{wt} + \alpha_0\phi^0_{wt}\phi_{wt}\mu'(\phi_{wt})$$

При $\mu(z) = z$ максимизируется сумма ковариаций $\text{cov}(\theta^0_d, \theta_d)$.

Преимущество ковариационного регуляризатора:
Если θ^0_{td} равномерно на T_d, то ковариация не накладывает ограничений на распределение θ_{td} между темами из T_d.

Регуляризатор №3: Разреживание

Гипотеза разреженности: среди ϕ_{wt}, θ_{td} много нулей.

Чем сильнее разрежено распределение, тем ниже его энтропия.
Максимальной энтропией обладает равномерное распределение.

Поэтому максимизируем дивергенцию между равномерным распределением и искомыми распределениями ϕ_{wt}, θ_{td}:

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

Подставляем, получаем «анти-LDA»:

$$\phi_{wt} \propto \big(n_{wt} - \beta\big)_+, \qquad \theta_{td} \propto \big(n_{dt} - \alpha\big)_+.$$

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions.

Постепенное разреживание распределений ϕ_{wt} и θ_{td}

Эвристика:
постепенно увеличивать коэффициенты регуляризации α, β

Реализация эвристики:
начиная с итерации i_0, через каждые δ итераций
обнуляем долю r наименьших значений
в каждом распределении ϕ_t и θ_d,
так, чтобы сумма обнуляемых значений
не превышала R_θ для распределений θ_d,
не превышала R_ϕ для распределений ϕ_t

Обозначения параметров эвристики:
$i_0{:}\delta{:}r$ (если R_θ и R_ϕ не используются)
$i_0{:}\delta{:}r$, th:R_θ, ph:R_ϕ

Альтернативное обоснование — OBD (Optimal Brain Damage)

Пусть алгоритм сошелся к локальному оптимуму.
Обнуление каких параметров меньше всего повлияет на значение правдоподобия?

Разложим правдоподобие в ряд Тейлора в окрестности точки локального максимума:

$$L(\Phi+\Delta\Phi, \Theta+\Delta\Theta) \approx L(\Phi,\Theta) + \frac{1}{2}\sum_{w,t}\sum_{u,s}\Delta\phi_{wt}\Delta\phi_{us}\frac{\partial^2 L(\Phi,\Theta)}{\partial\phi_{wt}\partial\phi_{us}} +$$

$$+\frac{1}{2}\sum_{t,d}\sum_{s,g}\Delta\theta_{td}\Delta\theta_{sg}\frac{\partial^2 L(\Phi,\Theta)}{\partial\theta_{td}\partial\theta_{sg}} + \sum_{w,t}\sum_{s,d}\Delta\phi_{wt}\Delta\theta_{sd}\frac{\partial^2 L(\Phi,\Theta)}{\partial\phi_{wt}\partial\theta_{sd}}$$

Y. LeCun, J. Denker, S. Solla R. E. Howard, L. D. Jackel.
Optimal Brain Damage // Advances in Neural Information Processing Systems II. Morgan Kauffman. — 1990.

Альтернативное обоснование — OBD (Optimal Brain Damage)

Оценим изменение функционала:

- при обнулении одного параметра ϕ_{wt}:

$$\Delta L = -\frac{1}{2}\sum_{d} n_{dw} p^2(t|d,w) \approx n_{wt}$$

- при обнулении параметров ϕ_{wt} для данного слова w:

$$\Delta L = -\frac{1}{2}\sum_{d,t,s} n_{dw} \frac{\phi_{wt}\theta_{td}}{p(w|d)} \frac{\phi_{ws}\theta_{sd}}{p(w|d)} = -\frac{1}{2}\sum_{t} n_{wt}$$

- при обнулении параметров ϕ_{wt}, θ_{td} для всех слов, тем и документов:

$$\Delta L = -\frac{1}{2}\sum_{wt} n_{wt} - \frac{1}{2}\sum_{td} n_{td}$$

Работает ли разреживание? Эксперимент...

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

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

Число итераций 40; число тем $|T| = 100$.

Воронцов К.В., Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных, 2013. — Т. 1, № 6. — С. 657–686.

Работает ли разреживание? Эксперимент...

Использовались две коллекции:

- NIPS:
  - $|D| = 1566$ статей конференции NIPS на английском языке;
  - суммарной длины $n \approx 2.3 \cdot 10^6$,
  - словарь $|W| \approx 1.3 \cdot 10^4$.
  - Контрольная коллекция: $|D'| = 174$.
- RuDis:
  - $|D| = 2000$ авторефератов диссертаций на русском языке;
  - суммарной длины $n \approx 8.7 \cdot 10^6$,
  - словарь $|W| \approx 3 \cdot 10^4$.
  - Контрольная коллекция: $|D'| = 200$.

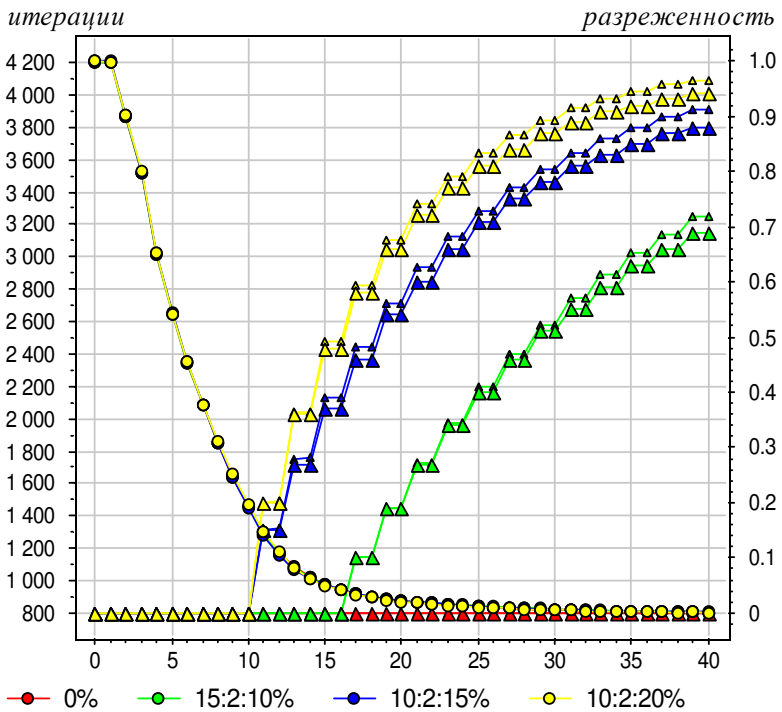
Перплексия тестовой коллекции D' (hold-out perplexity):

$$\mathcal{P}(D') = \exp\left(-\sum_{d\in D'}\sum_{w\in d''} n_{dw} \ln p(w|d) \Big/ \sum_{d\in D'}\sum_{w\in d''} n_{dw}\right)$$

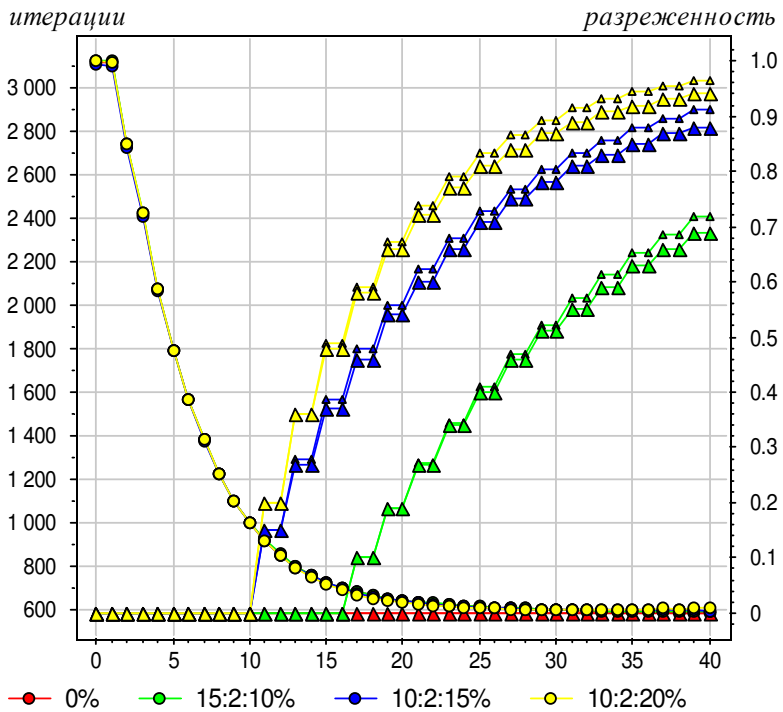
число итераций 40; число тем $|T| = 100$.

Разреживание распределений ϕ_{wt} и θ_{td}

робастная модель с фоном и шумом,
разреживание через 2 итерации



RuDis

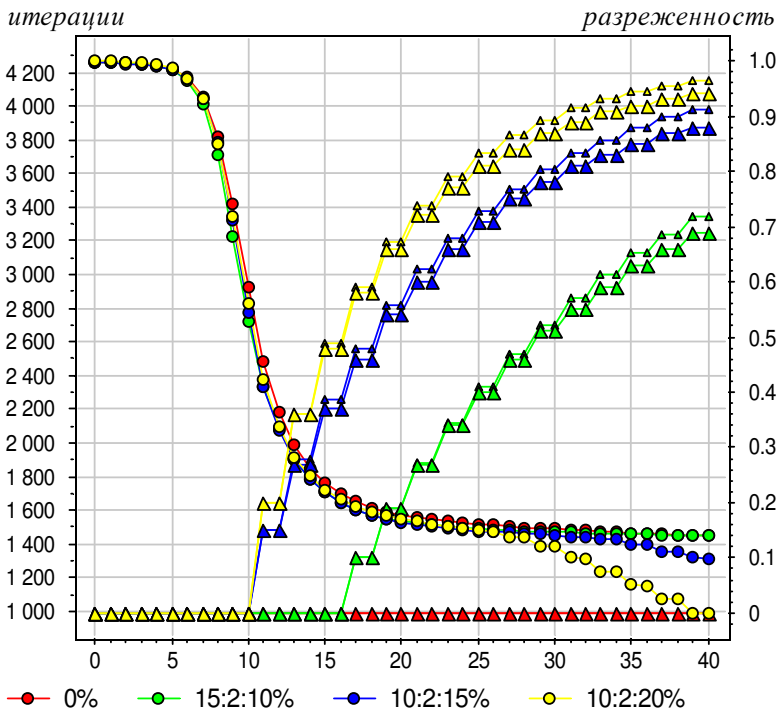


NIPS

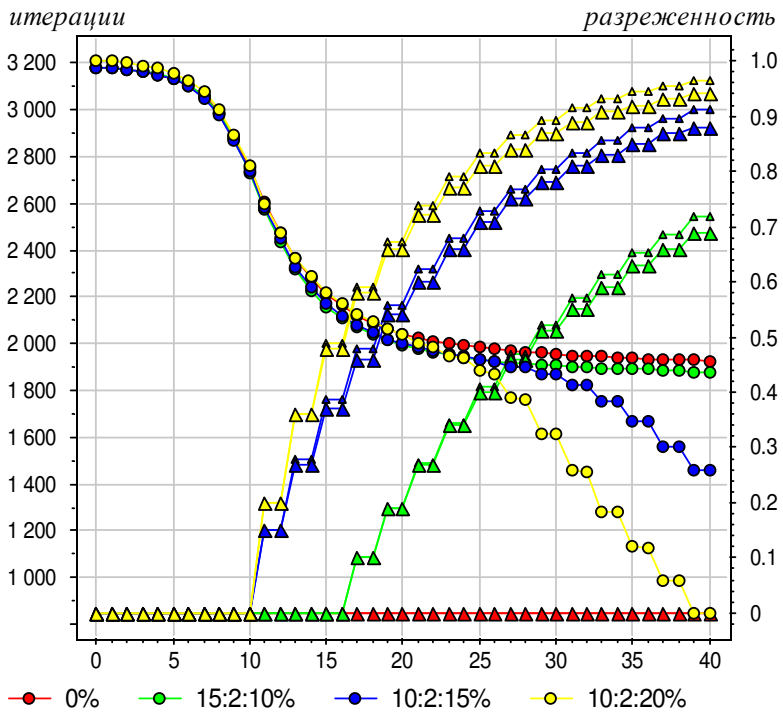
- Разреживание не ухудшает перплексию робастной модели

Разреживание распределений ϕ_{wt} и θ_{td}

упрощённая робастная модель с шумом без фона,
разреживание через 2 итерации



RuDis

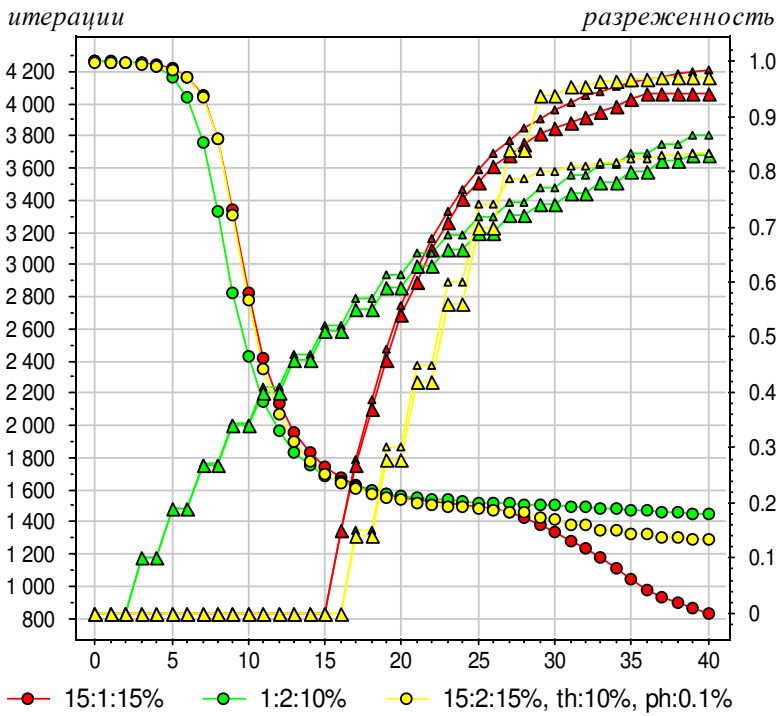


NIPS

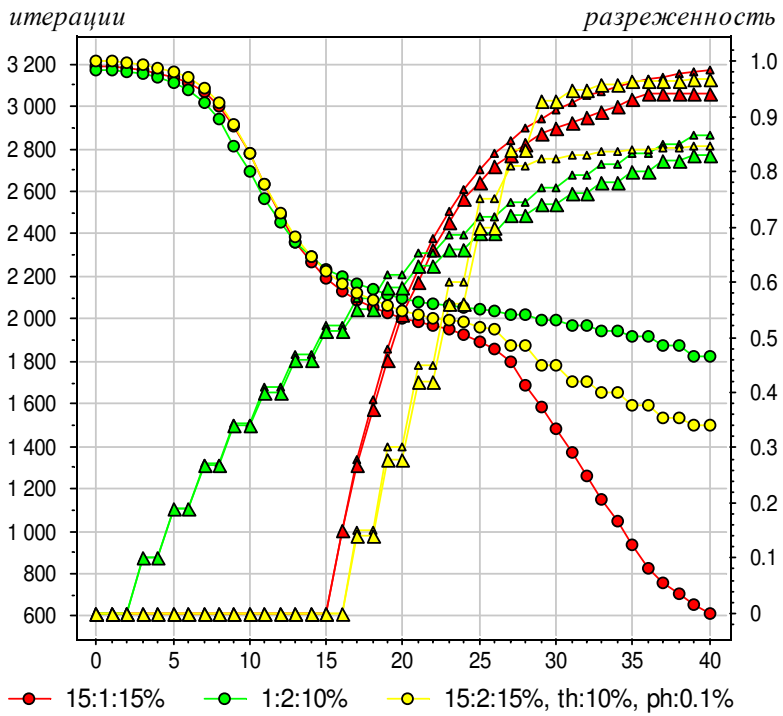
- Вполне достаточно упрощённой робастной модели

Разреживание распределений ϕ_{wt} и θ_{td}

упрощённая робастная модель с шумом без фона,
агрессивные стратегии разреживания $i_0{:}\delta{:}r$, th:R_θ, ph:R_ϕ



RuDis

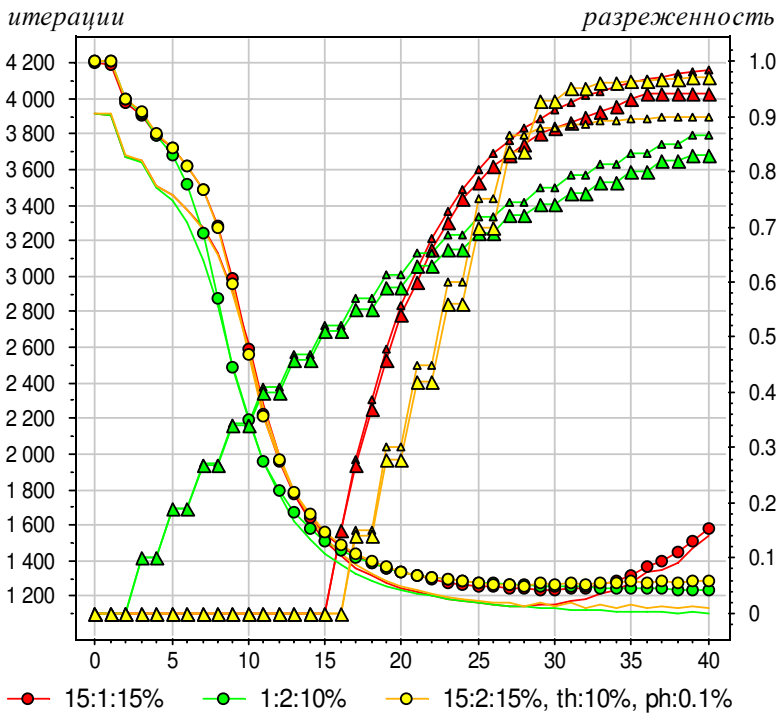


NIPS

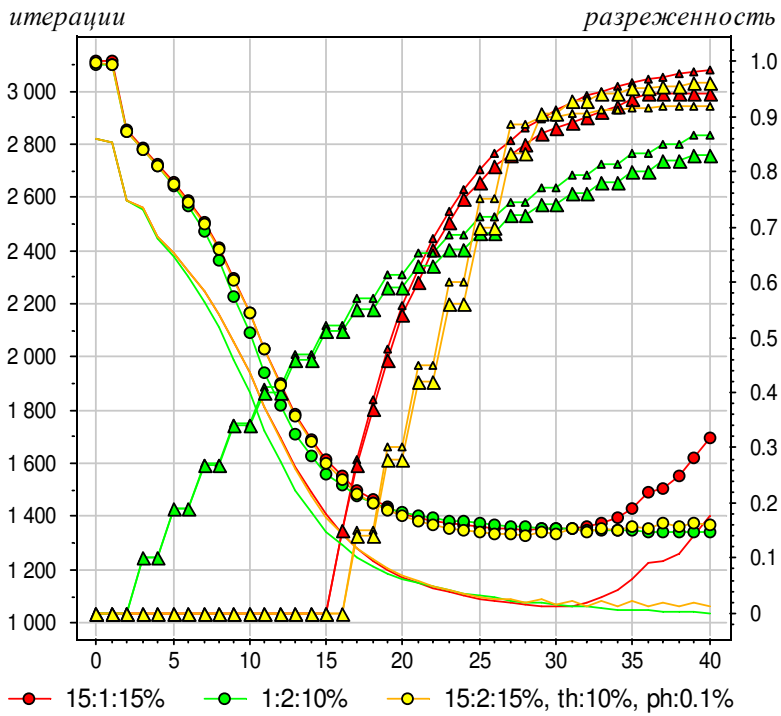
- Можно разреживать ещё агрессивнее

Разреживание распределений ϕ_{wt} и θ_{td}

робастная модель с шумом ($\gamma = 0.01$) и фоном ($\varepsilon = 0.01$),
агрессивные стратегии разреживания $i_0{:}\delta{:}r$, th:R_θ, ph:R_ϕ



RuDis



NIPS

- Разреживание может портить модель с низким шумом

Выводы

1. Возможно достигать разреженности 95–99% без ухудшения перплексии
2. При числе тем $|T| = 100$ это означает, что в среднем каждое слово относится к 1–5 темам
3. При этом многие строки матрицы Φ обнуляются, т.е. слово оказывается нетематическим
4. Можно использовать упрощённые робастные модели,
   — не требующие хранения матрицы $(\pi_{dw})_{D\times W}$,
   — не требующие задания параметров ε, γ

Воронцов К.В., Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных, 2013. — Т. 1, № 6. — С. 657–686.

Регуляризатор №4: Антикорреляция

Гипотеза некоррелированности тем:
чем различнее темы, тем лучше они интерпретируются.

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, получаем ещё один вариант разреживания — постепенное контрастирование строк матрицы Φ:

$$\phi_{wt} \propto \left(n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \right)_+ .$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp. 224–228.

Регуляризатор №5: Максимизация когерентности тем

Гипотеза: тема лучше интерпретируется, если она содержит *когерентные* (часто встречающиеся рядом) слова $u, v \in W$.

Пусть C_{uv} — оценка когерентности, например $\hat{p}(v|u) = N_{uv}/N_u$.

$$R(\Phi, \Theta) = \tau \sum_{t \in T} \sum_{(u,v)} C_{uv} n_{ut} \ln \phi_{vt} \to \max,$$

Подставляем, получаем ещё один вариант сглаживания: векторы ϕ_{wt} притягиваются к эмпирическим оценкам распределений $p(w|t)$, вычисляемым по когерентным словам:

$$\phi_{wt} \propto n_{wt} + \tau \sum_{u \in W \setminus w} C_{uw} n_{ut}.$$

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Empirical Methods in Natural Language Processing, EMNLP-2011. — Pp. 262–272.

Регуляризатор №6: Выделение стоп-слов

Гипотеза: нетематические слова имеют во всех документах одинаковое распределение, близкое к $\hat{p}(w) = n_w/n$

Пусть $S \subset T$ — подмножество тем для стоп-слов

Минимизируем сумму дивергенций $\mathrm{KL}_w(\hat{p}(w)\|\phi_{wt})$:

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

Подставляем, получаем «LDA только для стоп-слов»:

$$\phi_{wt} \propto \hat{n}_{wt} + \beta_0 \hat{p}(w), \quad t \in S.$$

Вместе с разреживанием, антикорреляцией, когерентностью сглаживание переводит стоп-слова из других тем в темы из S

Регуляризатор №7: Связи между документами

Гипотеза: чем больше n_{dc} — число ссылок из d на c, тем более близки тематики документов d и c.

Минимизируем ковариации между вектор-столбцами связанных документов θ_d, θ_c:

$$R(\Phi, \Theta) = \tau \sum_{d,c \in D} n_{dc} \operatorname{cov}(\theta_d, \theta_c) \to \max,$$

Подставляем, получаем ещё один вариант сглаживания:

$$\theta_{td} \propto \hat{n}_{dt} + \tau \theta_{td} \sum_{c \in D} n_{dc} \theta_{tc}.$$

Dietz L., Bickel S., Scheffer T. Unsupervised prediction of citation influences // ICML 2007. — Pp. 233–240.

Регуляризатор №8: Классификация

Пусть C — множество классов документов (категории, пользователи, авторы, ссылки, годы, конференции,...)

Гипотеза:
классификация документа d объясняется его темами:

$$p(c|d) = \sum_{t \in T} p(c|t)p(t|d) = \sum_{t \in T} \psi_{ct}\theta_{td}$$

Минимизируем дивергенцию между моделью $p(c|d)$
и «эмпирической частотой» классов в документах m_{dc}:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct}\theta_{td} \to \max$$

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 1–2.

Регуляризатор №8: Классификация (EM-алгоритм)

E-шаг. По формуле Байеса:

$$p(t|d,w) = \frac{\phi_{wt}\theta_{td}}{\sum_{s\in T}\phi_{ws}\theta_{sd}} \qquad p(t|d,c) = \frac{\psi_{ct}\theta_{td}}{\sum_{s\in T}\psi_{cs}\theta_{sd}}$$

M-шаг. Максимизация регуляризованного правдоподобия:

$$\phi_{wt} \propto n_{wt} \qquad n_{wt} = \sum_{d\in D} n_{dw} p(t|d,w)$$

$$\theta_{td} \propto n_{dt} + \tau m_{dt} \qquad n_{dt} = \sum_{w\in W} n_{dw} p(t|d,w) \qquad m_{dt} = \sum_{c\in C} m_{dc} p(t|d,c)$$

$$\psi_{ct} \propto m_{ct} \qquad m_{ct} = \sum_{d\in D} m_{dc} p(t|d,c)$$

Регуляризатор №9: Динамическая тематическая модель

Пусть классы C — это время публикации (год/месяц/день)

Гипотеза:
тематика меняется медленно, поэтому вероятности ψ_{ct} в последовательные годы $(c-1, c)$ должны быть близки:

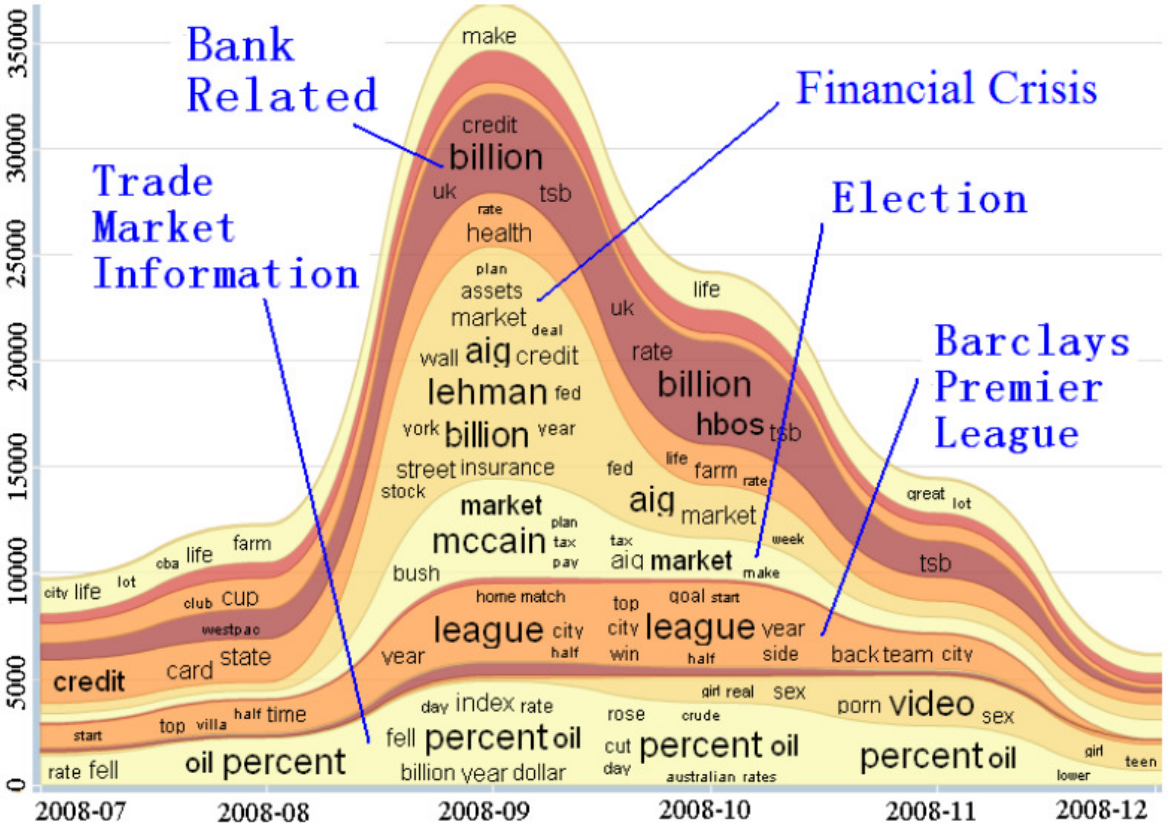
$$R_2(\Psi) = -\frac{\tau_2}{2} \sum_{c \in C} \sum_{t \in T} \left(\psi_{ct} - \psi_{c-1,t}\right)^2 \to \max.$$

Сглаживание распределений $\psi_{ct} = p(c|t)$:

$$\psi_{ct} \propto \tau_1 m_{ct} + \tau_2 \psi_{ct} \left(\psi_{c-1,t} + \psi_{c+1,t} - 2\psi_{ct}\right).$$

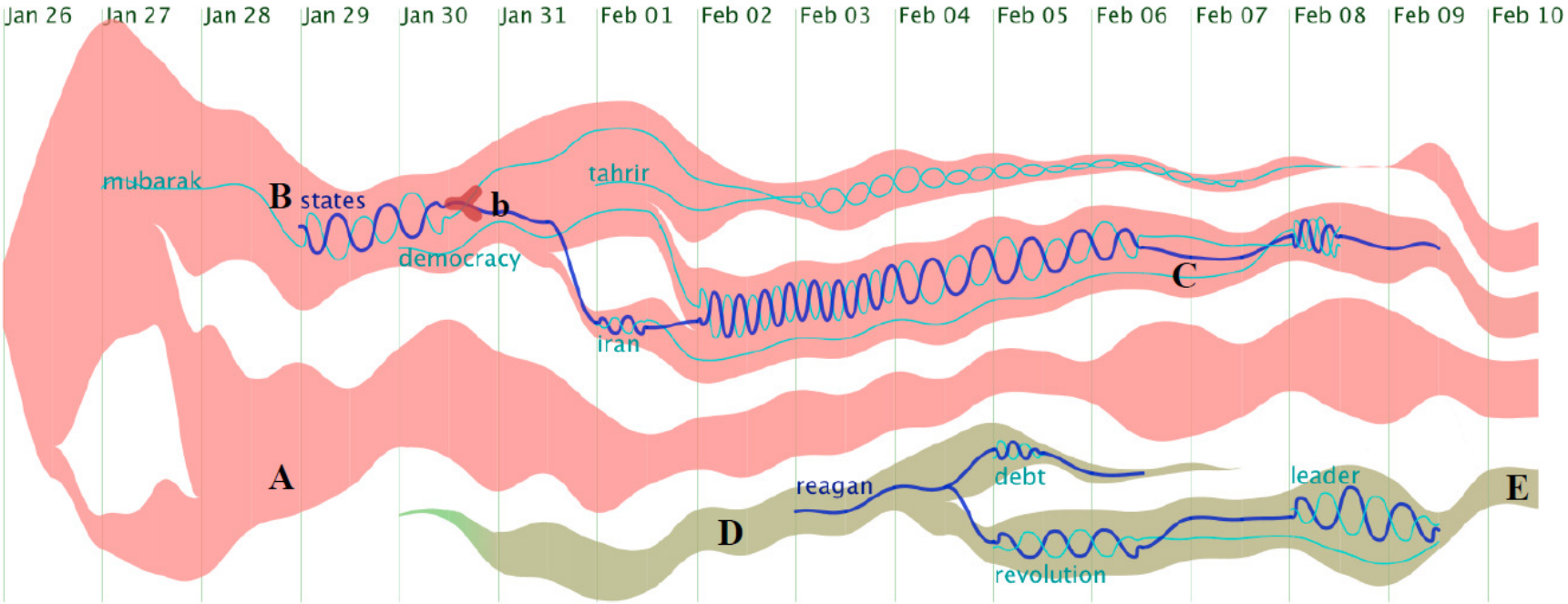
Если значение ψ_{ct} меньше полусуммы соседних вероятностей $\psi_{c-1,t}$, $\psi_{c+1,t}$, то оно увеличивается, иначе — уменьшается:

Пример динамической 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.

Ещё пример динамической 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.

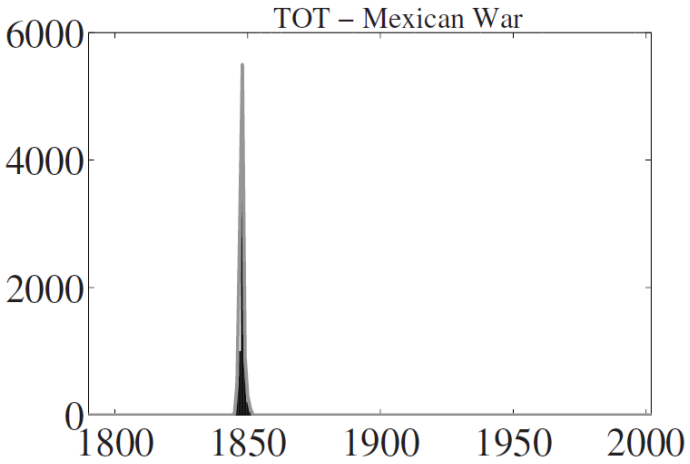
Обобщения и модификации тематических моделей

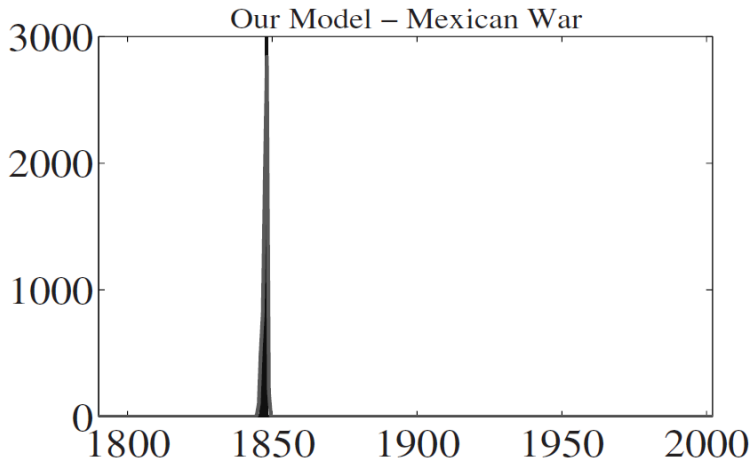
- Робастные и разреженные модели
- Онлайновые (динамические) модели
- Модели связей между документами (ссылки, цитирование)
- Модели классификации и категоризации документов
- Author-topic models, учитывающие авторство
- Entity-topic models, учитывающие именованные сущности
- Темпоральные модели, учитывающие годы публикаций
- Модели, учитывающие связи слов внутри документа
- Иерархические модели
- Многоязыковые модели

Daud A., Li J., Zhou L., Muhammad F. 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

Совмещение динамической и n-граммной модели



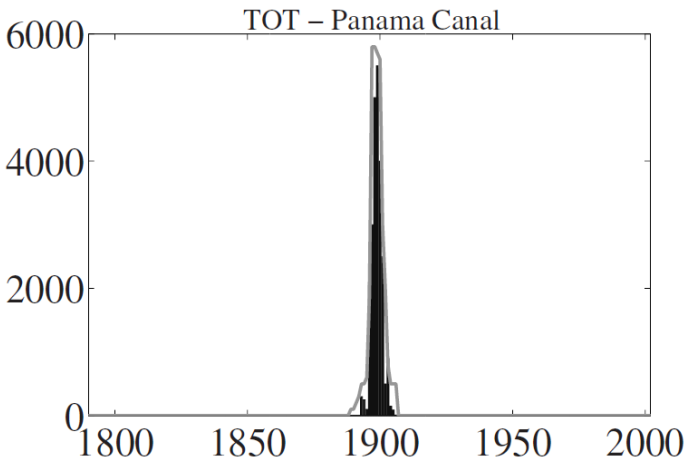


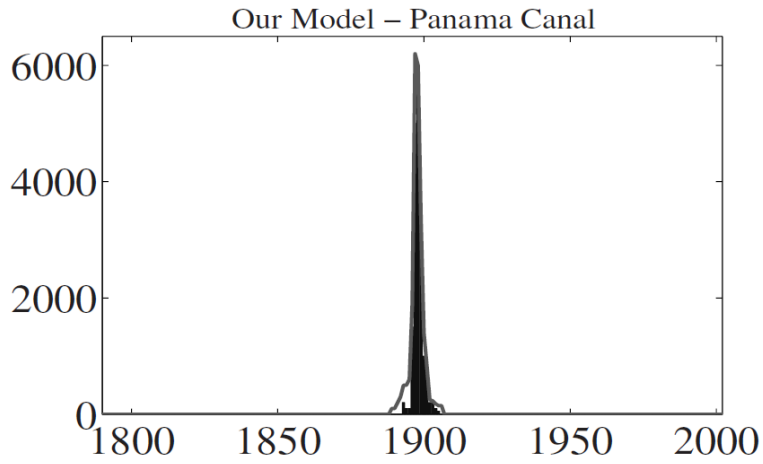
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents // 35'th ECIR 2013, Moscow, March 24–27. — pp. 292–304.

Совмещение динамической и n-граммной модели





1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

1. panama canal	8. united states senate
2. isthmian canal	9. french canal company
3. isthmus panama	10. caribbean sea
4. republic panama	11. panama canal bonds
5. united states government	12. panama
6. united states	13. american control
7. state panama	14. canal

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents // 35'th ECIR 2013, Moscow, March 24–27. — pp. 292–304.

Иерархические тематические модели

Для выявления иерархии тем используется модель HDP — иерархический процесс Дирихле, обобщение модели LDA.

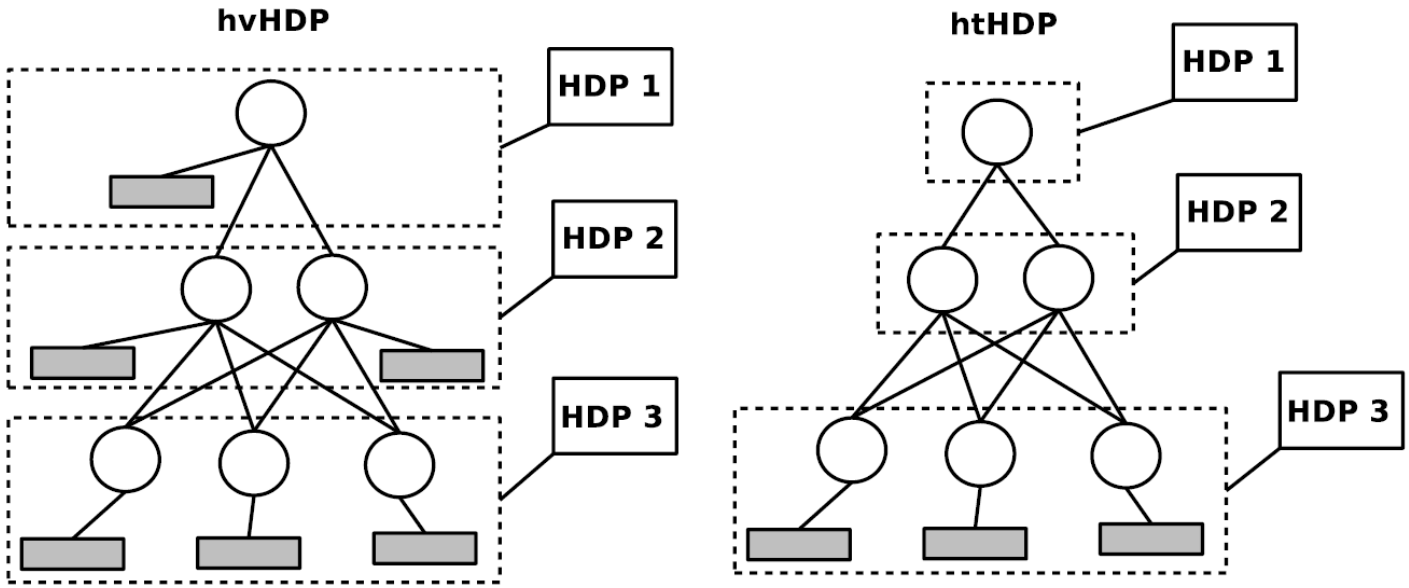
Задача построения иерархии и задача оценивания её качества признаются открытыми научными проблемами.

- "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."

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

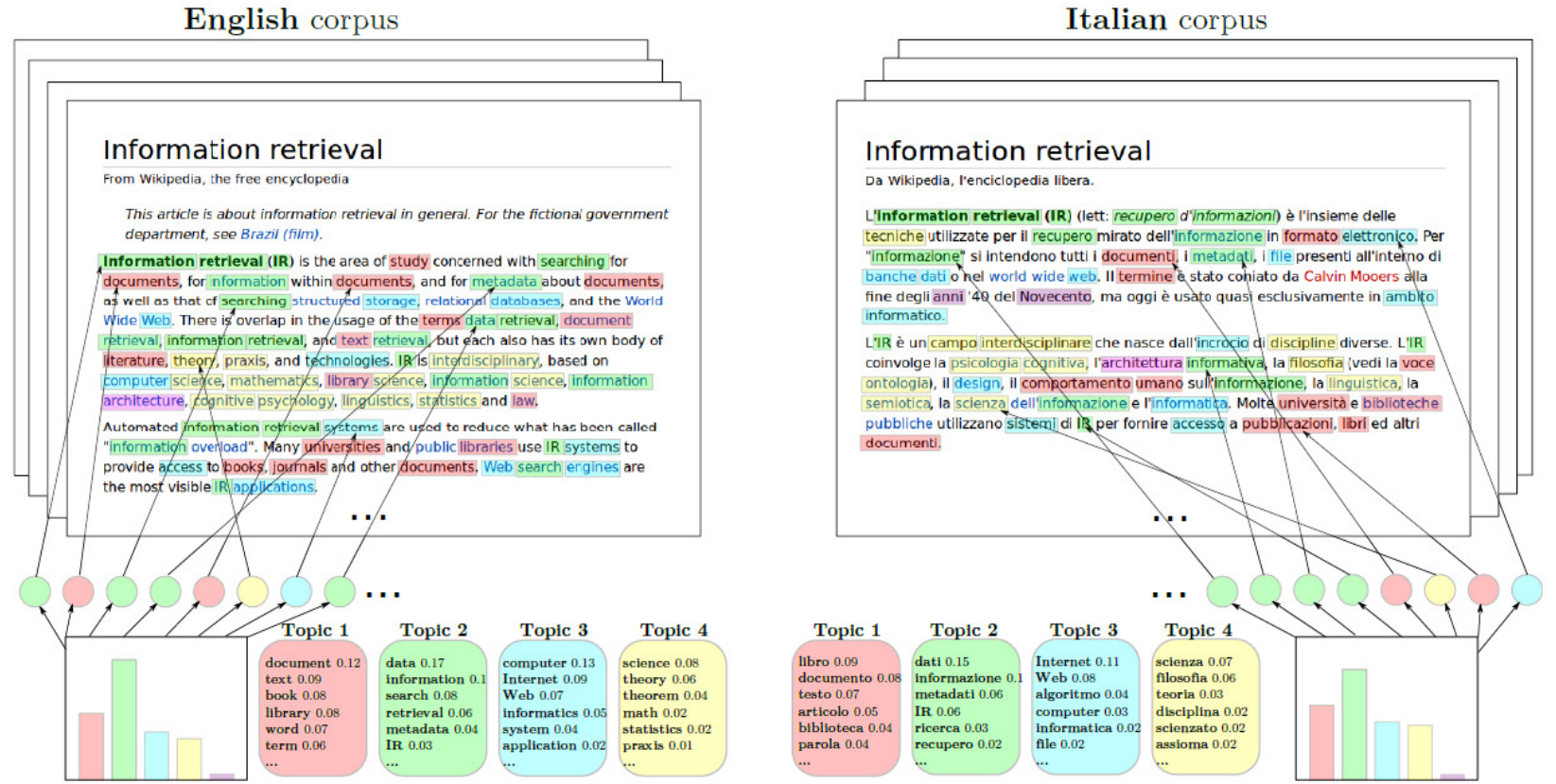
Две восходящие стратегии построения иерархии

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



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

Многоязычные модели



Vulić I., De Smet W., Tang J., Moens M.-F.. Probabilistic topic modeling in multilingual settings: a short overview of its methodology with applications // NIPS, 7–8 December 2012. — Pp. 1–11.

Многоязычные модели

Основные выводы:

- достаточно приравнять $p(t|d)$ параллельных текстов
- достаточно иметь относительно небольшой параллельный корпус для каждой пары языков
- достаточно выравнивания на уровне документов
- выравнивание на уровне предложений или абзацев не обязательно, но улучшает качество многоязычного поиска
- наличие словаря «много-ко-многим» не обязательно, но улучшает качество многоязычного поиска
- нет необходимости применять «тяжёлые» методы машинного перевода