# Матричные разложения и вероятностное тематическое моделирование текстовых коллекций

Воронцов Константин Вячеславович
ВЦ РАН • МФТИ • МГУ • ВШЭ • Яндекс • FORECSYS


• Традиционная молодёжная летняя школа •
27 июня 2014

# Содержание

# Задачи матричного разложения

**Дано:** матрица $Z = \|z_{ij}\|_{n\times m}$, $(i,j) \in \Omega \subseteq \{1..n\} \times \{1..m\}$

**Найти:** матрицы $X = \|x_{it}\|_{n\times k}$ и $Y = \|y_{tj}\|_{k\times m}$ такие, что

$$\|Z - XY\|_{\Omega,d} = \sum_{(i,j)\in\Omega} d\Big(z_{ij}, \sum_t x_{it}y_{tj}\Big) \to \min_{X,Y}$$

**Типы задач:**

- различные функции потерь:
  квадратичная $d(z,\hat{z}) = (z-\hat{z})^2$,
  Кульбака–Лейблера: $d(z,\hat{z}) = z\ln(z/\hat{z}) - z + \hat{z}$, и др.
- неотрицательные матричные разложения: $x_{it} \geqslant 0$, $y_{tj} \geqslant 0$
- стохастические матричные разложения: $\sum_i x_{it} = 1$, $\sum_t y_{tj} = 1$
- сильно разреженные данные: $|\Omega| \ll nm$

# Примеры прикладных задач матричного разложения

1. **Разделение смеси химических веществ по данным жидкостной хроматографии**

$$z_{t\lambda} = \sum_i x_{ti} y_{i\lambda}$$

**дано:** $z_{t\lambda}$ — выход сканирующего УФ-детектора;
**найти:** $x_{ti}$ — хроматограмма $i$-го вещества, $t$ — время;
$y_{i\lambda}$ — спектр $i$-го вещества, $\lambda$ — длина волны.

2. **Оценивание экспрессии генов по данным ДНК-микрочипов с учётом кросс-гибридизации**

$$z_{pk} = \sum_g a_{pg} c_{gk}$$

**дано:** $z_{pk}$ — интенсивность свечения $p$-й пробы на $k$-м чипе;
**найти:** $a_{pg}$ — коэффициент сродства $p$-й пробы $g$-му гену,
$c_{gk}$ — концентрация $g$-го гена на $k$-м чипе.

# Примеры прикладных задач матричного разложения

3. **Выявление интересов в рекомендующих системах (коллаборативная фильтрация)**

$$z_{iu} = \sum_t p_{it} q_{tu}$$

**дано:** $z_{iu}$ — рейтинги товаров $i$, поставленные пользователем $u$;
**найти:** $p_{it}$ — профиль интересов товара $i$;
$q_{tu}$ — профиль интересов пользователя $u$.

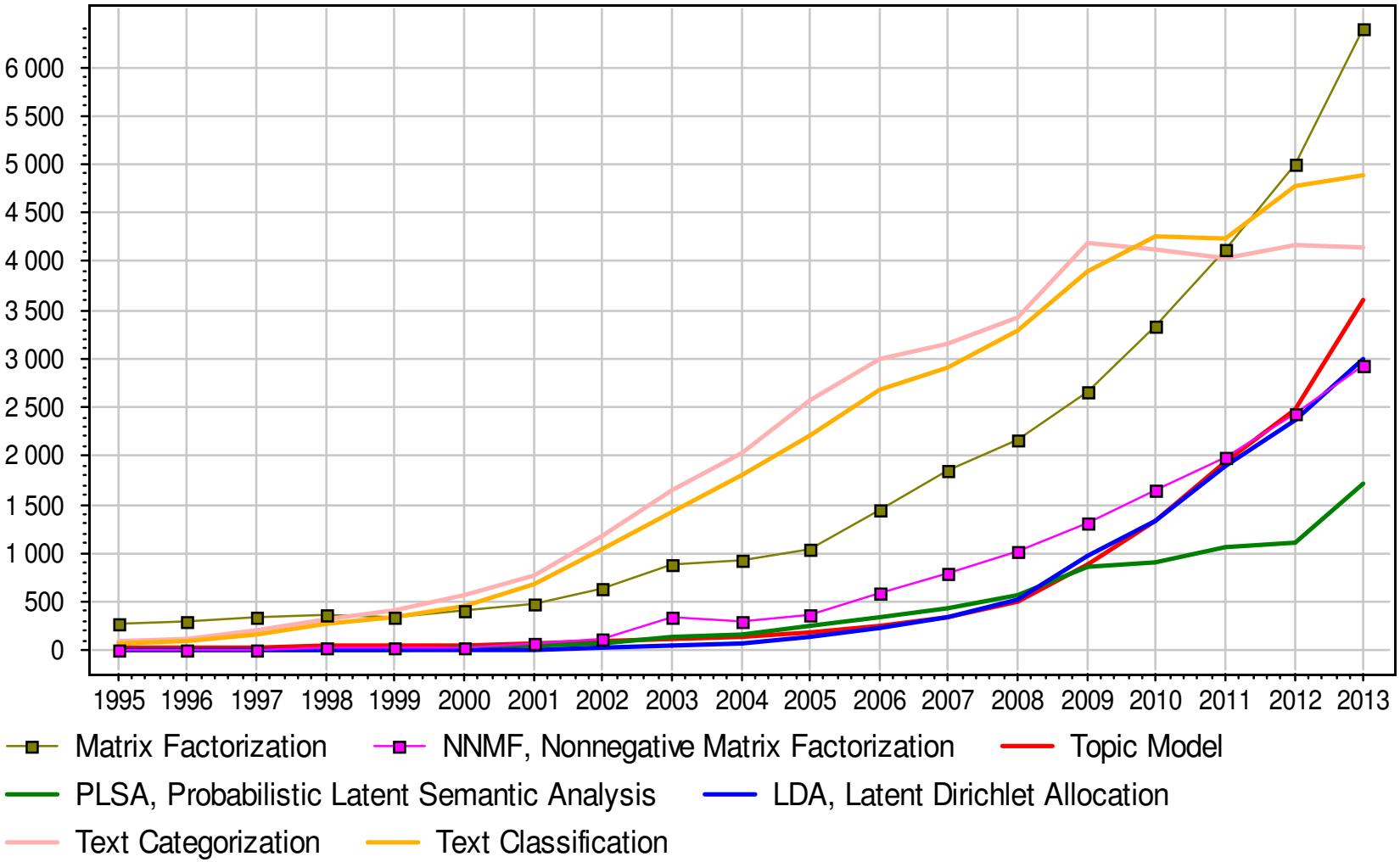
4. **Выявление латентных тем в коллекциях текстов (тематическое моделирование)**

$$z_{wd} = \sum_t \phi_{wt} \theta_{td}$$

**дано:** $z_{wd} = p(w|d)$ — частоты слов $w$ в документах $d$;
**найти:** $\phi_{wt} = p(w|t)$ — распределения слов $w$ в темах $t$,
$\theta_{td} = p(t|d)$ — распределения тем $t$ в документах $d$.

# Тематическое моделирование и близкие области исследований

Динамика цитирования, по данным Google Scholar:



# Понятие «латентной темы»

- *Тема* — специальная терминология предметной области.
- *Тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах.
- *Тема* — вероятностное распределение на терминах: $p(w|t)$ — вероятность встретить термин $w$ в теме $t$.

Документ $d$ состоит из наблюдаемых терминов $w_1, \ldots, w_{n_d}$, $p(w|d)$ — известная частота термина $w$ в документе $d$.

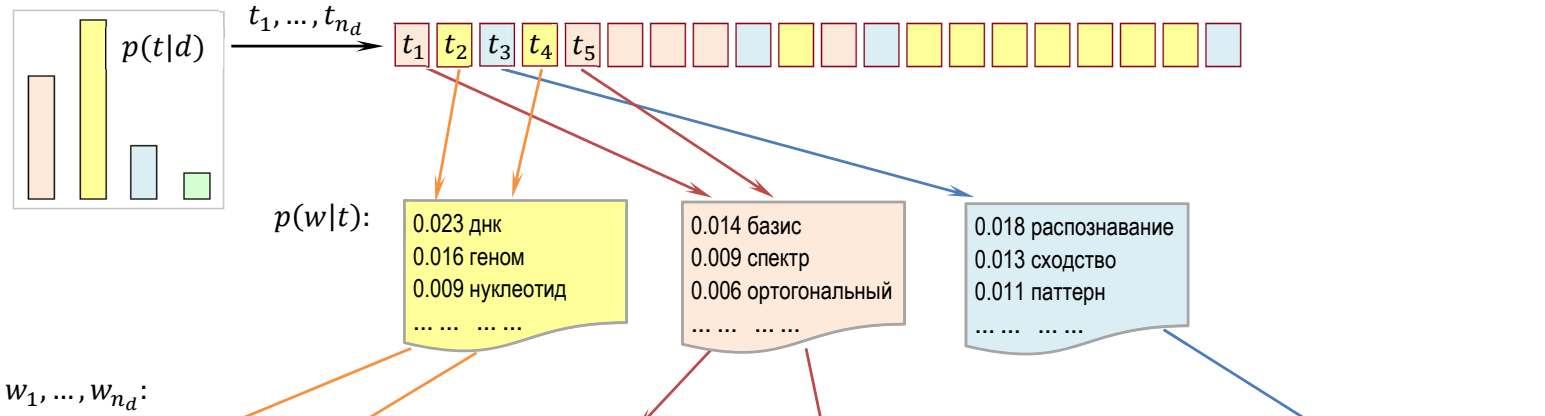
Документ имеет ненаблюдаемый *тематический профиль*: $p(t|d)$ — неизвестная частота темы $t$ в документе $d$.

Когда автор писал термин $w$ в документ $d$, он думал о теме $t$. *Тематическая модель* пытается выявить латентные темы.

# Вероятностная тематическая модель

описывает появление терминов $w$ в документах $d$ темами $t$:

$$p(w|d) = \sum_t p(w|t)p(t|d)$$




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).

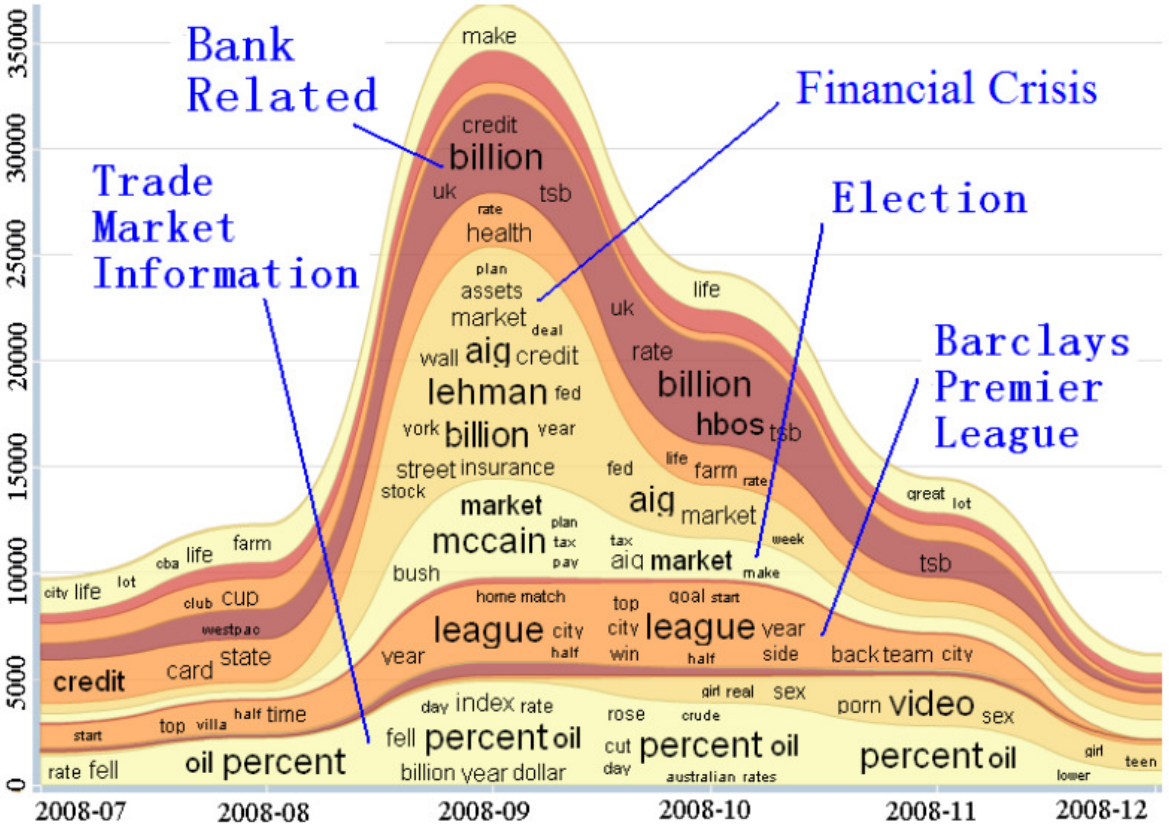
# Цели и приложения тематического моделирования

- Выявить скрытую тематическую структуру коллекции текстов
- Выявить семантический профиль каждого документа

**Приложения:**

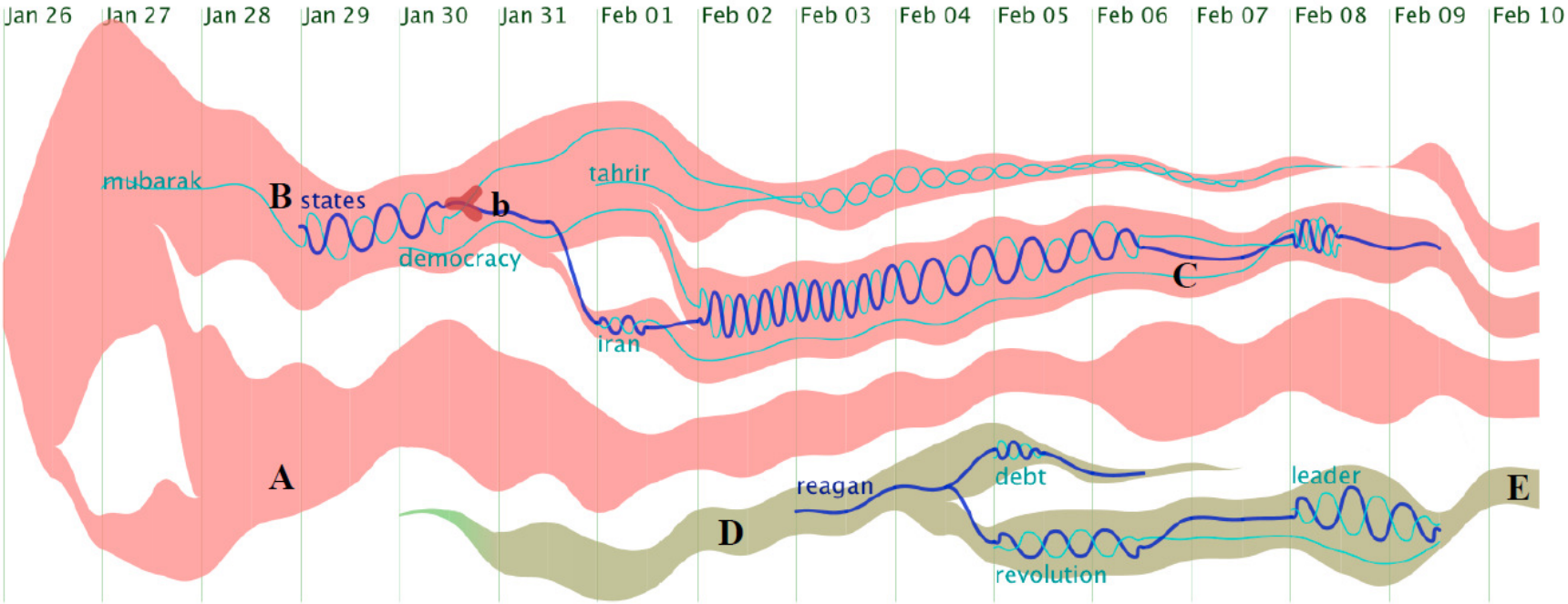
- Семантический поиск по текстовому запросу любой длины
- Категоризация, классификация, аннотирование, суммаризация, сегментация текстовых документов
- Поиск научной информации, трендов, фронта исследований
- Поиск специалистов (expert search), рецензентов, проектов
- Анализ и агрегирование новостных потоков
- Рубрикация документов, изображений, видео, музыки
- Рекомендующие системы, коллаборативная фильтрация
- Аннотация генома и другие задачи биоинформатики
- Анализ дискретизированных биомедицинских сигналов

# Динамические модели, учитывающие время



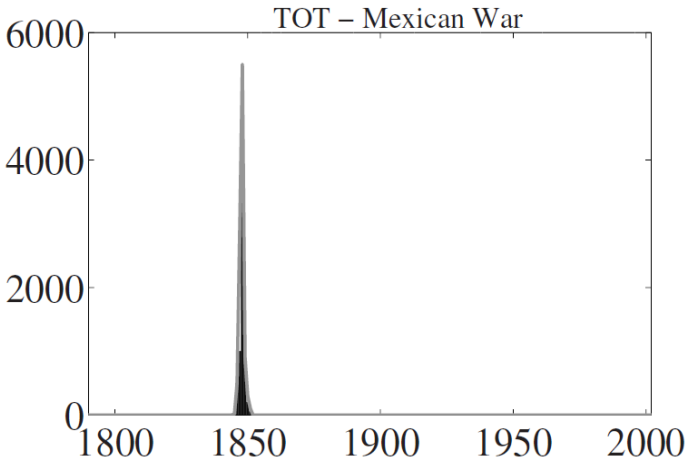

*Jianwen Zhang, Yangqiu Song, Changshui Zhang, Shixia Liu* Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora // KDD'10, July 25–28, 2010.

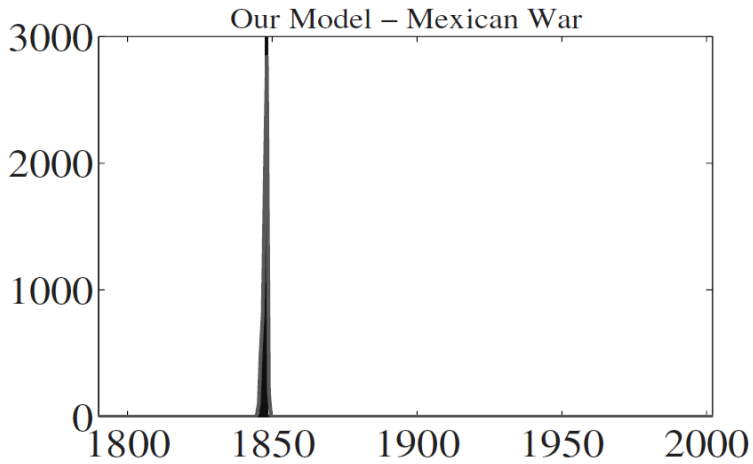
# Динамические модели эволюции тем




*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.

# Совмещение динамической и $n$-граммной модели



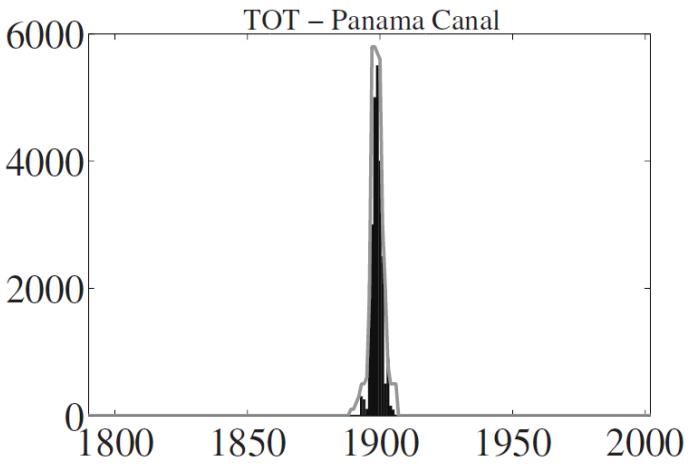



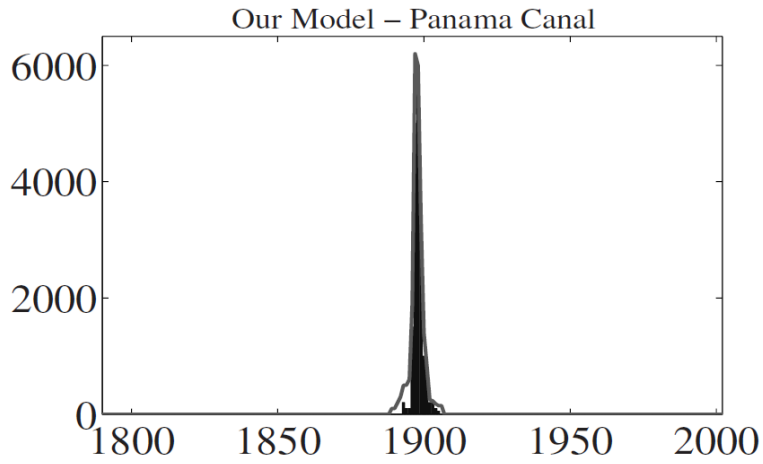

| | |
|---|---|
| 1. mexico | 8. territory |
| 2. texas | 9. army |
| 3. war | 10. peace |
| 4. mexican | 11. act |
| 5. united | 12. policy |
| 6. country | 13. foreign |
| 7. government | 14. citizens |

| | |
|---|---|
| 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 |

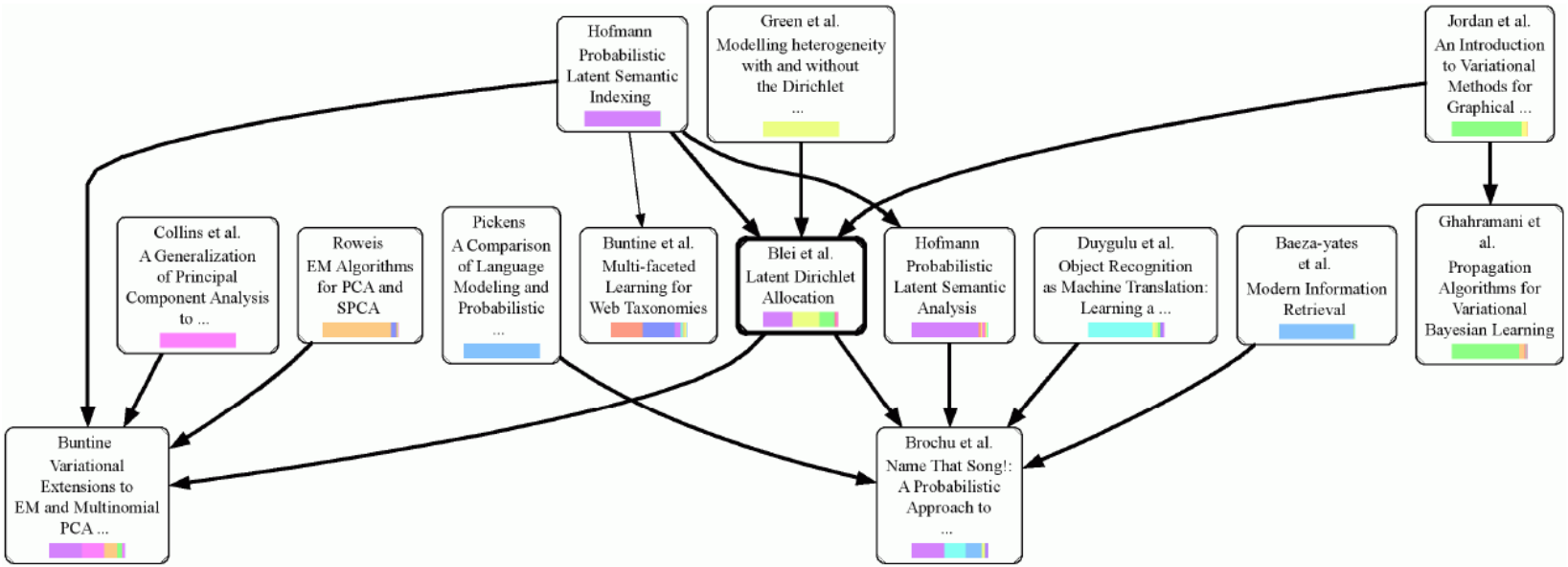
| | |
|---|---|
| 1. panama canal | 8. united states senate |
| 2. isthmian canal | 9. french canal company |
| 3. isthmus panama | 10. caribbean sea |
| 4. republic panama | 11. panama canal bonds |
| 5. united states government | 12. panama |
| 6. united states | 13. american control |
| 7. state panama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

# Модели, учитывающие цитирования или гиперссылки

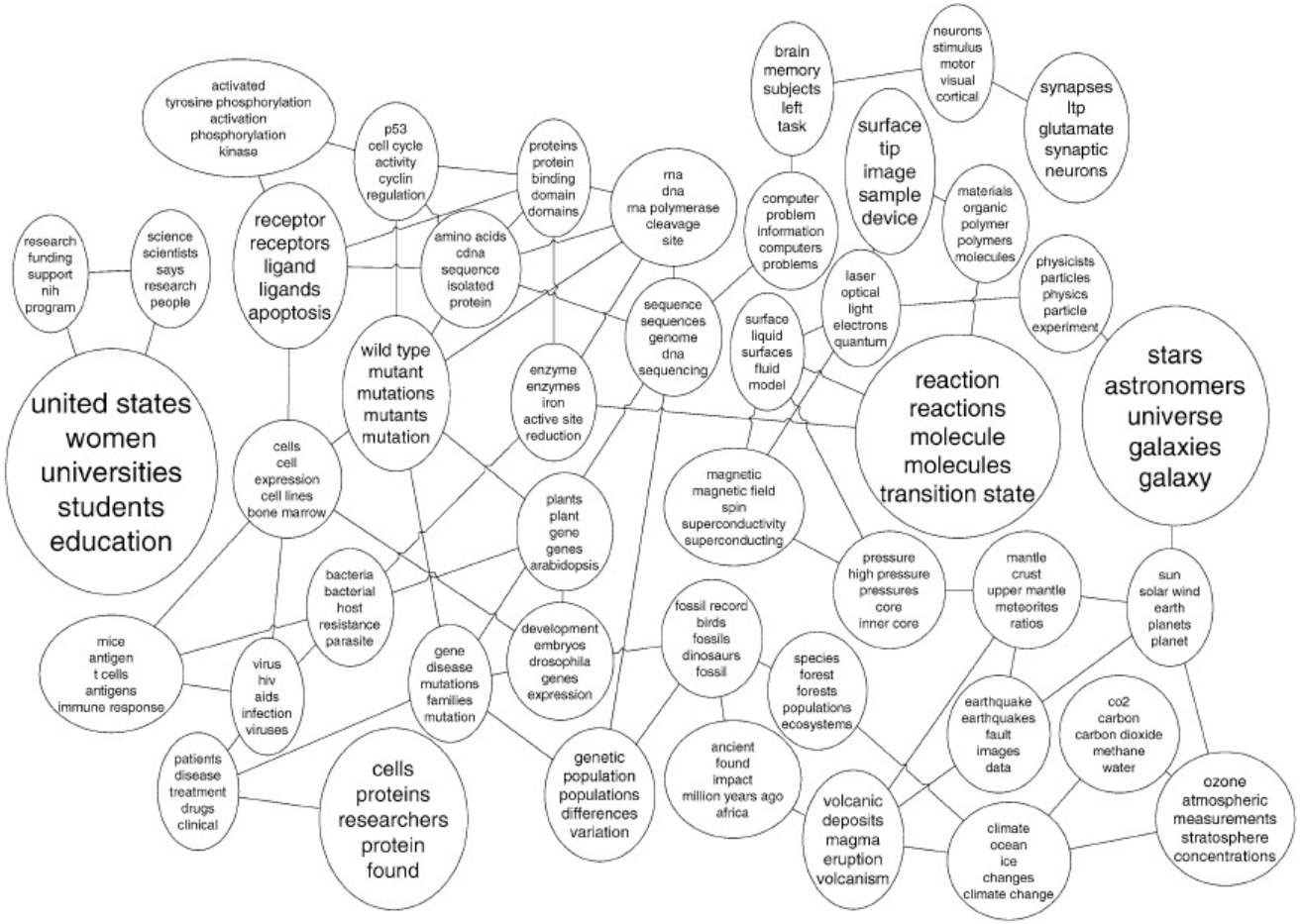
Учёт ссылок уточняет тематическую модель

Тематическая модель выявляет самые влиятельные ссылки



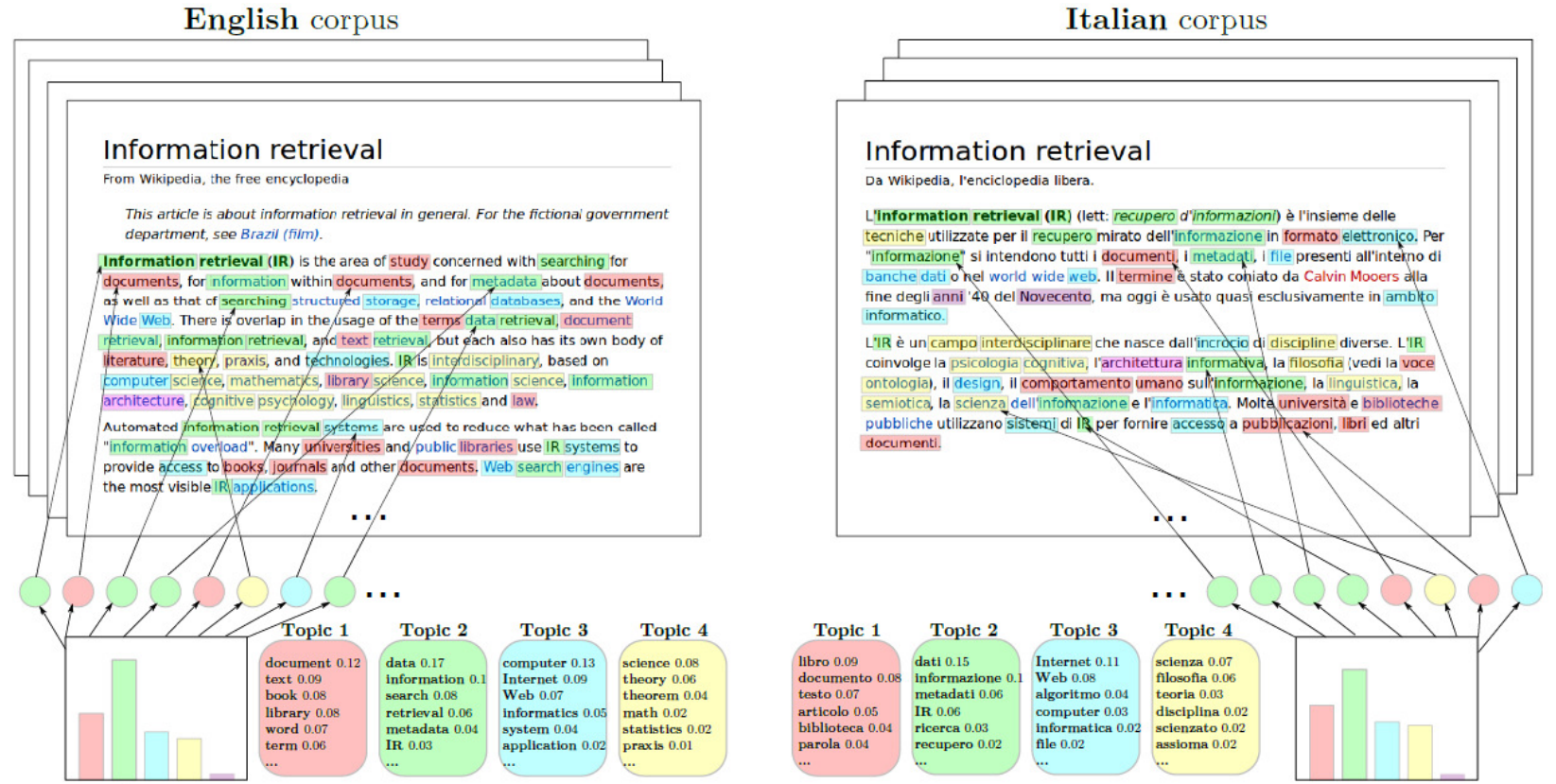

*Laura Dietz, Steffen Bickel, Tobias Scheffer*. Unsupervised prediction of citation influences // ICML-2007, Pp. 233–240.

# Выявление взаимосвязей между темами




*D. Blei, J. Lafferty*. A correlated topic model of Science // Annals of Applied Statistics, 2007. Vol. 1, Pp. 17-35.

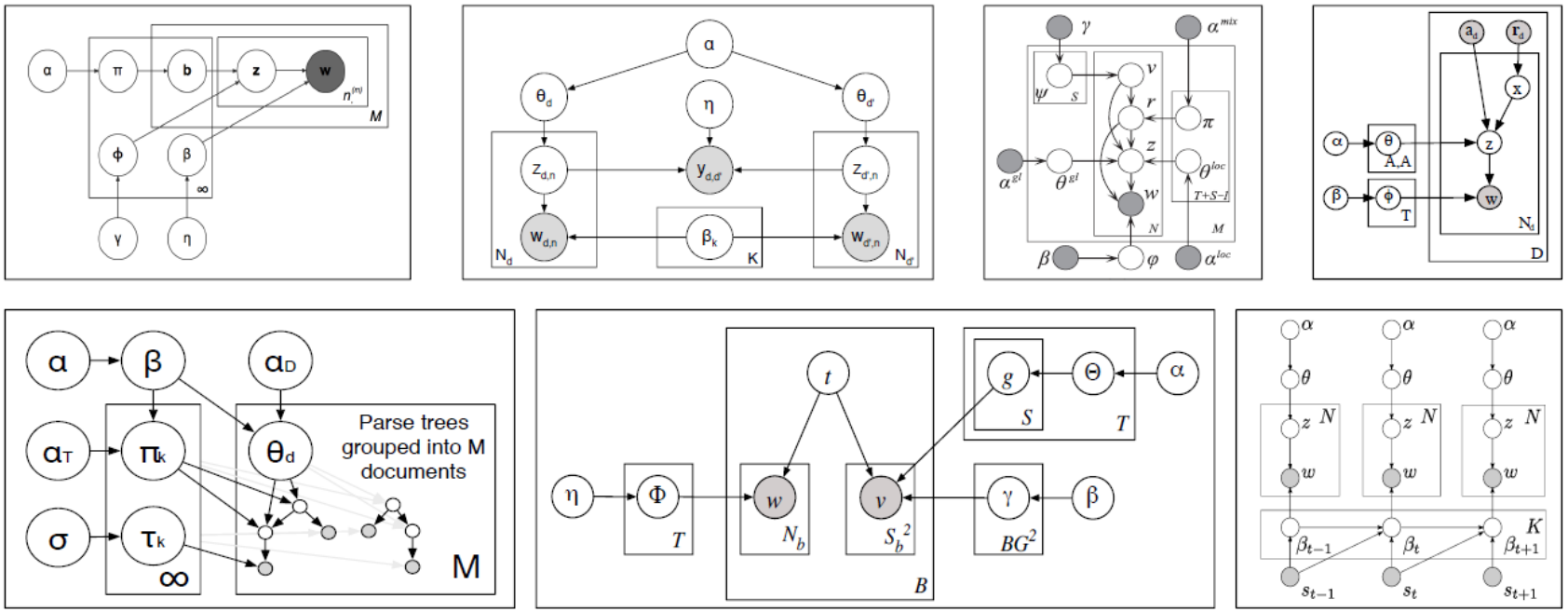
# Многоязычные модели




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

# Резюме по краткому обзору тематических моделей

- Их много и они разные :) Но их трудно комбинировать :(
- Математический аппарат местами выносит мозг :(




*David Blei.* Probabilistic topic models // Communications of the ACM. 2012. Vol. 55. No. 4. Pp. 77–84.

**Topic Modeling Bibliography:** http://mimno.infosci.cornell.edu/topics.html

# Вероятностная тематическая модель (BTM)

$W$ — словарь, множество терминов (слов, словосочетаний)
$D$ — множество (коллекция, корпус) текстовых документов

**Базовые предположения:**

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

**Вероятностная модель порождения документов:**

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

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

# Обратная задача BTM: стохастическое матричное разложение

**Прямая задача:** по распределениям $\phi_{wt} \equiv p(w|t)$, $\theta_{td} \equiv p(t|d)$ сгенерировать коллекцию документов $d = \{w_1, \ldots, w_{n_d}\}$, $d \in D$

**Обратная задача:** по коллекции $D$, заданной частотами $n_{dw}$ — сколько раз термин $w$ встречается в документе $d$, найти параметры модели $\phi_{wt}$, $\theta_{td}$.

Это задача стохастического матричного разложения:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) \quad \Longleftrightarrow \quad \underset{W \times D}{Z} = \underset{W \times T}{\Phi} \cdot \underset{T \times D}{\Theta}$$

$Z = \|p(w|d)\|_{W \times D}$ — известная матрица частот, $p(w|d) = \frac{n_{dw}}{n_d}$
$\Phi = \|\phi_{wt}\|_{W \times T}$ — искомая матрица терминов тем $\phi_{wt} = p(w|t)$,
$\Theta = \|\theta_{td}\|_{T \times D}$ — искомая матрица тем документов $\theta_{td} = p(t|d)$.

# PLSA — Probabilistic Latent Semantic Analysis [Hofmann 1999]

Принцип максимума правдоподобия: $\ln \prod_{d,w} p(d,w)^{n_{dw}} \to \max_{\Phi,\Theta}$

**Задача:** максимизировать логарифм правдоподобия

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

при ограничениях неотрицательности и нормировки

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

$\Longleftrightarrow$ минимизировать взвешенную сумму KL-дивергенций:

$$\sum_{d\in D} n_d \underbrace{\sum_{w\in d} \hat{p}(w|d) \ln \frac{\hat{p}(w|d)}{p(w|d)}}_{\mathrm{KL}(\hat{p}\|p)} \to \min_{\Phi,\Theta},$$

# EM-алгоритм для максимизации правдоподобия

## Теорема

Максимум $\mathscr{L}(\Phi, \Theta)$ удовлетворяет системе уравнений относительно основных переменных $\phi_{wt}$, $\theta_{td}$
и вспомогательных переменных $p_{tdw}$, $n_{wt}$, $n_{td}$

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_{t'} \phi_{wt'}\theta_{t'd}}; \\ \text{M-шаг:} & \phi_{wt} = \dfrac{n_{wt}}{\sum_{w'} n_{w't}}; \quad n_{wt} = \sum\limits_{d \in D} n_{dw} p_{tdw}; \\ & \theta_{td} = \dfrac{n_{td}}{\sum_{t'} n_{t'd}}; \quad n_{td} = \sum\limits_{w \in d} n_{dw} p_{tdw}; \end{cases}$$

EM-алгоритм — это чередование E- и M-шага до сходимости.
Это метод простых итераций для решения системы уравнений
[Hofmann 1999], [Asuncion 2009]

# Вероятностная интерпретация шагов EM-алгоритма

**E-шаг — это формула Байеса:**

$$p_{tdw} = p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T}\phi_{ws}\theta_{sd}}$$

$n_{dwt} = n_{dw}p(t|d,w)$ — оценка числа троек $(d,w,t)$ в коллекции

**M-шаг — это частотные оценки условных вероятностей:**

$$\phi_{wt} = \frac{n_{wt}}{n_t} \equiv \frac{\sum\limits_{d\in D} n_{dwt}}{\sum\limits_{d\in D}\sum\limits_{w\in d} n_{dwt}}, \qquad \theta_{td} = \frac{n_{td}}{n_d} \equiv \frac{\sum\limits_{w\in d} n_{dwt}}{\sum\limits_{w\in W}\sum\limits_{t\in T} n_{dwt}},$$

Краткая запись через знак пропорциональности $\propto$:

$$\phi_{wt} \propto n_{wt}; \qquad \theta_{td} \propto n_{td};$$

# Рациональный EM-алгоритм

**Идея:** E-шаг встраивается внутрь M-шага

**Вход**: коллекция $D$, число тем $|T|$, число итераций $i_{\max}$;
**Выход**: матрицы терминов тем $\Theta$ и тем документов $\Phi$;

инициализация $\phi_{wt}$, $\theta_{td}$ для всех $d \in D$, $w \in W$, $t \in T$;
**для всех** итераций $i = 1, \ldots, i_{\max}$
  $n_{wt}, n_{td}, n_t, n_d := 0$ для всех $d \in D$, $w \in W$, $t \in T$;
  **для всех** документов $d \in D$ и всех слов $w \in d$
    $p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$ для всех $t \in T$;
    $n_{wt}, n_{td}, n_t, n_d \mathrel{+}= n_{dw} p_{tdw}$ для всех $t \in T$;
  $\phi_{wt} := n_{wt}/n_t$ для всех $w \in W$, $t \in T$;
  $\theta_{td} := n_{td}/n_d$ для всех $d \in D$, $t \in T$;

Обычно $i_{\max} = 20..100$ итераций достаточно, $O(n|T|i_{\max})$

# LDA — Latent Dirichlet Allocation [Blei, Ng, Jordan 2003]

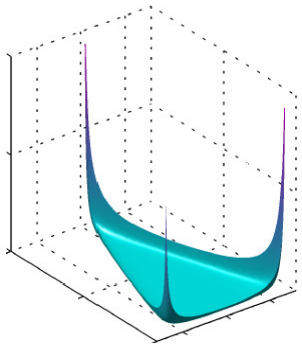
**Гипотеза.** Вектор-столбцы $\phi_t = (\phi_{wt})_{w \in W}$ и $\theta_d = (\theta_{td})_{t \in T}$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \beta_0 = \sum_w \beta_w, \ \beta_t \geqslant 0;$$

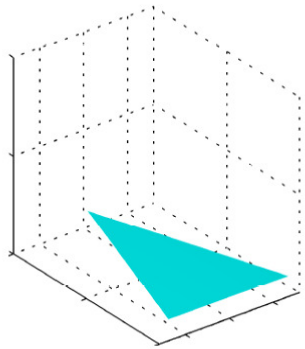
$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t \geqslant 0;$$

**Пример:**

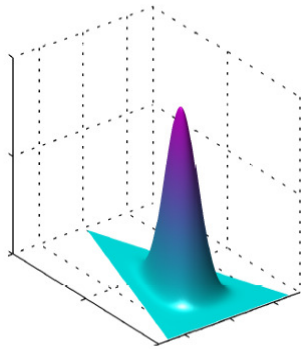
$\mathrm{Dir}(\theta|\alpha)$
$|T| = 3$
$\theta, \alpha \in \mathbb{R}^3$



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



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



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

# Главное отличие LDA от PLSA

Оценки условных вероятностей $\phi_{wt} \equiv p(w|t)$, $\theta_{td} \equiv p(t|d)$:

- в PLSA — несмещённые оценки максимума правдоподобия:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad \theta_{td} = \frac{n_{td}}{n_d}$$

- в LDA — сглаженные байесовские оценки:

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

LDA был придуман, чтобы уменьшить переобучение PLSA.
**Однако...** Различие проявляется только при малых $n_{wt}$, $n_{td}$.
Робастные LDA и PLSA строят модели одинакового качества.

---

*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.

*Potapenko A. A., Vorontsov K. V.* Robust PLSA Performs Better Than LDA. ECIR-2013, Moscow, Russia, 24-27 March 2013. LNCS, Springer. Pp. 784–787.

# Математики шутят...

Математика — часть физики. Физика — экспериментальная, естественная наука, часть естествознания. Математика — это та часть физики, в которой эксперименты дёшевы.

— *В. И. Арнольд.*

# Неустойчивость! Эксперимент на модельных данных

Модельные коллекции порождаются заданными матрицами $\Phi_0$ и $\Theta_0$ 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:

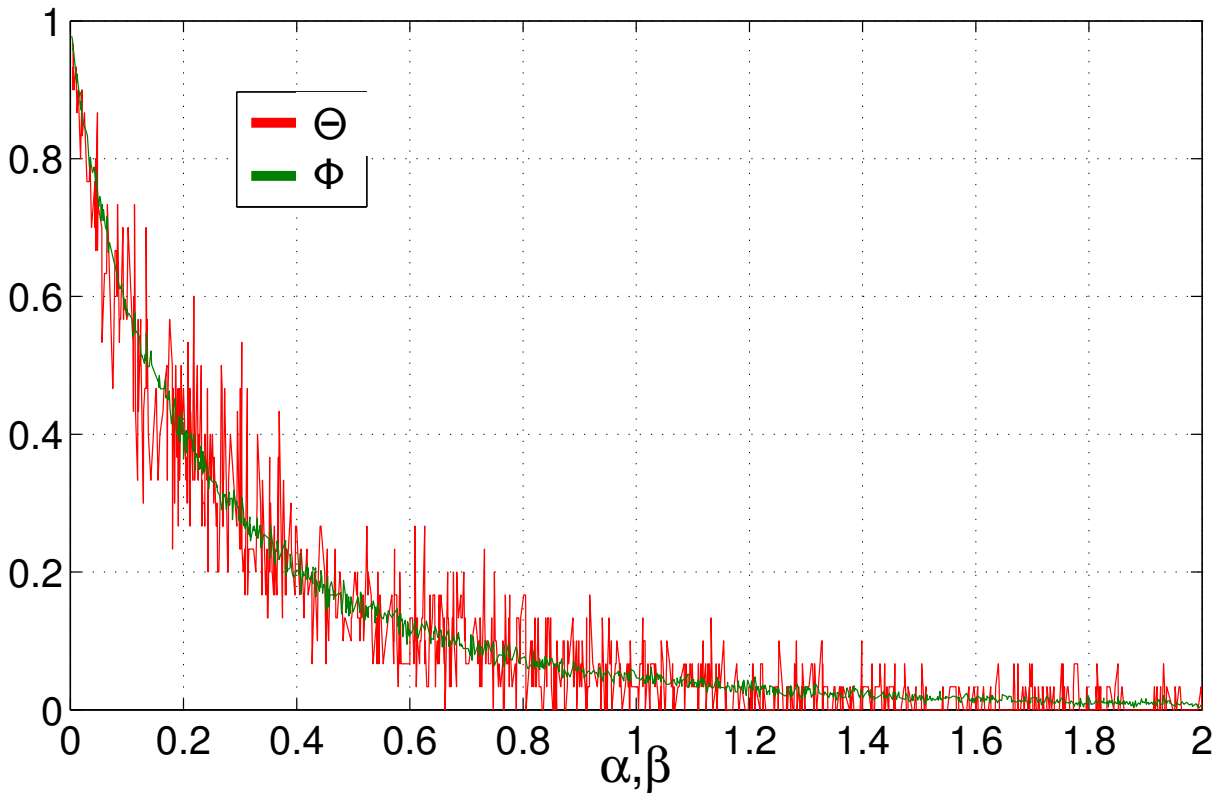
$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left( \sqrt{p(i|j)} - \sqrt{p_0(i|j)} \right)^2},$$

как для самих матриц $\Phi$ и $\Theta$, так и для их произведения:

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

# Генерация модельных данных различной разреженности

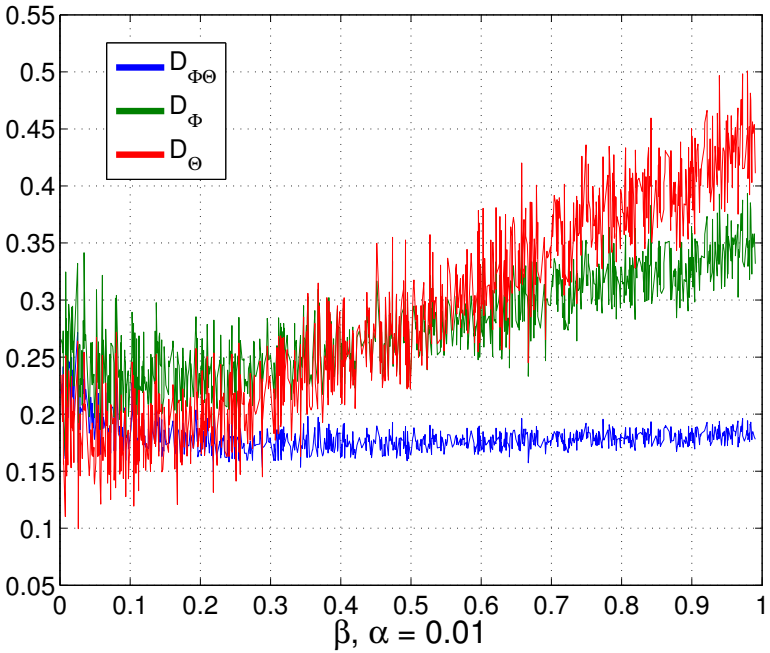
Зависимость разреженности (доли почти нулевых элементов) распределений $\theta_d^0 \sim \mathrm{Dir}(\alpha)$ и $\phi_t^0 \sim \mathrm{Dir}(\beta)$ от параметров $\alpha$ и $\beta$ симметричного распределения Дирихле:



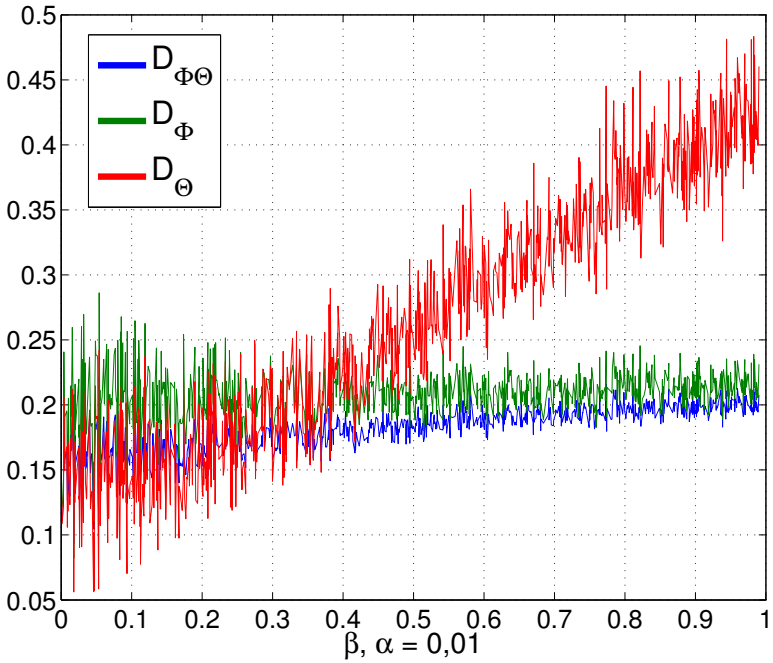
# Эксперимент: неустойчивость восстановления Φ, Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы $\Phi_0$

PLSA



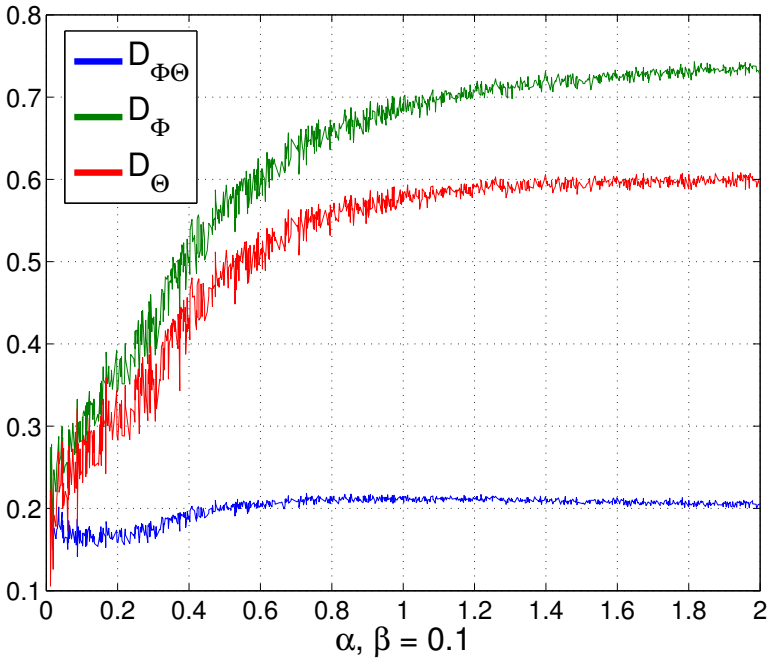

LDA



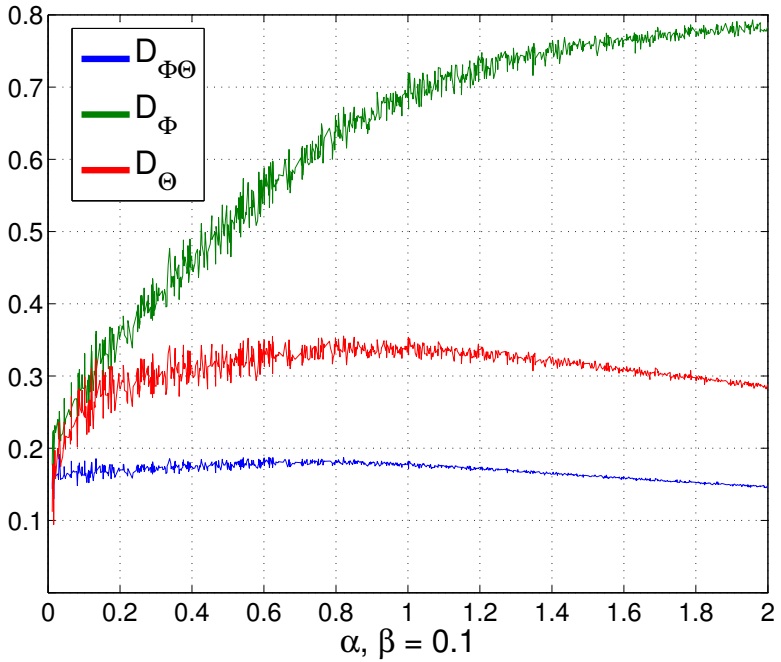
# Эксперимент: неустойчивость восстановления Φ, Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы $\Theta_0$

PLSA




LDA



# Выводы

1. Матрицы $\Phi$, $\Theta$ устойчиво восстанавливаются только при сильной разреженности $\Phi_0$, $\Theta_0$ (более 80% нулей)
2. Произведение $\Phi\Theta$ восстанавливается устойчиво, независимо от разреженности исходных $\Phi_0$, $\Theta_0$
3. Задача некорректно поставлена, нет единственности: для любых $S_{T\times T}$ таких, что $\Phi', \Theta'$ — стохастические,
$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'.$$
4. Поэтому необходима регуляризация, однако распределение Дирихле — слишком слабый регуляризатор

---

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

# Математики продолжают шутить...

«Представляется важной задача освобождения всюду, где это возможно, от излишних вероятностных допущений»

— *А. Н. Колмогоров*,
создатель современной теории вероятностей

(Теория информации и теория алгоритмов, 1987)

# Аддитивная регуляризация тематической модели

Пусть, наряду с правдоподобием, требуется максимизировать ещё $n$ критериев $R_i(\Phi, \Theta)$, $i = 1, \ldots, n$ — *регуляризаторов*.

Метод многокритериальной оптимизации — *скаляризация*.

**Задача** максимизации регуляризованного правдоподобия

$$\underbrace{\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}}_{\text{log-likelihood } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

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

где $\tau_i > 0$ — *коэффициенты регуляризации*.

# EM-алгоритм с регуляризацией M-шага

## Теорема

Максимум $\mathscr{L}(\Phi,\Theta)+R(\Phi,\Theta)$ удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw}$, $n_{wt}$, $n_{td}$,

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_{t'}\phi_{wt'}\theta_{t'd}}; \\ \text{M-шаг:} & \phi_{wt} \propto \left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\right)_+; \quad n_{wt} = \sum\limits_{d\in D} n_{dw}p_{tdw}; \\ & \theta_{td} \propto \left(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\right)_+; \quad n_{td} = \sum\limits_{w\in d} n_{dw}p_{tdw}; \end{cases}$$

где $(x)_+ = \max(x,0)$ — операция положительной срезки.

PLSA: $R(\Phi,\Theta) = 0$

LDA: $R(\Phi,\Theta) = \sum\limits_{t,w}\beta_w \ln\phi_{wt} + \sum\limits_{d,t}\alpha_t \ln\theta_{td}$

# Справочные сведения. Условия Каруша–Куна–Таккера

Задача математического программирования:

$$\begin{cases} f(x) \to \min\limits_x; \\ g_i(x) \leqslant 0, \quad i = 1, \ldots, m; \\ h_j(x) = 0, \quad j = 1, \ldots, k. \end{cases}$$

Необходимые условия. Если $x$ — точка локального минимума, то существуют множители $\mu_i$, $i = 1, \ldots, m$, $\lambda_j$, $j = 1, \ldots, k$:

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{j=1}^{k} \lambda_j g_j(x); \\ g_i(x) \leqslant 0; \ h_j(x) = 0; \ \text{(исходные ограничения)} \\ \mu_i \geqslant 0; \ \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; \ \text{(условие дополняющей нежёсткости)} \end{cases}$$

# Доказательство Теоремы о регуляризации M-шага

1. Условия ККТ для $\phi_{wt}$:

$$\sum_d n_{dw} \frac{\theta_{td}}{p(w|d)} + \frac{\partial R}{\partial \phi_{wt}} = \lambda_t - \lambda_{wt}; \quad \lambda_{wt} \geqslant 0; \quad \lambda_{wt}\phi_{wt} = 0.$$

2. Умножим обе части равенства на $\phi_{wt}$ и выделим $p_{tdw}$:

$$\phi_{wt}\lambda_t = \sum_d n_{dw} \frac{\phi_{wt}\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} = n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}.$$

3. Учтём ограничение $\phi_{wt} \geqslant 0$ и предположение $\lambda_t > 0$:

$$\phi_{wt}\lambda_t = \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

4. Суммируем обе части равенства по $w \in W$:

$$\lambda_t = \sum_{w \in W} \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

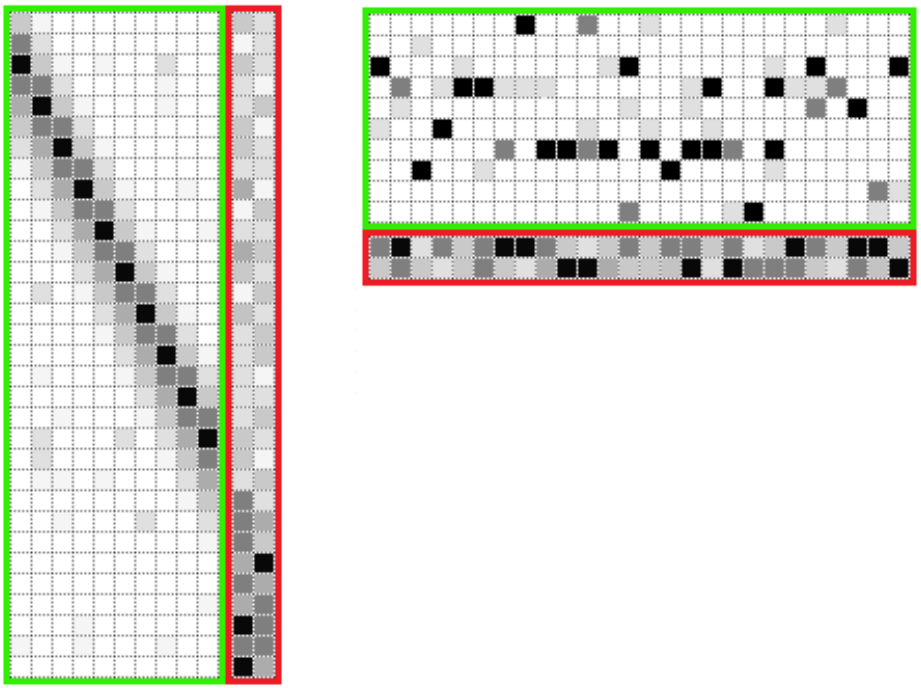
5. Подставим $\lambda_t$ из (4) в (3), получим требуемое. ■

# Гипотеза о структуре интерпретируемых тем

$S \subset T$ — предметные темы содержат термины предметных областей, разрежены, существенно различны.

$B \subset T$ — фоновые темы содержат слова общей лексики.

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



# Справочные сведения. Дивергенция Кульбака–Лейблера

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

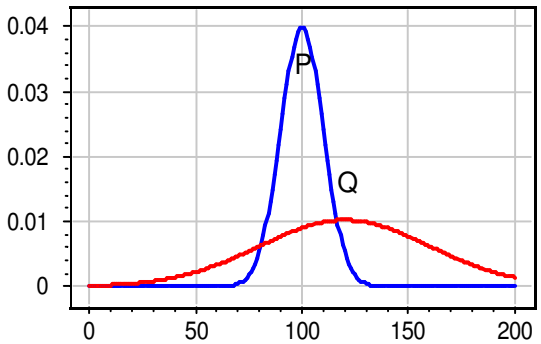
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

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

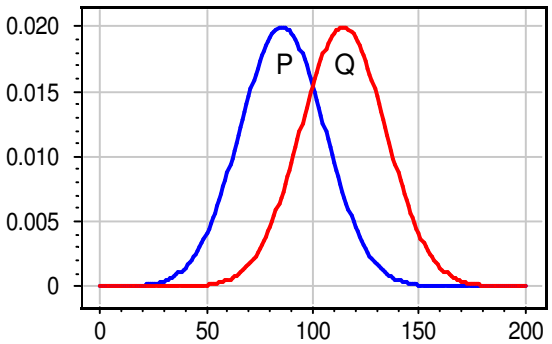
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

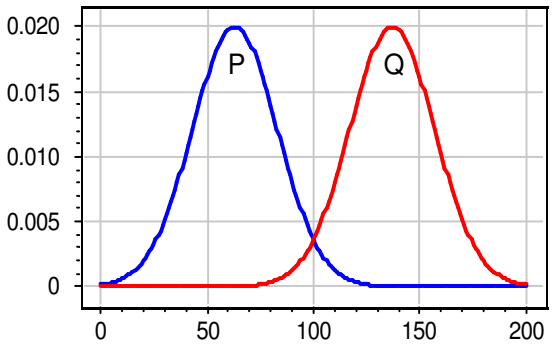
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то $P$ сильнее вложено в $Q$, чем $Q$ в $P$:




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




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




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

# Регуляризатор сглаживания (переосмысление LDA)

**Гипотеза** сглаженности фоновых тем $t \in B$:
распределения $\phi_{wt}$ близки к заданному распределению $\beta_w$
распределения $\theta_{td}$ близки к заданному распределению $\alpha_t$

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

Максимизируем сумму этих регуляризаторов:

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

Подставляем, получаем формулы M-шага LDA, для всех $t \in B$:

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

Это новая, не-байесовская интерпретация LDA [Blei 2003].

# Физики не шутят...

«Я понимаю явление, если нахожу ему несколько объяснений»

— *Р. Фейнман*

# Регуляризатор для разреживания предметных тем

**Гипотеза** разреженности предметных тем $t \in S$:
среди $\phi_{wt}$, $\theta_{td}$ много нулевых значений.

Максимизируем дивергенцию между заданными распределениями $\beta_w$, $\alpha_t$ и искомыми $\phi_{wt}$, $\theta_{td}$:

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

Подставляем, получаем «анти-LDA» для всех $t \in S$:

$$\phi_{wt} \propto \left(n_{wt} - \beta_0\beta_w\right)_+, \qquad \theta_{td} \propto \left(n_{td} - \alpha_0\alpha_t\right)_+.$$

---

*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.

# Регуляризатор для декоррелирования предметных тем

**Гипотеза** некоррелированности предметных тем $t \in S$:
чем различнее темы, тем лучше они интерпретируются.

Минимизируем ковариации между вектор-столбцами $\phi_t$:

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

Подставляем, получаем ещё один вариант разреживания —
постепенное контрастирование строк матрицы $\Phi$:

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

---

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

# Регуляризатор для сокращения числа тем

**Гипотеза:** если в теме слишком мало слов, то она не нужна.

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

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

Подставляем, получаем:

$$\theta_{td} \propto \left( n_{td} - \tau \frac{n_d}{n_t} \theta_{td} \right)_+.$$

**Эффект:**
строки матрицы $\Theta$ могут целиком обнуляться для тем $t$, собравших мало слов по коллекции, $n_t = \sum_d \sum_w n_{dwt}$.

# Эксперимент по комбинированию регуляризаторов

**Задача:** улучшить интерпретируемость, не ухудшив перплексию

**Набор регуляризаторов:**

1. сглаживание фоновых тем — столбцов $\Phi$, строк $\Theta$
2. разреживание предметных тем — столбцов $\Phi$, строк $\Theta$
3. декоррелирование предметных тем — столбцов $\Phi$
4. удаление незначимых тем — строк $\Theta$

**Данные:** NIPS (Neural Information Processing System)

- $|D| = 1566$ 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: $|D'| = 174$.

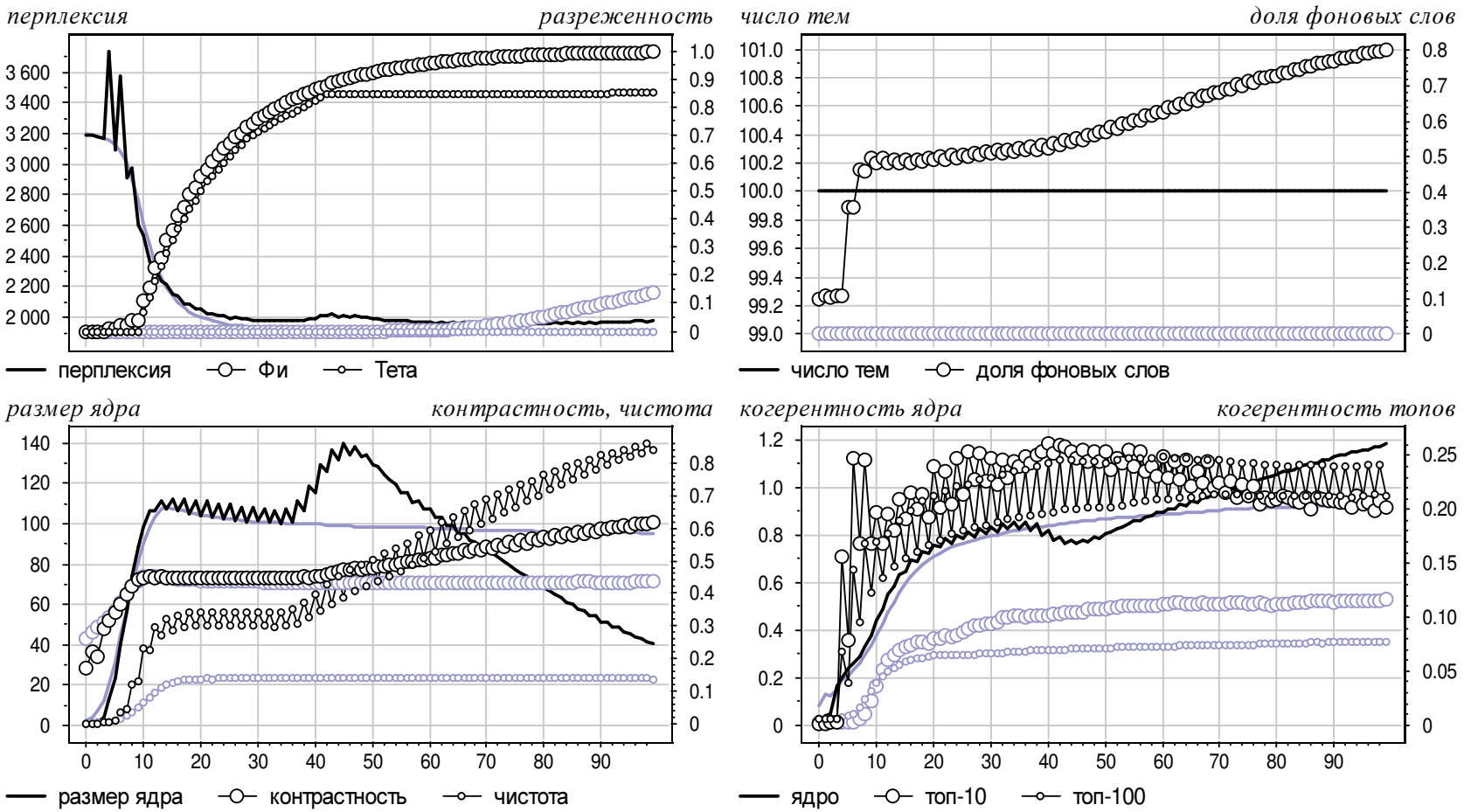
# Критерии качества модели

Построение BTM — многокритериальная оптимизация.
Поэтому критериев для контроля качества модели тоже много.

- Перплексия контрольной коллекции: $\mathcal{P} = \exp(-\frac{1}{n}\mathscr{L})$
- Разреженность — доля нулевых элементов в $\Phi$ и $\Theta$
- Характеристики интерпретируемости тем:
  - когерентность темы: [Newman, 2010]
  - размер ядра темы: $|W_t|$, ядро $W_t = \{w\colon p(t|w) > 0.25\}$
  - чистота темы: $\sum\limits_{w \in W_t} p(w|t)$
  - контрастность темы: $\frac{1}{|W_t|} \sum\limits_{w \in W_t} p(t|w)$
- Вырожденность тематической модели:
  - число тем: $|T|$
  - доля фоновых слов: $\frac{1}{n} \sum\limits_{d \in D} \sum\limits_{w \in d} \sum\limits_{t \in B} p(t|d,w)$

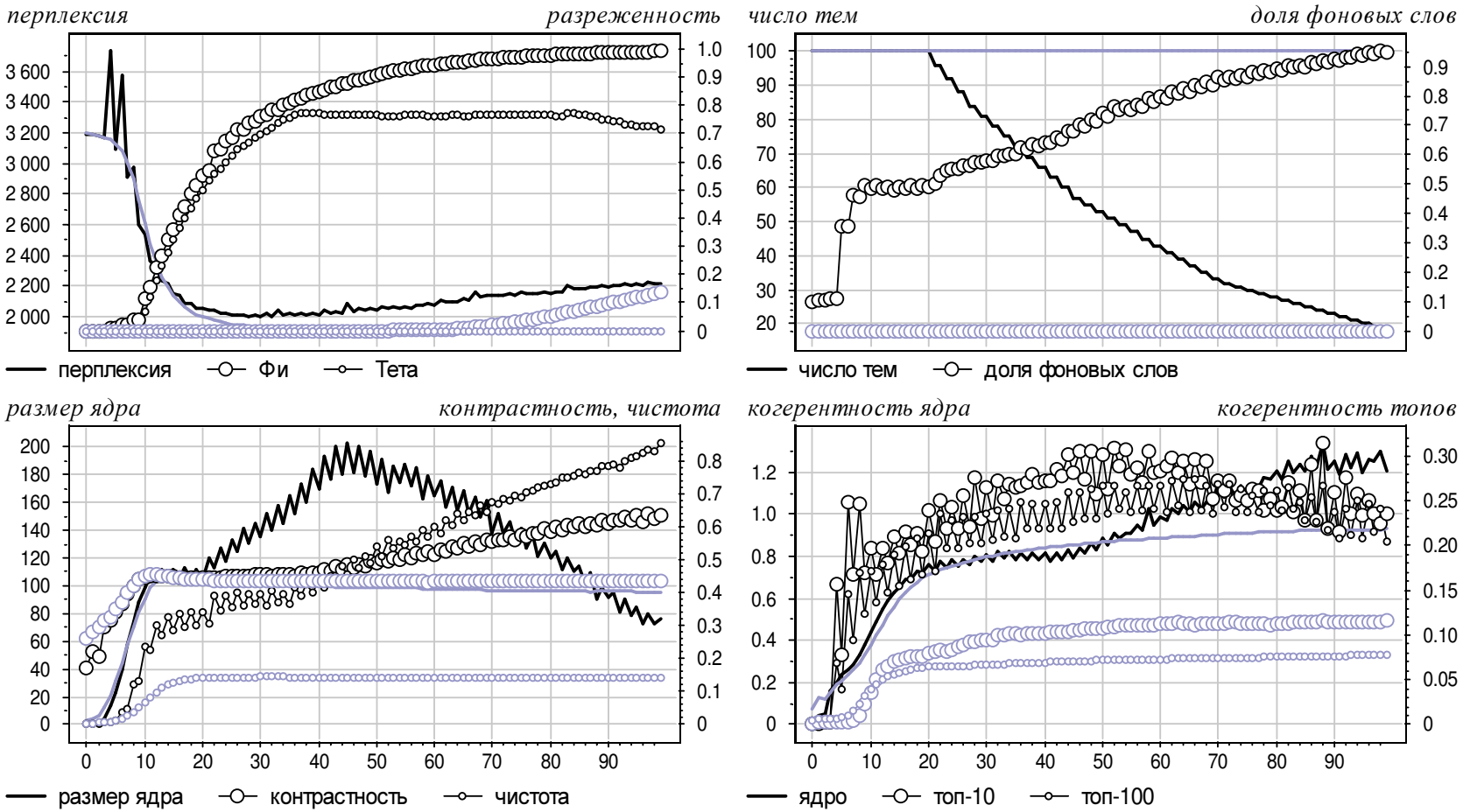
# Разреживание, сглаживание, декорреляция, сокращение тем

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



# Все те же, с удалением незначимых тем

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



# Выводы

**Одновременное улучшение многих показателей:**

- *разреженность* выросла от 0 до 95%–98%
- *когерентность тем* выросла от 0.1 до 0.3
- *чистота тем* выросла от 0.15 до 0.8
- *контрастность тем* выросла от 0.4 до 0.6
- *размер ядер тем* вырос от 0 до 150 терминов
- почти без потери *перплексии* (правдоподобия) модели

**Подобраны траектории регуляризации:**

- разреживание включать постепенно после 10-20 итераций
- сглаживание включать сразу
- декорреляцию включать сразу и как можно сильнее
- сокращение числа тем включать постепенно,
- никогда не совмещая с декорреляцией на одной итерации

# Лингвистический анализ электрокардиосигналов

**Дано:**
20 тысяч кодограмм ЭКГ (строки в 6-буквенном алфавите), каждая отнесена к некоторым из 40 заболеваний,
**важно** учесть случаи сочетания заболеваний.

**Найти:**

- темы классов (диагностические эталоны заболеваний)
- алгоритм классификации (диагностики заболеваний)

**Регуляризаторы:**

- разреживание, сглаживание, антикоррелирование
- привязка документов к классам (категоризация)
- учёт различий в степени доверия диагнозам
- учёт несбалансированности классов

# Регуляризатор для классификации документов

Пусть $C$ — множество классов (для ЭКГ — заболевания, для текстов — категории, авторы, ссылки, годы, читатели)

**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 для классификация документов

EM-алгоритм дополняется оцениванием параметров $\psi_{ct}$.

**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:

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

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

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

# Регуляризатор для категоризации документов

Снова регуляризатор для классификации:

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

**Недостаток:** за «эмпирическую частоту классов» не вполне обоснованно принимается равномерное распределение:

$$m_{dc} = n_d \tfrac{1}{|C_d|}[c \in C_d]$$

**Ковариационный регуляризатор:**

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

приводит к естественному аналитическому решению

$$\psi_{ct} = \big[c = c^*(t)\big], \quad c^*(t) = \arg\max_{c\in C} \sum_{d\in D} m_{dc}\theta_{td}$$

**Эффект:** Каждая категория $c$ распадается на свои темы.

# Задача анализа потока пресс-релизов

**Дано:** коллекция пресс-релизов МИДов ряда стран.
Более 40 тыс. сообщений, 180Мб текста.

**Найти:**

- какие темы общие, какие специфичны для источников?
- какие темы «вечные», а какие привязаны к событиям?
- какие темы, и когда коррелируют с заданной темой?

**Регуляризаторы:**

- разреживание, сглаживание, антикоррелирование
- привязка документов к источникам и моментам времени
- сглаживание тематик во времени
- частичное обучение: привязка ключевых терминов к темам

# Регуляризаторы для динамической тематической модели

$Y$ — моменты времени (например, годы публикаций),
$y(d)$ — метка времени документа $d$,
$D_y \subset D$ — все документы, относящиеся к моменту $y \in Y$.

**Гипотеза 1:** распределение $p(t|y) = \sum_{d \in D_y} \theta_{td} p(d)$ разрежено:

$$R_1(\Theta) = -\tau_1 \sum_{y \in Y} \sum_{t \in T} \ln p(t|y) \to \max.$$

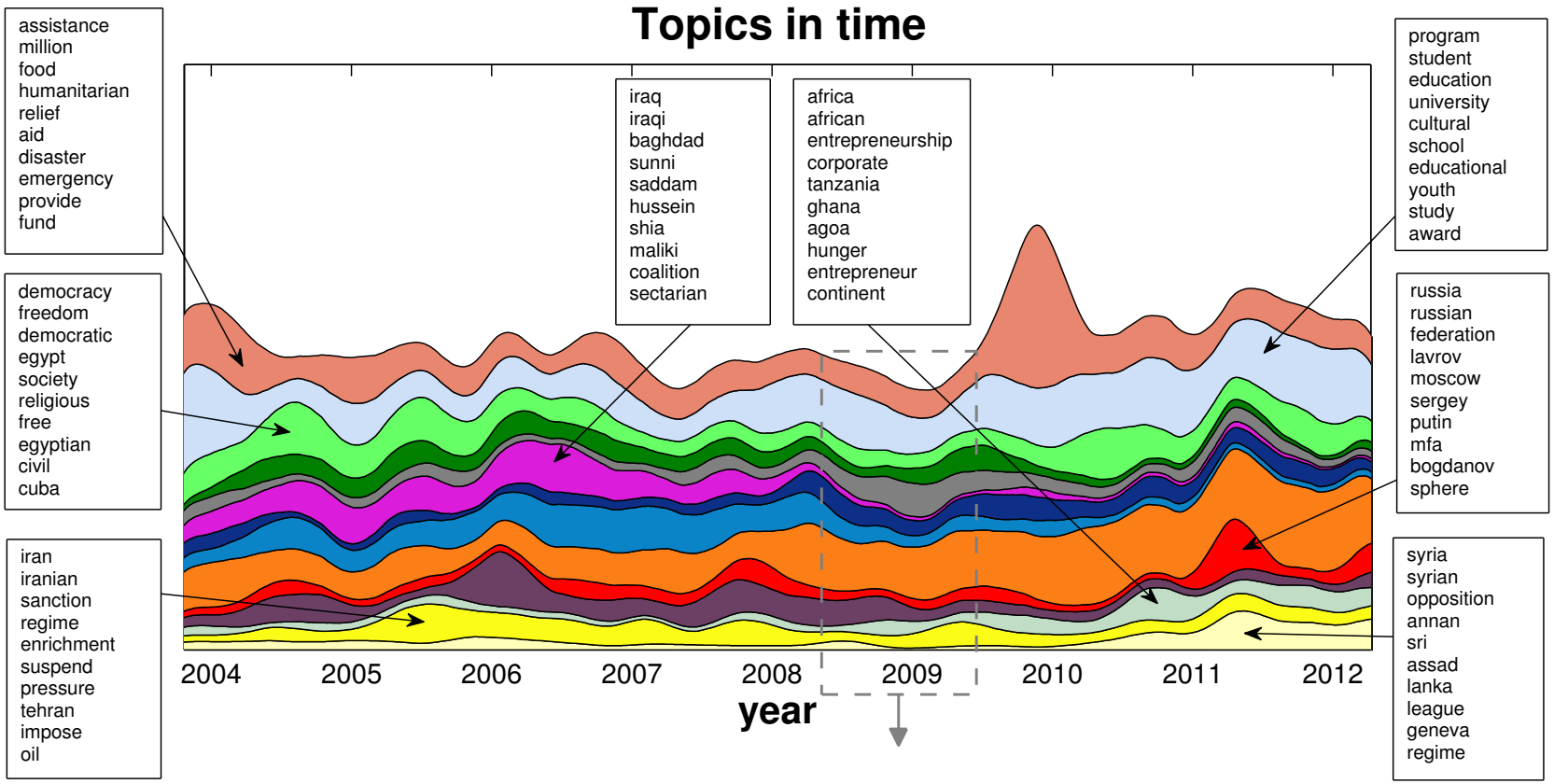
**Эффект** — разреживание тем $t$ с малым $p(t|y(d))$:

$$\theta_{td} \propto \left( n_{td} - \tau_1 \frac{\theta_{td} p(d)}{p(t|y(d))} \right)_+.$$

**Гипотеза 2:** $p(t|y)$ меняются плавно, с редкими скачками:

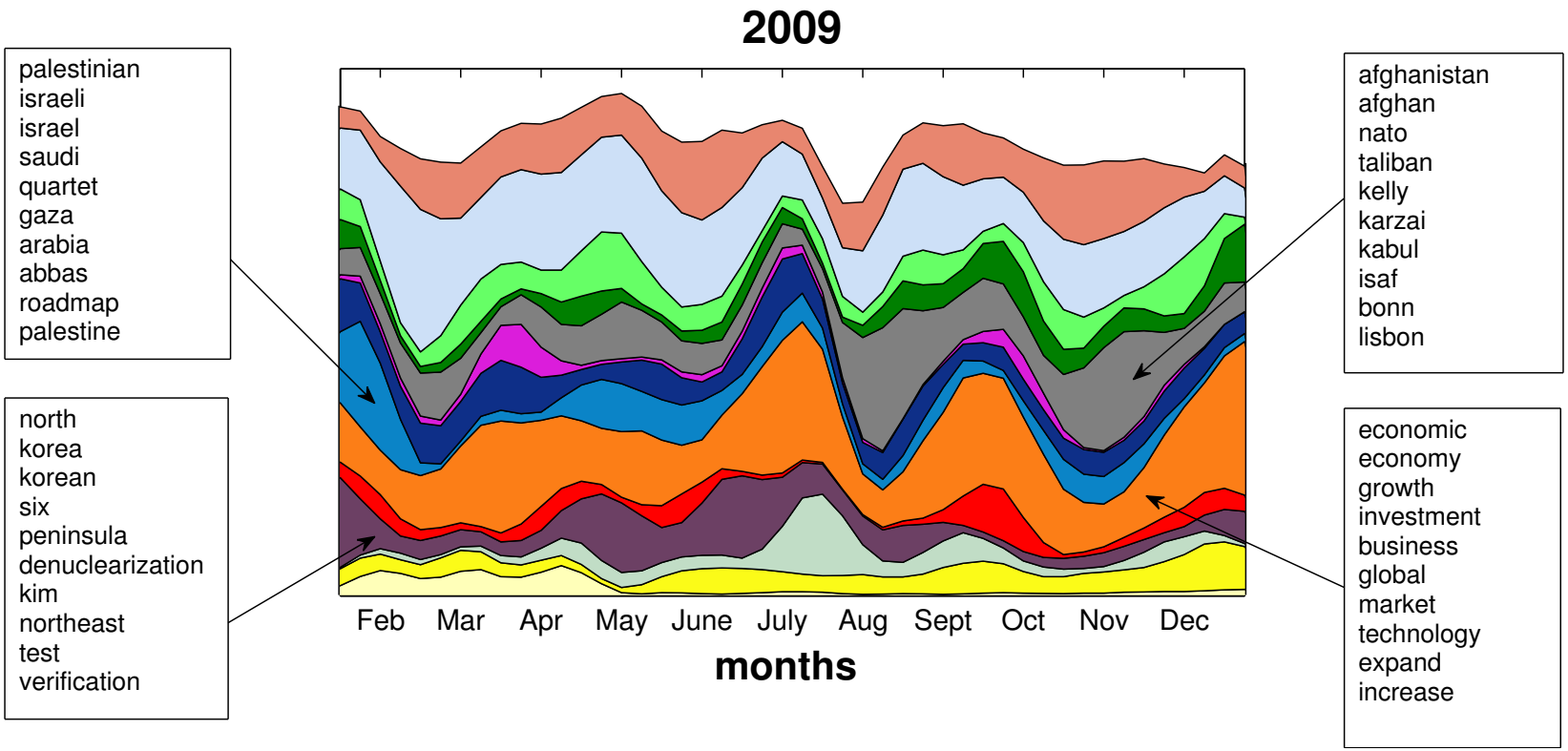
$$R_2(\Theta) = -\tau_2 \sum_{y \in Y} \sum_{t \in T} \left| p(t|y) - p(t|y-1) \right| \to \max.$$

# Эксперименты с динамической тематической моделью




*Никита Дойков*. Курсовая работа, ВМК МГУ, 2014

# Эксперименты с динамической тематической моделью




*Никита Дойков*. Курсовая работа, ВМК МГУ, 2014

# Литература


- *Hofmann T.* Probabilistic Latent Semantic Indexing. SIGIR, 1999.
- *Blei D., Ng A., Jordan M.* Latent Dirichlet Allocation // Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.
- *Teh Y. W., Newman D., Welling M.* A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. NIPS, 2006, Pp. 1353–1360.
- *Porteous I., Newman D., Ihler A., Asuncion A., Smyth P., Welling M.* Fast Collapsed Gibbs Sampling For Latent Dirichlet Allocation. KDD 2008.
- *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.
- *Newman D., Lau J. H., Grieser K., Baldwin T.* Automatic Evaluation of Topic Coherence // Human Language Technologies, HLT-2010, Pp. 100–108.
- *Yi Wang.* Distributed Gibbs Sampling of Latent Dirichlet Allocation: The Gritty Details. 2011.
- *Sato I., Nakagawa H.* Rethinking Collapsed Variational Bayes Inference for LDA. Int'l Conf. on Machine Learning ICML, 2012.
- *Vorontsov K. V.* Additive Regularization for Topic Models of Text Collections // Doklady Mathematics. Pleiades Publisher, 2014. Vol. 88, No. 3.
- *Vorontsov K. V., Potapenko A. A.*, Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // AIST'14. Springer. 2014. (to appear)

Воронцов Константин Вячеславович
voron@forecsys.ru

Страницы на вики www.MachineLearning.ru:

- Участник:Vokov
- Вероятностные тематические модели (курс лекций, К. В. Воронцов)
- Тематическое моделирование

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

Почему в регуляризаторе разреживания

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

не возникает проблем с $\ln \phi_{wt}$ при $\phi_{wt} \to 0$?

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

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

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

$$\phi_{wt} \propto \left( n_{wt} - \beta_0 \beta_w \frac{\phi_{wt}}{\phi_{wt}+\varepsilon} \right)_+$$

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

# LDA. Принцип максимума апостериорной вероятности

$$\ln \prod_{d\in D}\prod_{w\in d} p(d,w)^{n_{dw}} \prod_{t\in T} \mathrm{Dir}(\phi_t;\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d;\alpha) \;\to\; \max_{\Phi,\Theta}$$

**Задача** максимизации регуляризованного правдоподобия:

$$\tilde{\mathscr{L}}(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + \\ + \sum_{t\in T}\sum_{w\in W} (\beta_w - 1)\,[\phi_{wt} > 0] \ln \phi_{wt} + \\ + \sum_{d\in D}\sum_{t\in T} (\alpha_t - 1)\,[\theta_{td} > 0] \ln \theta_{td} \;\to\; \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

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

# EM-алгоритм для максимизации апостериорной вероятности

## Теорема

Максимум $\tilde{\mathscr{L}}(\Phi, \Theta)$ удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw}$, $n_{wt}$, $n_{td}$,

$$\text{E-шаг:} \quad p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{t'} \phi_{wt'}\theta_{t'd}};$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \dfrac{\big(n_{wt} + \beta_w - 1\big)_+}{\sum_{w'} \big(n_{w't} + \beta_{w'} - 1\big)_+}; & n_{wt} = \sum\limits_{d \in D} n_{dw} p_{tdw}; \\ \theta_{td} = \dfrac{\big(n_{td} + \alpha_t - 1\big)_+}{\sum_{t'} \big(n_{t'd} + \alpha_{t'} - 1\big)_+}; & n_{td} = \sum\limits_{w \in d} n_{dw} p_{tdw}; \end{cases}$$

где $(x)_+ = \max(x, 0)$ — операция положительной срезки.

EM-алгоритм — это чередование E- и M-шага до сходимости.
Это метод простых итераций для решения системы уравнений.