

# Вероятностные тематические модели без интегралов и распределений Дирихле

Воронцов Константин Вячеславович  
ВЦ РАН • МФТИ • ВМК МГУ • ШАД Яндекс

Семинар в одной крупной компании • 21 марта 2014

## Содержание

- 1 Основы вероятностного тематического моделирования**
  - Цели, задачи, проблемы
  - Формализация
  - Модели PLSA и LDA
- 2 Обзор тематических моделей**
  - Время. N-граммы. Цитирования.
  - Взаимосвязи тем. Иерархии тем.
  - Многоязычность. Визуализация. Графмодели.
- 3 Аддитивная регуляризация тематических моделей**
  - Проблема неединственности решения
  - Общая формула и примеры регуляризаторов
  - Эксперименты с многокритериальными моделями

## 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:

$p(w|t)$  — какими терминами  $w$  определяется каждая тема  $t$

$p(t|d)$  — к каким темам  $t$  относится каждый документ  $d$

### Критерии:

*внутренний* — точность описания коллекции моделью  $p(w|d)$

*внешний* — качество решения конечной задачи

## 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
- Рубрикация документов, изображений, видео, музыки
- Рекомендательные сервисы (коллаборативная фильтрация)
- Аннотация генома и другие задачи биоинформатики

## Дополнительная информация

- повышает адекватность тематической модели
- позволяет выявить темы нетекстовых объектов

### Виды дополнительной информации:

- последовательность слов документа  $d$ :  $\{w_1, \dots, w_{n_d}\}$
- разбиение документа на предложения, разделы
- метаданные: год, авторы, источник, и т.д.
- цитаты и/или гиперссылки: исходящие, входящие
- рубрикатор(ы)
- словари, тезаурусы, онтологии предметных областей
- изображения внутри документов
- именованные сущности в тексте документов
- пользователи документов
- теги, ключевые слова, привязанные к документам

## Проблемы

- как строить комбинированные и многоцелевые модели?
- как учитывать все дополнительные данные сразу?
- как обеспечивать интерпретируемость тем?
- как определять правильное число тем?
- как восстанавливать иерархию тем?
- как автоматически именовать темы?
- как учитывать лингвистические знания?
- как строить модели на десятки тысяч тем?
- как строить модели сверхбольших коллекций?
- как делать визуализацию и навигацию по темам?

## 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 $d$ :

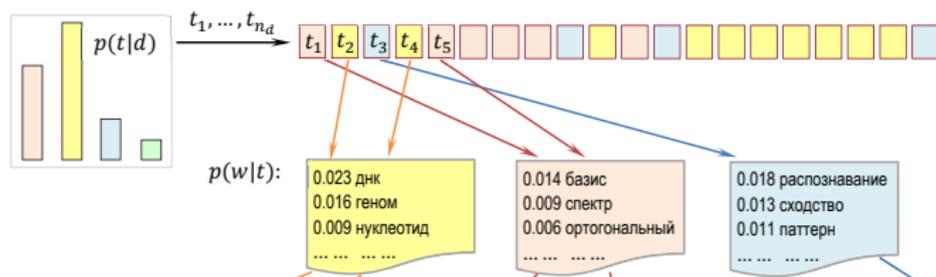
$$p(w|d) = \sum_{t \in T} p(w|d, t) p(t|d) = \sum_{t \in T} p(w|t) p(t|d)$$

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



$w_1, \dots, w_{n_d}$ :

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).

## Задача максимизации правдоподобия

**Задача:** найти максимум правдоподобия

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

при ограничениях неотрицательности и нормировки

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

**Интерпретация:** найти стохастическое матричное разложение

$$\|F - \Phi\Theta\|_{\text{KL}} \rightarrow \min_{\Phi, \Theta}$$

$F = (\hat{p}(w|d))_{W \times D}$  — известная матрица исходных данных,

$\Phi = (\phi_{wt})_{W \times T}$  — искомая матрица терминов тем  $\phi_{wt} = p(w|t)$ ,

$\Theta = (\theta_{td})_{T \times D}$  — искомая матрица тем документов  $\theta_{td} = p(t|d)$ .

## EM-алгоритм, вероятностный латентный семантический анализ PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

**E-шаг.** Выразим  $p(t|d, w)$  через  $\phi_{wt}$ ,  $\theta_{td}$  по формуле Байеса:

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

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

**M-шаг.** Частотные оценки условных вероятностей:

$$\phi_{wt} = \frac{n_{wt}}{n_t} \equiv \frac{\sum_{d \in D} n_{dwt}}{\sum_{d \in D} \sum_{w \in d} n_{dwt}}, \quad \theta_{td} = \frac{n_{dt}}{n_d} \equiv \frac{\sum_{w \in d} n_{dwt}}{\sum_{w \in W} \sum_{t \in T} n_{dwt}},$$

или краткая запись:

$$\phi_{wt} \propto n_{wt}$$

$$\theta_{td} \propto n_{dt}$$

## Недостатки классического PLSA и способы их устранения

- 1 PLSA вынужден хранить 3D-матрицу  $p(t|d, w)$ ;  
PLSA медленно сходится на больших коллекциях;  
PLSA искажает модель при добавлении документа;  
— рациональный алгоритм, онлайн-алгоритм
- 2 PLSA неверно оценивает вероятности новых слов:  
если  $n_w = 0$ , то  $\hat{p}(w|t) = 0$  для всех  $t \in T$   
— робастные модели с шумом и фоном
- 3 PLSA переобучается, т.к.  $\dim(\Phi, \Theta) = |D| \cdot |T| + |W| \cdot |T|$   
— регуляризация (LDA), привлечение внешних данных
- 4 PLSA не позволяет управлять разреженностью  $\Phi$  и  $\Theta$ , т.к.  
(в начале  $\phi_{wt} = 0$ )  $\Leftrightarrow$  (в финале  $\phi_{wt} = 0$ )  
(в начале  $\theta_{td} = 0$ )  $\Leftrightarrow$  (в финале  $\theta_{td} = 0$ )  
— эвристики постепенного разреживания

## Рациональный EM-алгоритм

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

**Вход:** коллекция  $D$ , число тем  $|T|$ , число итераций  $i_{\max}$ ;

**Выход:** распределения  $\Theta$  и  $\Phi$ ;

инициализация  $\phi_{wt}, \theta_{td}$  для всех  $d \in D, w \in W, t \in T$ ;

**для всех** итераций  $i = 1, \dots, i_{\max}$

$n_{wt}, n_{dt}, n_t, n_d := 0$  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  $t \in T$ ;

$n_{wt}, n_{dt}, n_t, n_d += n_{dw}p(t|d, w)$  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

инициализировать  $\phi_{wt}$  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  $D_j$ ,  $j = 1, \dots, J$

$\tilde{n}_{wt} := 0$ ,  $\tilde{n}_t := 0$  для всех  $w \in W$ ,  $t \in T$ ;

**для всех**  $d \in D_j$

инициализировать  $\theta_{td}$  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**  $\theta_d$  не сойдётся;

$\tilde{n}_{wt}, \tilde{n}_t += n_{dw} p(t|d, w)$  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  $w \in W$ ,  $t \in T$ ;

## Латентное размещение Дирихле

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

- в PLSA — несмещённые оценки максимума правдоподобия:

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

- в LDA — сглаженные байесовские оценки:

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

- различия между LDA и PLSA практически исчезают
  - на больших данных
  - при использовании робастного алгоритма

---

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

*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.

## Робастная модель с фоновой и шумовой компонентами

**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  
(SWB — Special Words with Background):

$$p(w|d) = \frac{p_0(w|d) + \gamma\pi_{dw} + \varepsilon\pi_w}{1 + \gamma + \varepsilon}; \quad p_0(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td},$$

$\pi_{dw} \equiv p_{\text{ш}}(w|d)$  — шумовая компонента,  $\gamma$  — параметр;

$\pi_w \equiv p_{\text{ф}}(w)$  — фоновая компонента,  $\varepsilon$  — 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.*

## Упрощённая робастная тематическая модель

Если PLSA не может предсказать слово  $w$  в документе  $d$

$$p_0(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td} = 0,$$

то естественно считать такие слова  $w$  нетематическими:

$$p(w|d) = \nu_d p_0(w|d) + [p_0(w|d) = 0] \pi_{dw},$$

где  $\pi_{dw} = \frac{n_w}{n}$  — униграммная модель языка.

Нормировочный множитель  $\nu_d$  находится аналитически:

$$\nu_d = \sum_{w \in W} [p_0(w|d) > 0] \pi_{dw}$$

## Внутренний критерий для сравнения моделей

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

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

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

параметры  $\phi_{wt}$  оцениваются по обучающей коллекции  $D$ ;

параметры  $\theta_{td}$  оцениваются по первой половине  $d'$ ;

перплексия вычисляется по второй половине  $d''$ .

**Интерпретация перплексии:**

1) чем меньше, тем лучше;

2) если документ состоит из  $n_d$  равновероятных слов, то  $\mathcal{P} = n_d$

## Эксперименты с робастными PLSA и LDA

Использовались две коллекции:

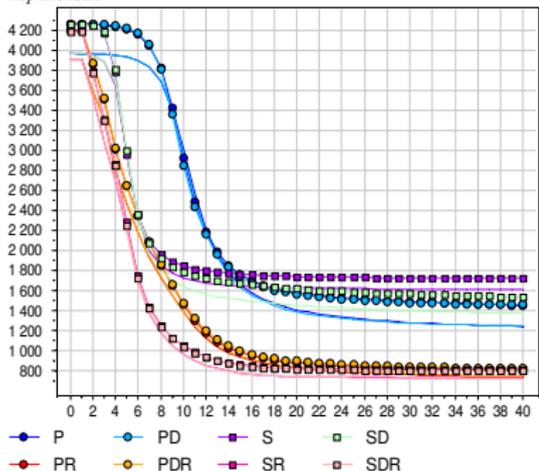
- NIPS:
  - $|D| = 1566$  статей конференции NIPS на английском языке;
  - суммарной длины  $n \approx 2.3 \cdot 10^6$ ,
  - словарь  $|W| \approx 1.3 \cdot 10^4$ .
  - Контрольная коллекция:  $|D'| = 174$ .
- RuDis:
  - $|D| = 2000$  авторефератов диссертаций на русском языке;
  - суммарной длины  $n \approx 8.7 \cdot 10^6$ ,
  - 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  $|T| = 100$ ;

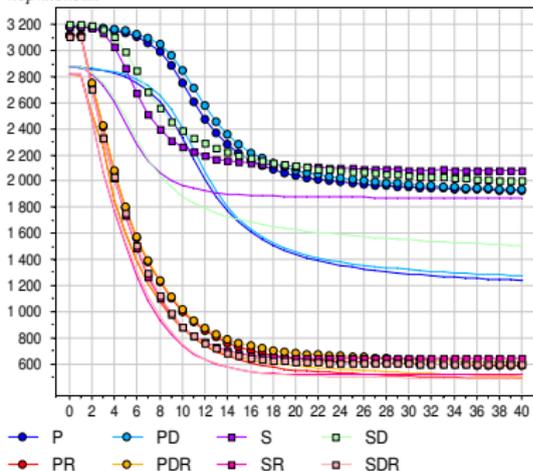
## Эксперименты с робастными PLSA и LDA

перplexия



RuDis

перplexия



NIPS

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

## Выводы

- 1 LDA не уменьшает переобучение,  
а лишь точнее моделирует вероятности редких слов
- 2 Робастный PLSA лучше, чем LDA

### Недостатки LDA:

- 1 Противоречит гипотезе разреженности
- 2 Не имеет убедительных лингвистических обоснований
- 3 Усложняет построение композитных моделей

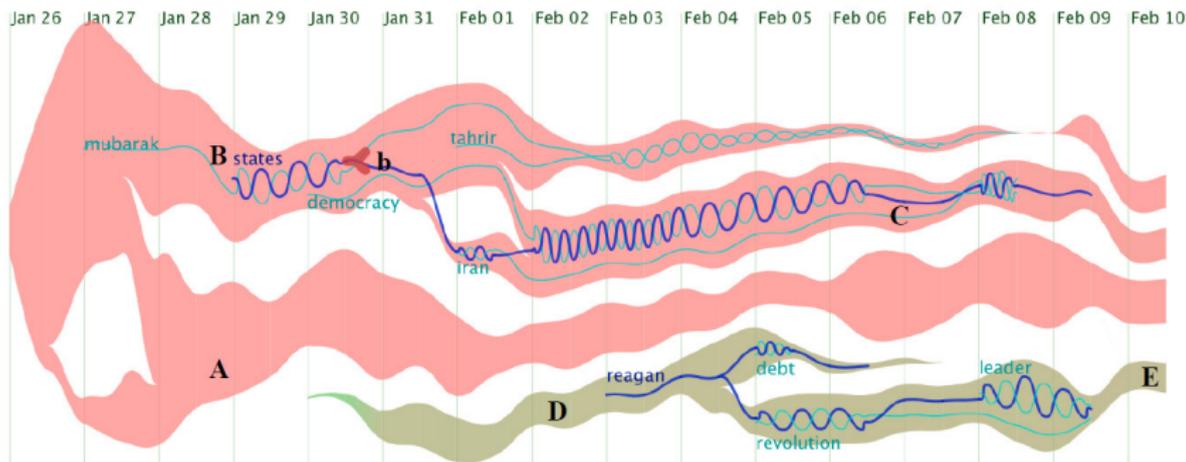
---

*Potapenko A. A., Vorontsov K. V., Robust PLSA Performs Better Than LDA // European Conference on Information Retrieval ECIR-2013, Moscow, 24–27 March 2013. — Pp. 784–787.*

*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.*

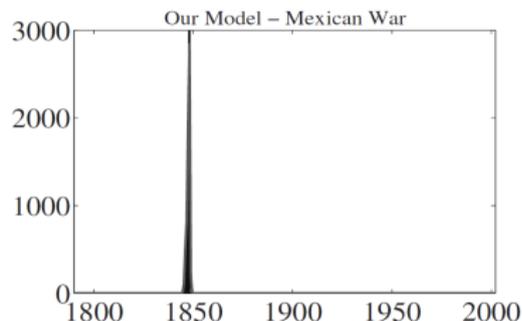
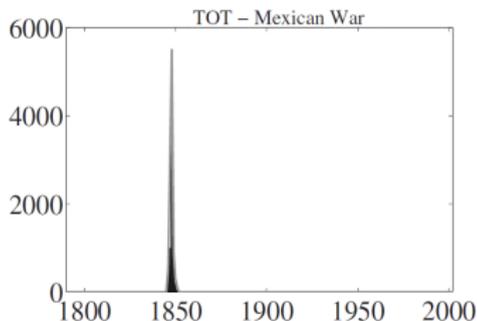


## Динамические модели эволюции тем



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

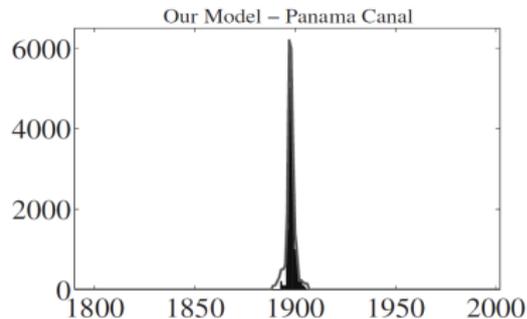
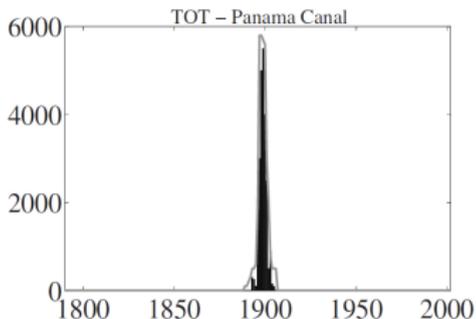


|               |              |
|---------------|--------------|
| 1. mexico     | 8. territory |
| 2. texas      | 9. army      |
| 3. war        | 10. peace    |
| 4. mexican    | 11. act      |
| 5. united     | 12. policy   |
| 6. country    | 13. foreign  |
| 7. government | 14. citizens |

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

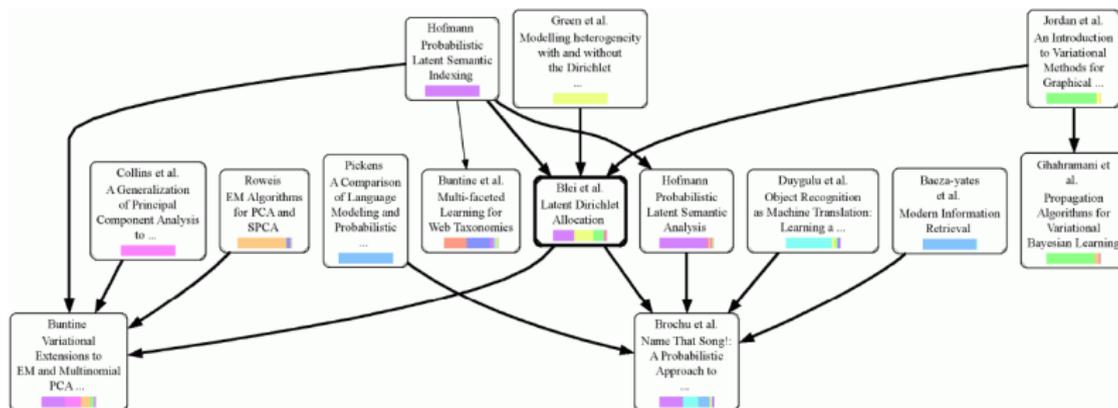
|                             |                         |
|-----------------------------|-------------------------|
| 1. panama canal             | 8. united states senate |
| 2. isthmian canal           | 9. french canal company |
| 3. isthmus panama           | 10. caribbean sea       |
| 4. republic 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

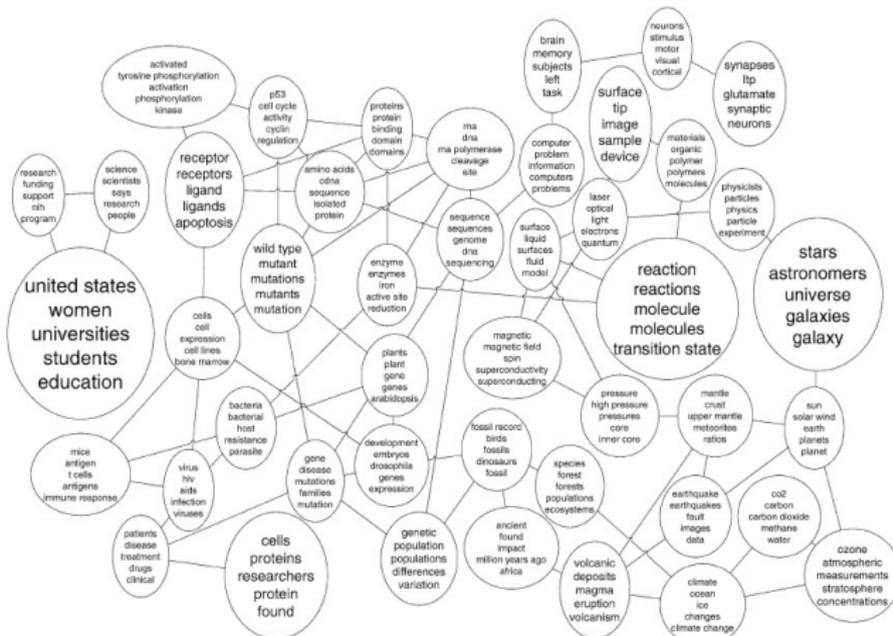
Учёт ссылок уточняет тематическую модель

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.*

## Иерархические тематические модели

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

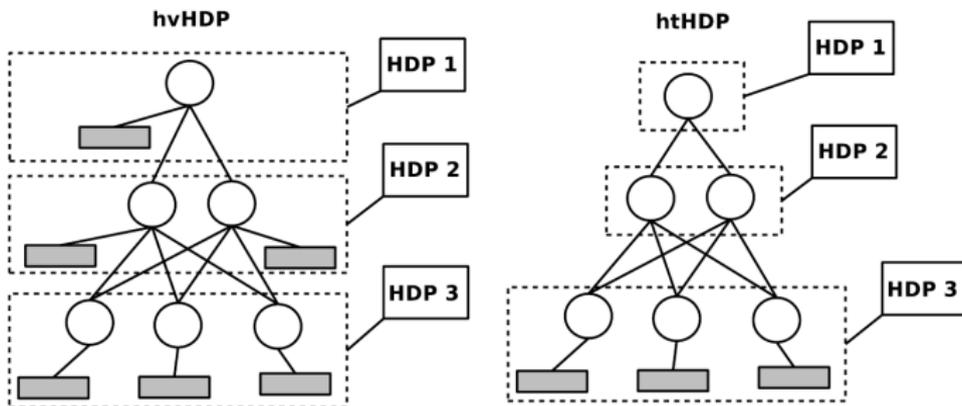
Задача построения иерархии и задача оценивания её качества признаются открытыми научными проблемами.

- “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.”

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

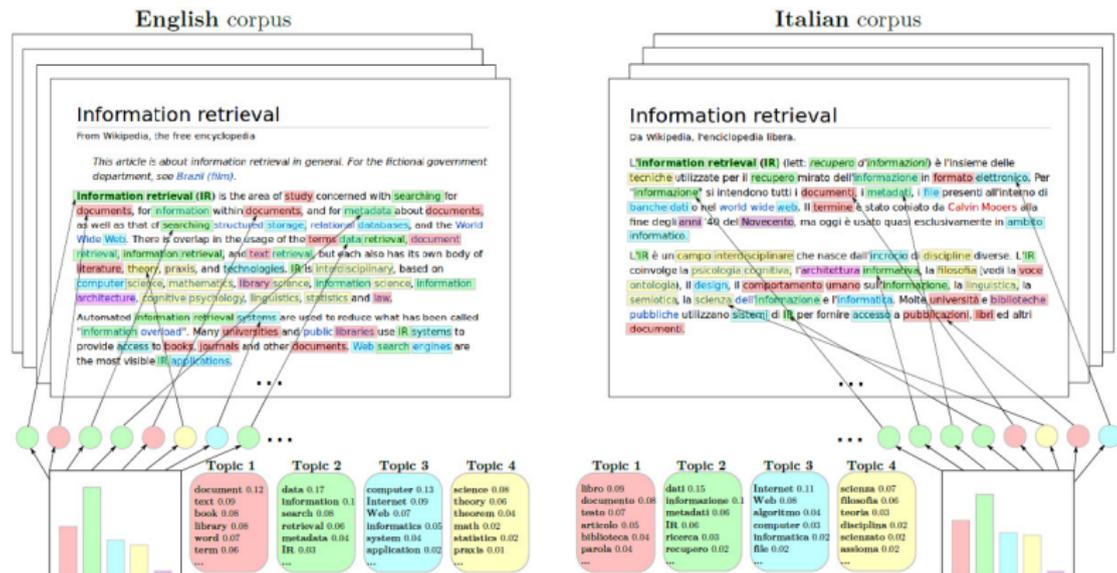
## Две восходящие стратегии построения иерархии

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



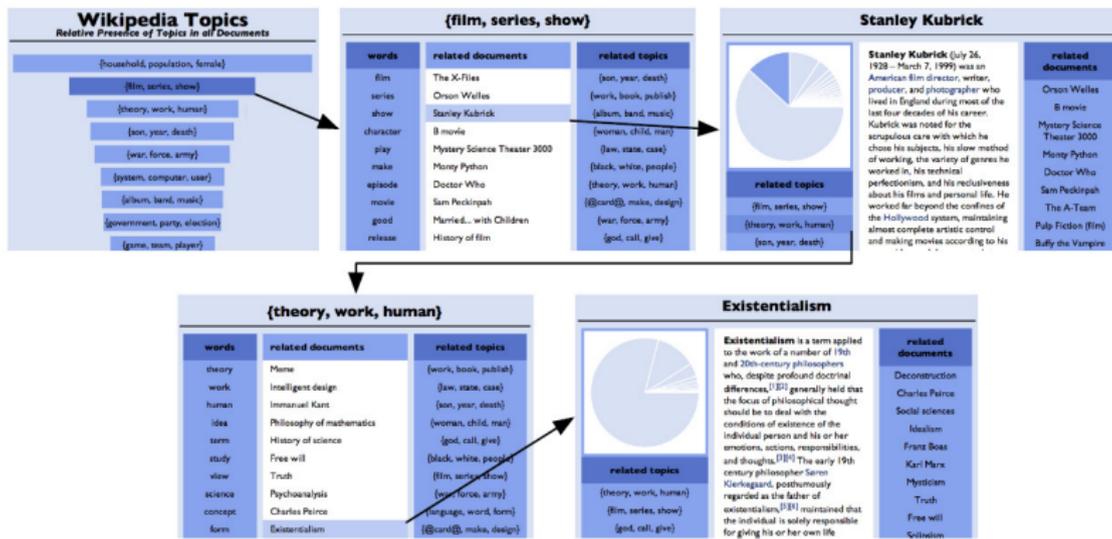
*E. Zavitsanos, G. Paliouras, G. A. Vouros. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes // Journal of Machine Learning Research 12 (2011) 2749-2775.*

## Многоязычные модели



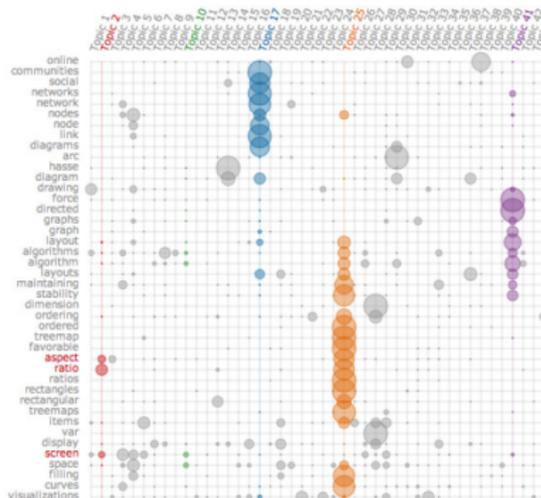
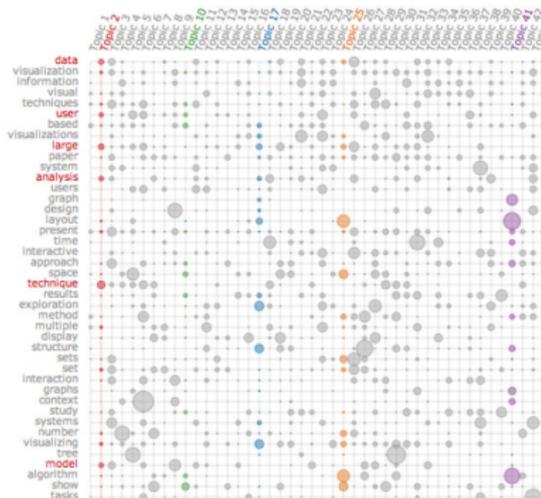
I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: a short overview of its methodology with applications // NIPS, 7–8 December 2012. — Pp. 1–11.

# Визуализация тематической модели



A. Chaney, D. Blei. Visualizing topic models // International AAAI Conference on Social Media and Weblogs, 2012.

## Визуализация тематической модели

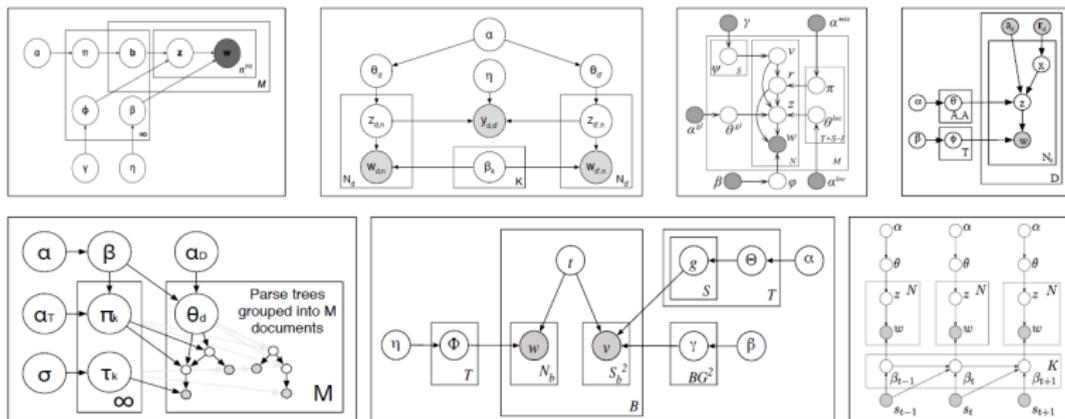


*Jason Chuang, Christopher D. Manning, Jeffrey Heer.*

Termite: Visualization Techniques for Assessing Textual Topic Models //  
Advanced Visual Interfaces, 2012

## Генеральная линия развития тематических моделей

- Основа — распределения Дирихле, процессы Дирихле
- Математический аппарат — байесовский вывод
- Язык описания — графические модели



*Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad.*

Knowledge discovery through directed probabilistic topic models: a survey.  
 Frontiers of Computer Science in China, Vol. 4, No. 2., 2010, Pp. 280–301.  
 (русский перевод на [www.MachineLearning.ru](http://www.MachineLearning.ru))

## 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:

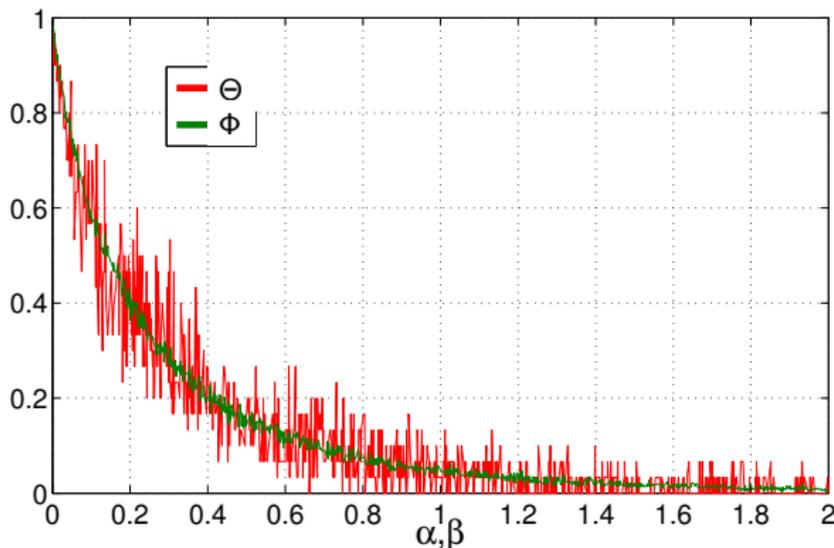
$$D_\Phi(\Phi, \Phi_0) = H(\Phi, \Phi_0);$$

$$D_\Theta(\Theta, \Theta_0) = H(\Theta, \Theta_0);$$

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

## Генерация модельных данных различной степени разреженности

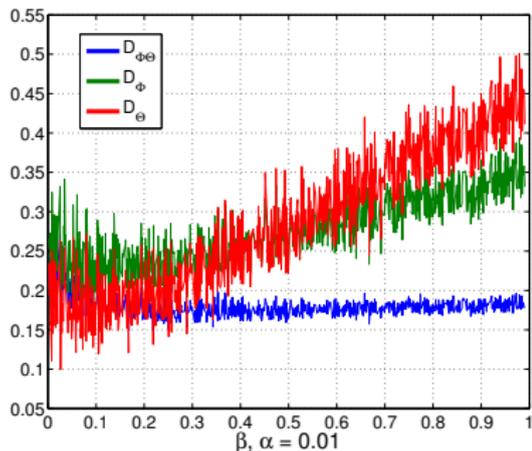
Зависимость разреженности (доли почти нулевых элементов) распределений  $\theta_d^0 \sim \text{Dir}(\alpha)$  и  $\phi_t^0 \sim \text{Dir}(\beta)$  от параметров  $\alpha$  и  $\beta$  симметричного распределения Дирихле:



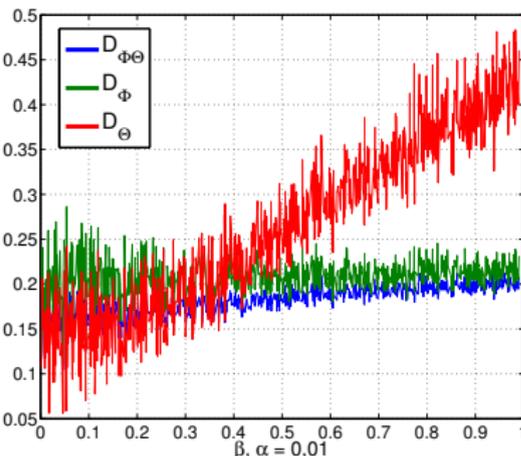
## Эксперимент: неустойчивость восстановления $\Phi$ , $\Theta$

Зависимость точности восстановления матриц  $\Phi$ ,  $\Theta$  и  $\Phi\Theta$  от разреженности матрицы  $\Phi_0$

PLSA



LDA

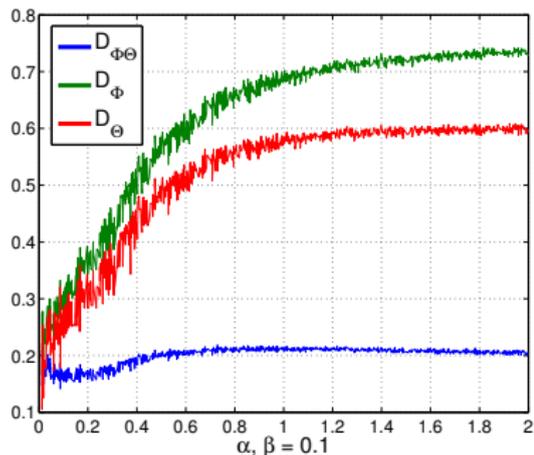


Чем больше нулей в  $\Phi$ ,  $\Theta$ , тем точнее они восстанавливаются

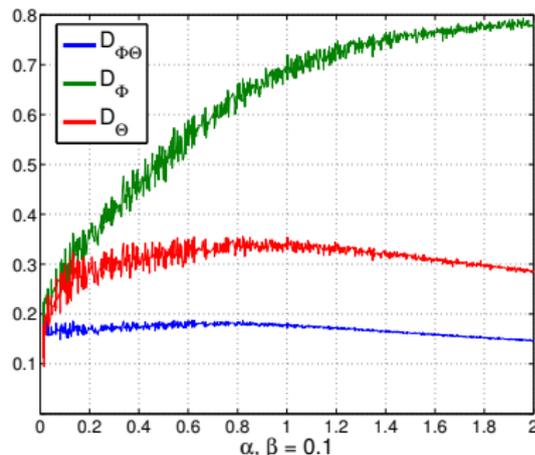
## Эксперимент: неустойчивость восстановления $\Phi$ , $\Theta$

Зависимость точности восстановления матриц  $\Phi$ ,  $\Theta$  и  $\Phi\Theta$  от разреженности матрицы  $\Theta_0$

PLSA



LDA



## Выводы

- 1 Произведение  $\Phi\Theta$  восстанавливается устойчиво, точность восстановления не зависит от разреженности исходных модельных данных  $\Phi_0, \Theta_0$
- 2 Матрицы  $\Phi, \Theta$  восстанавливаются неустойчиво, результат зависит от случайной инициализации
- 3 Методы PLSA и LDA одинаково неустойчивы (сглаживание не спасает от неединственности)
- 4 Устойчивое восстановление матриц  $\Phi, \Theta$  происходит только при сильной разреженности (более 80% нулей)

---

Реализация экспериментов:

*Виталий Глушаченков.* Магистерская диссертация. МФТИ, 2013.

*Михаил Колупаев.* Курсовая работа. ВШЭ, 2013.

## Причина неустойчивости тематических моделей

Задача стохастического матричного разложения:

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

$\hat{F} = (n_{dw}/n_d)_{W \times D}$  — известная матрица исходных данных;

$F = (p(w|d))_{W \times D}$  — матрица тематической модели;

$\Phi = (\phi_{wt})_{W \times T}$  — искомая матрица терминов тем  $\phi_{wt} = p(w|t)$ ;

$\Theta = (\theta_{td})_{T \times D}$  — искомая матрица тем документов  $\theta_{td} = p(t|d)$ .

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

## Многокритериальная оптимизация тематической модели

Пусть, наряду с правдоподобием, требуется максимизировать ещё  $n$  критериев  $R_i(\Phi, \Theta)$ ,  $i = 1, \dots, n$  — регуляризаторов. Метод многокритериальной оптимизации — скаляризация.

Задача максимизации регуляризованного правдоподобия:

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

при ограничениях неотрицательности и нормировки

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

где  $\tau_i > 0$  — коэффициенты регуляризации.

## Обоснование регуляризованного EM-алгоритма PLSA

### Теорема

Если  $\Phi, \Theta$  — решение задачи максимизации регуляризованного правдоподобия, то оно удовлетворяет системе уравнений

$$\left\{ \begin{array}{l} \text{E-шаг: } n_{dwt} = n_{dw} \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}; \\ \text{M-шаг: } n_{wt} = \sum_{d \in D} n_{dwt}; \quad \phi_{wt} \propto \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+; \\ \quad \quad \quad n_{td} = \sum_{w \in D} n_{dwt}; \quad \theta_{td} \propto \left( n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+ \end{array} \right.$$

При  $R(\Phi, \Theta) = 0$  получаем формулы обычного EM-PLSA.

## Дивергенция Кульбака–Лейблера

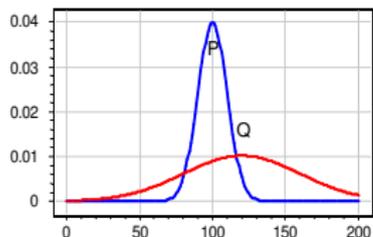
Функция расстояния между распределениями  $P = (p_i)_{i=1}^n$  и  $Q = (q_i)_{i=1}^n$ :

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

1.  $KL(P\|Q) \geq 0$ ;  $KL(P\|Q) = 0 \Leftrightarrow P = Q$ ;
2. Минимизация  $KL$  эквивалентна максимизации правдоподобия:

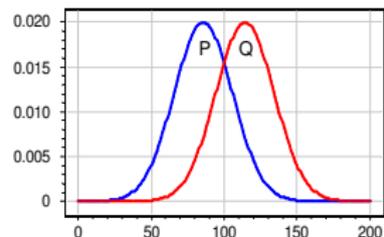
$$KL(P\|Q(\alpha)) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i(\alpha)} \rightarrow \min_{\alpha} \iff \sum_{i=1}^n p_i \ln q_i(\alpha) \rightarrow \max_{\alpha}.$$

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



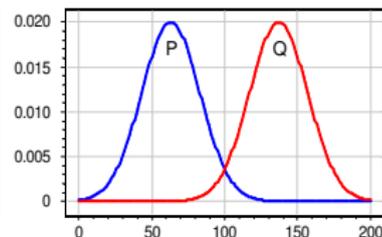
$$KL(P\|Q) = 0.442$$

$$KL(Q\|P) = 2.966$$



$$KL(P\|Q) = 0.444$$

$$KL(Q\|P) = 0.444$$



$$KL(P\|Q) = 2.969$$

$$KL(Q\|P) = 2.969$$

## Регуляризатор №1: Сглаживание (совпадает с LDA)

Гипотеза сглаженности:

распределения  $\phi_{wt}$  близки к заданным распределениям  $\beta_w$

распределения  $\theta_{td}$  близки к заданным распределениям  $\alpha_t$

$$\sum_{t \in T} \text{KL}_w(\beta_w \parallel \phi_{wt}) \rightarrow \min_{\Phi}; \quad \sum_{d \in D} \text{KL}_t(\alpha_t \parallel \theta_{td}) \rightarrow \min_{\Theta}.$$

Максимизируем сумму этих регуляризаторов:

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

Подставляем, получаем формулы M-шага LDA:

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

---

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 (совпадает с LDA)

### Выводы:

- Найдено альтернативное обоснование LDA:  
оказывается, это всего лишь притягивание столбцов  $\Phi$ ,  $\Theta$   
к заданным распределениям  $(\beta_w)$ ,  $(\alpha_t)$ ,  
причём  $\beta_0\beta_w$  и  $\alpha_0\alpha_t$  — параметры распределений Дирихле
- 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 свою «особую роль»,  
это один из многих регуляризаторов, и не самый лучший

## Регуляризатор №2: Частичное обучение (обобщение LDA)

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  $W_t \subset W$ .

$\phi_{wt}^0$  — распределение, равномерное на  $W_t$

$\theta_{td}^0$  — распределение, равномерное на  $T_d$

Максимизируем сумму этих регуляризаторов:

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

Подставляем, получаем обобщение LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0 \phi_{wt}^0 \quad \theta_{td} \propto n_{dt} + \alpha_0 \theta_{td}^0$$

---

*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 (второе обобщение LDA)

**Гипотеза:** вместо логарифма можно взять любую другую монотонно возрастающую функцию  $\mu$

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

Подставляем, получаем ещё одно обобщение LDA:

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

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

**Преимущество** ковариационного регуляризатора:

Если  $\theta_{td}^0$  равномерно на  $T_d$ , то ковариация не накладывает ограничений на распределение  $\theta_{td}$  между темами из  $T_d$ .

## Регуляризатор №3: Разреживание (третье обобщение LDA)

Гипотеза разреженности: среди  $\phi_{wt}$ ,  $\theta_{td}$  много нулей.

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

Максимизируем дивергенцию между распределениями  $\beta_w$ ,  $\alpha_t$   
(равномерными?) и искомыми распределениями  $\phi_{wt}$ ,  $\theta_{td}$ :

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

Подставляем, получаем «анти-LDA»:

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

---

*Varadarajan J., Emonet R., Odohez 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 $\phi_{wt}$ и $\theta_{td}$

### Эвристика:

постепенно увеличивать коэффициенты регуляризации  $\alpha$ ,  $\beta$

### Реализация эвристики:

начиная с итерации  $i_0$ , в конце каждой  $\delta$ -й итерации  
обнуляем долю  $r$  наименьших значений в  $\phi_t$  и  $\theta_d$ ,  
так, чтобы сумма обнуляемых значений  
не превышала  $R_\theta$  для распределений  $\theta_d$ ,  
не превышала  $R_\phi$  для распределений  $\phi_t$

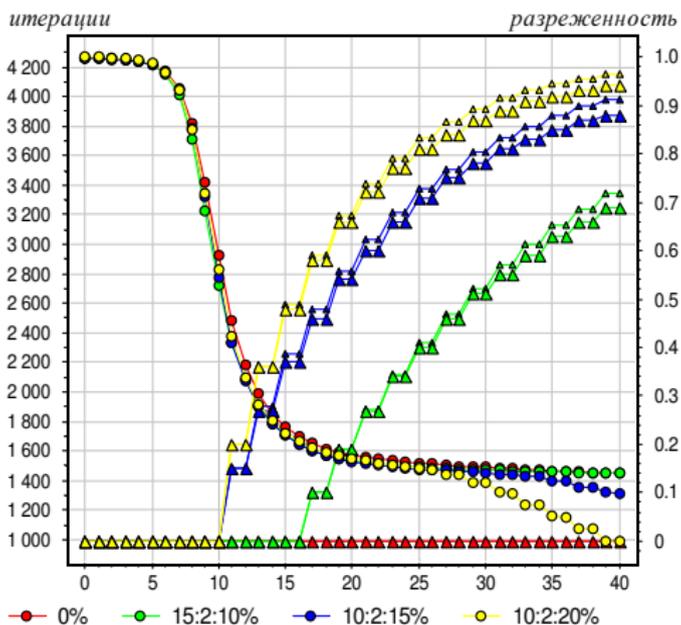
Обозначения параметров эвристики:  $i_0:\delta:r$ ,  $th:R_\theta$ ,  $ph:R_\phi$

---

*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.*

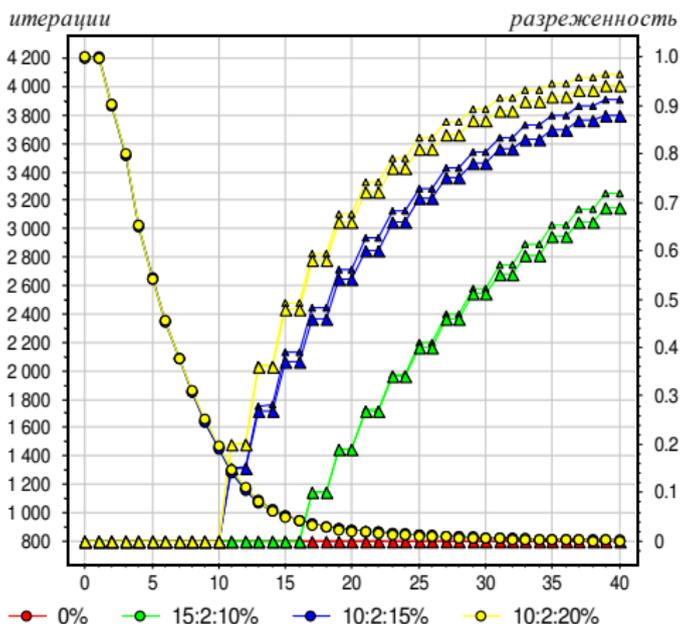
## Разреживание распределений $\phi_{wt}$ и $\theta_{td}$

упрощённая робастная модель,  
разреживание через 2 итерации



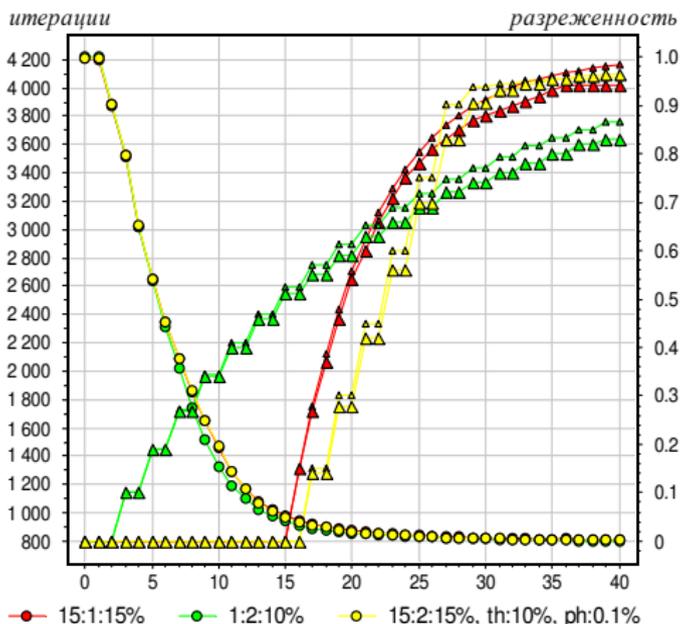
## Разреживание распределений $\phi_{wt}$ и $\theta_{td}$

робастная модель с фоном и шумом,  
разреживание через 2 итерации



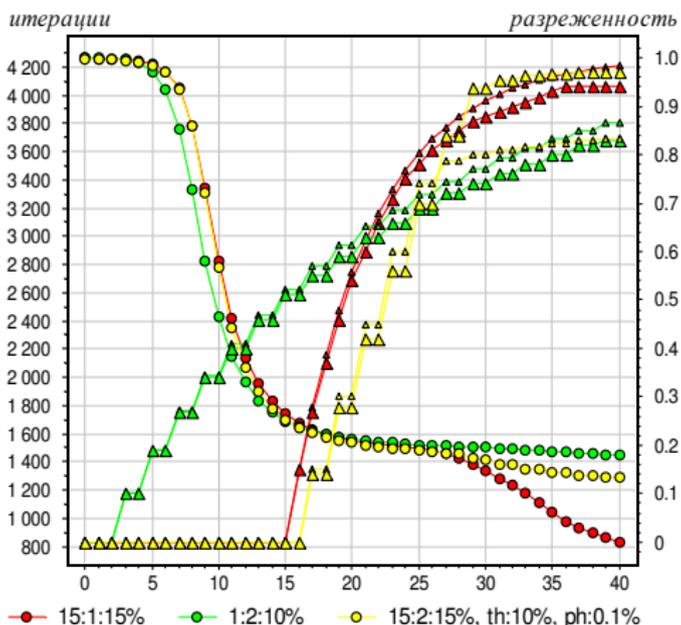
## Разреживание распределений $\phi_{wt}$ и $\theta_{td}$

робастная модель с фоном и шумом,  
агрессивные стратегии разреживания  $i_0:\delta:r$ ,  $th:R_\theta$ ,  $ph:R_\phi$



## Разреживание распределений $\phi_{wt}$ и $\theta_{td}$

упрощённая робастная модель без фона,  
агрессивные стратегии разреживания  $i_0:\delta:r$ ,  $th:R_\theta$ ,  $ph:R_\phi$



## 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  $\Phi$  обнуляются, т.е. слово оказывается нетематическим

## Регуляризатор №4: Анतिकорреляция

**Гипотеза некоррелированности тем:**

чем различнее темы, тем лучше они интерпретируются.

Минимизируем ковариации между вектор-столбцами  $\phi_t$ :

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

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

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

---

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  $\hat{p}(v|u) = N_{uv}/N_u$ .

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

Подставляем, получаем ещё один вариант сглаживания:  
векторы  $\phi_{wt}$  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: 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  $\theta_d, \theta_c$ :

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

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 №7: Классификация документов

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  $m_{dc}$ :

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

---

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

## Регуляризатор №7: Классификация (EM-алгоритм)

**E-шаг.** По формуле Байеса:

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

**M-шаг.** Максимизация регуляризованного правдоподобия:

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

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

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

## Регуляризатор №8: Категоризация документов

Снова регуляризатор для классификации:

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

Недостаток: приходится задавать равномерное распределение

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

Ковариационный регуляризатор:

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

приводит к естественному аналитическому решению

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

## Регуляризатор №9: Динамическая тематическая модель

Пусть классы  $C$  — это годы публикации

**Гипотеза:**

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

$$R_1(\Psi) = -\tau_1 \sum_{c \in C} \sum_{t \in T} |\psi_{ct} - \psi_{c-1,t}| \rightarrow \max.$$

Второй регуляризатор — разреживающий

$$R_2(\Psi) = -\tau_2 \sum_{c \in C} \sum_{t \in T} \ln \psi_{ct} \rightarrow \max.$$

## Подбор траекторий регуляризации

Пусть задана линейная комбинация регуляризаторов:

$$R(\Phi, \Theta) = \sum_{i=1}^n \tau_i R_i(\Phi, \Theta)$$

**Задача:** выбрать вектор коэффициентов  $\tau = (\tau_i)_{i=1}^n$

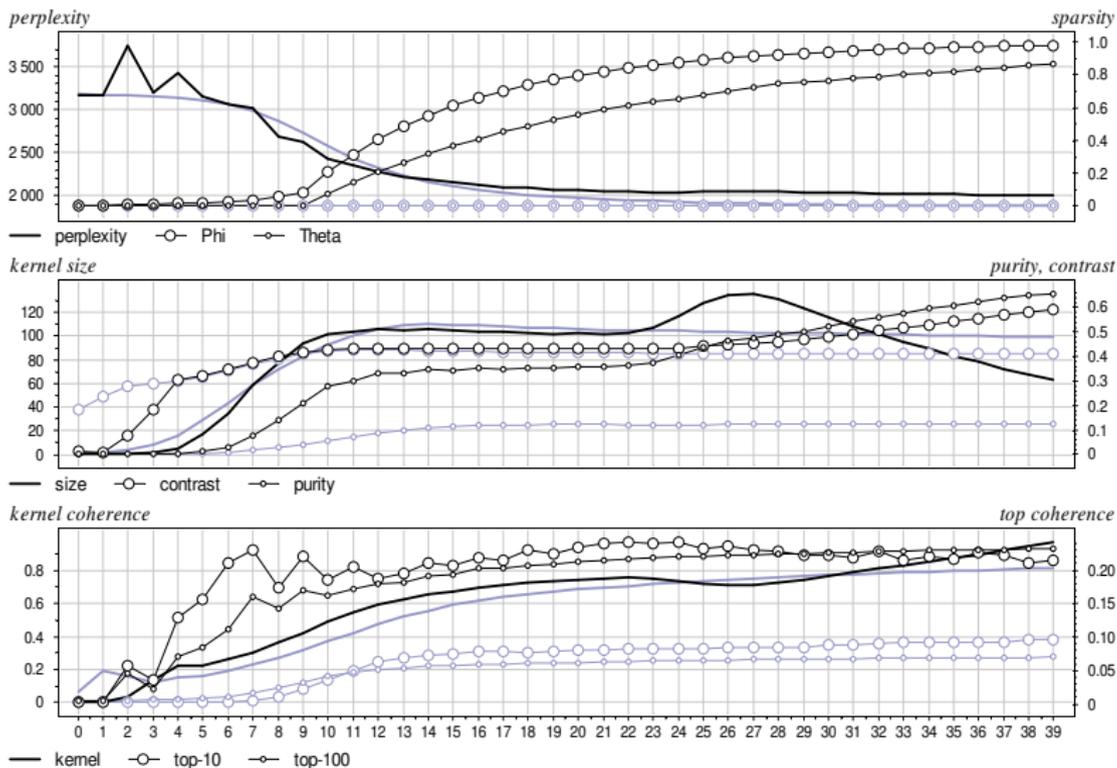
**Ближайшие аналоги:**

- Построение «Regularization Path» в задачах регрессии с двумя регуляризаторами  $L_1$  и  $L_2$  (Elastic Net)
- Постепенное разреживание тематической модели

**Идея** построения траектории в пространстве коэффициентов  $\tau$ :

- 1) достичь сходимости нерегуляризованного PLSA,
- 2) усиливать регуляризаторы постепенно, в определённом порядке.

# Комбинирование разреживания, сглаживания и декорреляции



## Оценки интерпретируемости: когерентность

Когерентность темы  $t$

$$\text{PMI}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i}^k \text{PMI}(w_i, w_j)$$

где  $w_i$  —  $i$ -й термин в порядке убывания  $\phi_{wt}$ .

$\text{PMI}(u, v) = \ln \frac{|D|N_{uv}}{N_u N_v}$  — поточечная взаимная информация (pointwise mutual information),

$N_{uv}$  — число документов, в которых термины  $u, v$  хотя бы один раз встречаются рядом (в окне 10 слов),

$N_u$  — число документов, в которых термин  $u$  встречается хотя бы один раз.

## Оценки интерпретируемости: чистота и контрастность

Термин  $w \in W$  *тематический*, если в распределении  $p(t|w) \propto \phi_{wt} n_t$  доля вероятностной массы  $\kappa = 50\%$  сконцентрирована не более чем в  $\delta = 2\%$  тем.

*Ядро* темы  $W_t$  — все её тематические термины.

Три показателя качества темы  $t$ :

- размер ядра:  $|W_t|$ ;
- чистота темы:  $\sum_{t \in W_t} p(w|t)$ ;
- контрастность темы:  $|W_t|^{-1} \sum_{t \in W_t} p(t|w)$ .

---

Vorontsov K. V., Potapenko A. A., Additive Regularization of Topic Models // Machine Learning Journal (подано).

Воронцов Константин Вячеславович  
[voron@forecsys.ru](mailto:voron@forecsys.ru)

Страницы на [www.MachineLearning.ru](http://www.MachineLearning.ru):

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

---

*Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН, №3, 2014 (в печати).*