

# Многокритериальный тематический анализ текстовых коллекций

Воронцов Константин Вячеславович

(ФИЦ ИУ РАН • МФТИ • ВШЭ • МГУ • Яндекс • FORECSYS • Aithea)

Коллоквиум ФКН НИУ ВШЭ • 5 октября 2017

- 1 Вероятностное тематическое моделирование**
  - Цели, приложения, постановка задачи
  - Аддитивная регуляризация тематических моделей
  - Классические модели: PLSA и LDA
- 2 Регуляризаторы и их комбинирование**
  - Обобщения LDA
  - Мешок регуляризаторов
  - Примеры приложений
- 3 Разведочный информационный поиск**
  - Концепция разведочного поиска
  - Оценивание качества тематического поиска
  - Оптимизация параметров модели

## Что такое «тематическое моделирование» (Topic Modeling)

- Одно из направлений обработки естественного языка
- Разновидность статистического анализа текстов
- Технология поиска информации не по словам, а по смыслу
- Выявление скрытых интересов по наблюдаемым данным
- «Мягкая кластеризация» текстовых документов
- Би-кластеризация слов и документов по кластерам-темам
- Модель машинного обучения без учителя  
(но есть и тематические модели, обучаемые с учителем)
- Модель языка, основанная на гипотезе «мешка слов»  
(но есть и модели, преодолевающие это ограничение)
- Сотни моделей, тысячи публикаций, тысячи приложений

## Приложения тематического моделирования

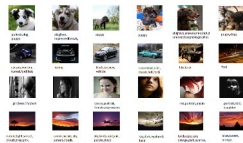
разведочный поиск в  
электронных библиотеках



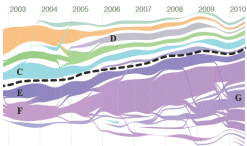
персонализированный  
поиск в соцсетях



мультимодальный поиск  
текстов и изображений



детектирование и трекинг  
новостных сюжетов



навигация по большим  
текстовым коллекциям

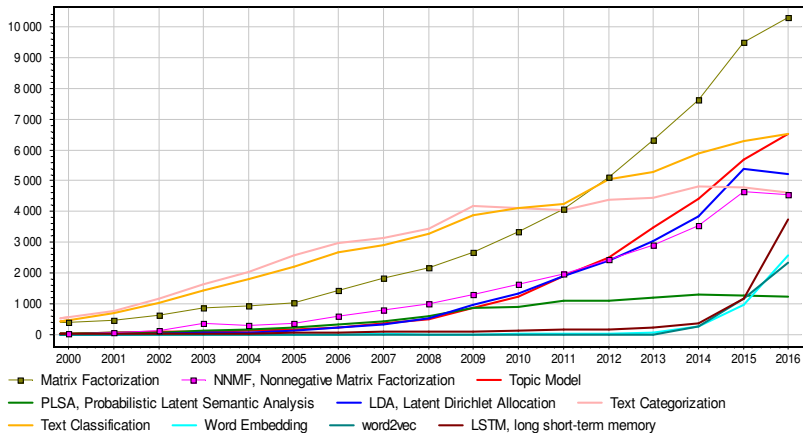


управление диалогом в  
разговорном интеллекте



## Тематическое моделирование и смежные области исследований

Динамика цитирования, по данным Google Scholar:



## Пусть

- $W$  — конечное множество слов (терминов, токенов)
- $D$  — конечное множество текстовых документов
- $T$  — конечное множество тем
- каждое слово  $w$  в документе  $d$  связано с некоторой темой  $t$
- $D \times W \times T$  — дискретное вероятностное пространство
- порядок слов в документе не важен (bag of words)
- порядок документов в коллекции не важен
- коллекция — это i.i.d. выборка  $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- $d_i, w_i$  — наблюдаемые, темы  $t_i$  — скрытые
- гипотеза условной независимости:  $p(w|d, t) = p(w|t)$

Тематическая модель, по формуле полной вероятности:

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

## Постановка задачи тематического моделирования

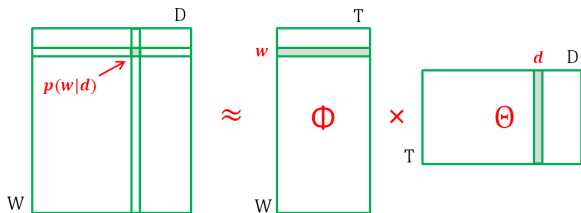
**Дано:** коллекция текстовых документов

- $n_{dw}$  — частоты терминов в документах,  $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

**Найти:** параметры тематической модели  $p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$

- $\phi_{wt} = p(w|t)$  — вероятности терминов  $w$  в каждой теме  $t$
- $\theta_{td} = p(t|d)$  — вероятности тем  $t$  в каждом документе  $d$

Это задача стохастического матричного разложения:



## Принцип максимума правдоподобия

Правдоподобие — плотность распределения выборки  $(d_i, w_i)_{i=1}^n$ :

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

Максимизация логарифма правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d)p(d) \rightarrow \max_{\Phi, \Theta}$$

эквивалентна максимизации функционала

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

при ограничениях неотрицательности и нормировки

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



## Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена*,  
если её решение

- существует,
- единственно,
- устойчиво.



Жак Саломон Адамар  
(1865–1963)

Наша задача матричного разложения *некорректно поставлена*:  
если  $\Phi, \Theta$  — решение, то стохастические  $\Phi', \Theta'$  — тоже решения

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta)$ ,  $\text{rank} S = |T|$
- $\mathcal{L}(\Phi', \Theta') = \mathcal{L}(\Phi, \Theta)$
- $\mathcal{L}(\Phi', \Theta') \leq \mathcal{L}(\Phi, \Theta) + \varepsilon$  — приближённые решения

**Регуляризация** — стандартный приём доопределения решения  
с помощью дополнительных критериев.

## ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}; \quad R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta)$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} \equiv p(t|d, w) = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \operatorname{norm}_{w \in W} \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), & n_{wt} = \sum_{d \in D} n_{dw} p_{tdw} \\ \theta_{td} = \operatorname{norm}_{t \in T} \left( n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), & n_{td} = \sum_{w \in W} n_{dw} p_{tdw} \end{cases} \end{cases}$$

где  $\operatorname{norm}_{t \in T}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$  — операция нормировки вектора.

## Элементарная интерпретация EM-алгоритма

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

**E-шаг:** условные вероятности тем  $p(t|d, w)$  для всех  $t, d, w$  вычисляются через  $\phi_{wt}, \theta_{td}$  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:** при  $R = 0$  частотные оценки условных вероятностей вычисляются суммированием счётчика  $n_{tdw} = n_{dw}p(t|d, w)$ :

$$\begin{aligned} \phi_{wt} &= \frac{n_{wt}}{n_t}, & n_{wt} &= \sum_{d \in D} n_{tdw}, & n_t &= \sum_{w \in W} n_{wt}; \\ \theta_{td} &= \frac{n_{td}}{n_d}, & n_{td} &= \sum_{w \in D} n_{tdw}, & n_d &= \sum_{t \in T} n_{td}. \end{aligned}$$

## Условия вырожденности модели для тем и документов

Решение может быть вырожденным для некоторых тем (столбцов матриц  $\Phi$ ) и документов (столбцов матрицы  $\Theta$ ).

*Тема  $t$  вырождена, если для всех терминов  $w \in W$*

$$n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \leq 0.$$

Если тема  $t$  вырождена, то  $p(w|t) = \phi_{wt} \equiv 0$ ; это означает, что тема исключается из модели (происходит отбор тем).

*Документ  $d$  вырожден, если для всех тем  $t \in T$*

$$n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \leq 0.$$

Если документ  $d$  вырожден, то  $p(t|d) = \theta_{td} \equiv 0$ ; это означает, что модель не в состоянии описать данный документ.

## Напоминания. Условия Каруша–Куна–Таккера

Задача математического программирования:

$$\begin{cases} f(x) \rightarrow \min_x; \\ g_i(x) \leq 0, & i = 1, \dots, m; \\ h_j(x) = 0, & j = 1, \dots, k. \end{cases}$$

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

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial x} = 0, & \mathcal{L}(x; \mu, \lambda) = f(x) + \sum_{i=1}^m \mu_i g_i(x) + \sum_{j=1}^k \lambda_j h_j(x); \\ g_i(x) \leq 0; h_j(x) = 0; & \text{(исходные ограничения)} \\ \mu_i \geq 0; & \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; & \text{(условие дополняющей нежёсткости)} \end{cases}$$

## Вывод системы уравнений из условий Каруша–Куна–Таккера

1. Условия ККТ для  $\phi_{wt}$  (для  $\theta_{td}$  всё аналогично):

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

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. Если  $\lambda_t \leq 0$ , то тема  $t$  вырождена,  $\phi_{wt} \equiv 0$  для всех  $w$ .

4. Если  $\lambda_t > 0$ , то либо  $\phi_{wt} = 0$ , либо  $n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} > 0$ :

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

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

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

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

## Рациональный EM-алгоритм

**Идея:** E-шаг встраивается внутрь M-шага, чтобы не хранить трёхмерный массив значений  $n_{dwt}$ .

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

$n_{wt}, n_{td} := 0$  для всех  $d \in D, w \in W, t \in T$ ;

**для всех** документов  $d \in D$  и всех слов  $w \in d$

$n_{tdw} := n_{dw} \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td})$  для всех  $t \in T$ ;

$n_{wt} += n_{tdw}; n_{td} += n_{tdw}$  для всех  $t \in T$ ;

$\phi_{wt} := \operatorname{norm}_{w \in W} \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$  для всех  $w \in W, t \in T$ ;

$\theta_{td} := \operatorname{norm}_{t \in T} \left( n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$  для всех  $d \in D, t \in T$ ;

## Онлайновый EM-алгоритм (реализован в BigARTM)

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

**Выход:** матрицы терминов тем  $\Theta$  и тем документов  $\Phi$ ;

инициализировать  $n_{wt} := 0$  и  $\phi_{wt}$ ;

**для всех**  $i = 1, \dots, i_{\max}$  (для больших коллекций  $i_{\max} = 1$ )

**для всех** документов  $d \in D$

инициализировать  $\theta_{td} := \frac{1}{|T|}$ ;

**для всех**  $j = 1, \dots, j_{\max}$  (итерации по документу)

$n_{tdw} := n_{dw} \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td})$  для всех  $w \in d$ ;

$\theta_{td} := \operatorname{norm}_{t \in T} \left( \sum_w n_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$ ;

$n_{wt} := \gamma n_{wt} + n_{tdw}$ ;

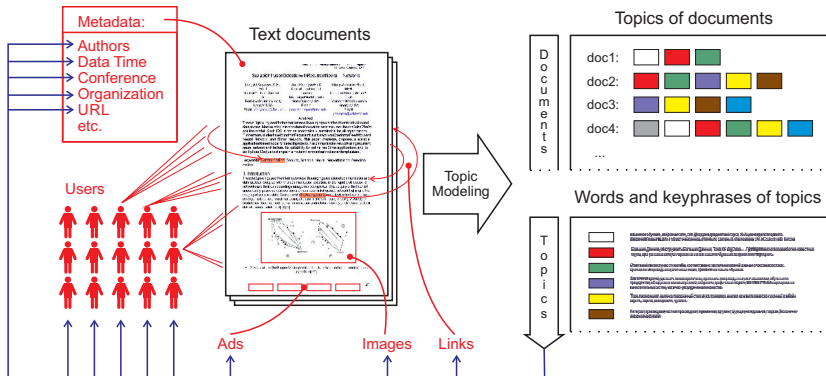
**если** пора обновить матрицу  $\Phi$  **то**

$\phi_{wt} := \operatorname{norm}_{w \in W} \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ ;



## Задачи мультимодального тематического моделирования

Темы определяют распределения не только терминов  $p(w|t)$ , но и других *модальностей*:  $p(\text{автор}|t)$ ,  $p(\text{время}|t)$ ,  $p(\text{ссылка}|t)$ ,  $p(\text{баннер}|t)$ ,  $p(\text{элемент\_изображения}|t)$ ,  $p(\text{пользователь}|t)$ , ...



## Мультимодальная ARTM

$W^m$  — словарь токенов  $m$ -й модальности,  $m \in M$

Максимизация суммы  $\log$  правдоподобий с регуляризацией:

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \operatorname{norm}_{w \in W^m} \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), & n_{wt} = \sum_{d \in D} \tau_{m(w)} n_{dw} p_{tdw} \\ \theta_{td} = \operatorname{norm}_{t \in T} \left( n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), & n_{td} = \sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw} \end{cases} \end{cases}$$

## BigARTM: библиотека тематического моделирования

### Ключевые возможности:

- Онлайн-овый параллельный мультимодальный ARTM
- Большие данные: коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов и мер качества

### Сообщество:

- Открытый код <https://github.com/bigartm>  
(discussion group, issue tracker, pull requests)
- Документация <http://bigartm.org>



### Лицензия и среда разработки:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

## BigARTM упрощает разработку тематических моделей


Для построения сложных моделей в BigARTM не нужны ни математические выкладки, ни программирование «с нуля».


### Этапы моделирования

#### Bayesian TM

#### ARTM

	Bayesian TM	ARTM
	Анализ требований	Анализ требований
Формализация:	Вероятностная порождающая модель данных	Стандартные критерии   Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Общий регуляризованный EM-алгоритм для любых моделей
Реализация:	Исследовательский код (Matlab, Python, R)	Промышленный код BigARTM (C++, Python API)
Оценивание:	Исследовательские метрики, исследовательский код	Стандартные метрики   Свои метрики
	Внедрение	Внедрение

 -- нестандартизируемые этапы, уникальная разработка для каждой задачи

 -- стандартизуемые этапы

## Классические модели PLSA и LDA

**PLSA**: probabilistic latent semantic analysis [Hofmann, 1999]  
(вероятностный латентный семантический анализ):

$$R(\Phi, \Theta) = 0.$$

M-шаг — частотные оценки условных вероятностей:

$$\phi_{wt} = \underset{w}{\text{norm}}(n_{wt}), \quad \theta_{td} = \underset{t}{\text{norm}}(n_{td}).$$

**LDA**: latent Dirichlet allocation (латентное размещение Дирихле):

$$R(\Phi, \Theta) = \sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td}.$$

M-шаг — сглаженные частотные оценки с параметрами  $\beta_w, \alpha_t$ :

$$\phi_{wt} = \underset{w}{\text{norm}}(n_{wt} + \beta_w - 1), \quad \theta_{td} = \underset{t}{\text{norm}}(n_{td} + \alpha_t - 1).$$

---

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

Blei D., Ng A., Jordan M. Latent Dirichlet allocation. 2003.

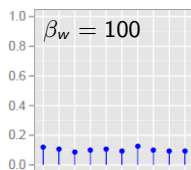
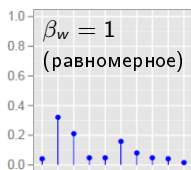
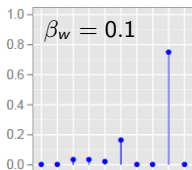
## Вероятностная байесовская интерпретация LDA [Blei, 2003]

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

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

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

**Пример.** Распределение  $\phi \sim \text{Dir}(\beta)$  при  $|W| = 10$ ,  $\phi, \beta \in \mathbb{R}^{10}$ :



## Максимизация апостериорной вероятности для модели LDA

Совместное правдоподобие данных и модели:

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

Регуляризатор — логарифм априорного распределения:

$$R(\Phi, \Theta) = \sum_{t, w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d, t} (\alpha_t - 1) \ln \theta_{td}$$

M-шаг — сглаженные или слабо разреженные оценки:

$$\phi_{wt} = \text{norm}_w(n_{wt} + \beta_w - 1), \quad \theta_{td} = \text{norm}_t(n_{td} + \alpha_t - 1).$$

при  $\beta_w > 1$ ,  $\alpha_t > 1$  — сглаживание,

при  $0 < \beta_w < 1$ ,  $0 < \alpha_t < 1$  — слабое разреживание,

при  $\beta_w = 1$ ,  $\alpha_t = 1$  априорное распределение равномерно, PLSA.

## Почему именно распределение Дирихле?

### Плюсы:

- удобно для байесовского вывода, т. к. является сопряжённым к мультиномиальному распределению
- описывает широкий класс распределений на симплексе
- позволяет управлять разреженностью  $\phi_{wt}$  и  $\theta_{td}$
- при малых  $n_{wt}$ ,  $n_{td}$  уменьшает переобучение

### Минусы:

- не имеет лингвистических обоснований
- не даёт выигрыша против PLSA на больших коллекциях
- слабый разреживатель: запрещены  $\beta_w \leq 0$ ,  $\alpha_t \leq 0$
- слабый регуляризатор: проблема неединственности остаётся

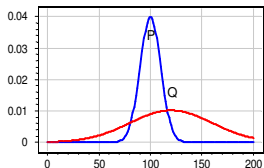


## Напоминание. Дивергенция Кульбака–Лейблера

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

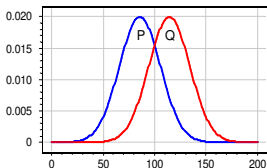
$$KL(P\|Q(\alpha)) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i(\alpha)} \rightarrow \min_{\alpha} \Leftrightarrow \sum_{i=1}^n p_i \ln q_i(\alpha) \rightarrow \max_{\alpha}$$

- Если  $KL(P\|Q) < KL(Q\|P)$ , то  $P$  вложено в  $Q$ :



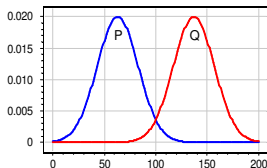
$$KL(P\|Q) = 0.44$$

$$KL(Q\|P) = 2.97$$



$$KL(P\|Q) = 0.44$$

$$KL(Q\|P) = 0.44$$



$$KL(P\|Q) = 2.97$$

$$KL(Q\|P) = 2.97$$

## Обобщённая не-байесовская интерпретация LDA

Сглаживание распределений по KL-дивергенции:

приблизить  $\phi_{wt} \equiv p(w|t)$  к заданным распределениям  $\beta_t(w)$ ,  
приблизить  $\theta_{td} \equiv p(t|d)$  к заданным распределениям  $\alpha_d(t)$ :

$$\sum_{t \in T} \tau_t \text{KL}(\beta_t(w) \parallel \phi_{wt}) \rightarrow \min_{\Phi}; \quad \sum_{d \in D} \tau_d \text{KL}(\alpha_d(t) \parallel \theta_{td}) \rightarrow \min_{\Theta}.$$

Взвешенная сумма регуляризаторов:

$$R(\Phi, \Theta) = \sum_{t \in T} \tau_t \sum_{w \in W} \beta_t(w) \ln \phi_{wt} + \sum_{d \in D} \tau_d \sum_{t \in T} \alpha_d(t) \ln \theta_{td}.$$

Формулы M-шага:

$$\phi_{wt} = \underset{w}{\text{norm}} \left( n_{wt} + \underbrace{\tau_t \beta_t(w)}_{\beta_{wt}} \right), \quad \theta_{td} = \underset{t}{\text{norm}} \left( n_{td} + \underbrace{\tau_d \alpha_d(t)}_{\alpha_{td}} \right).$$

## Сглаживание, разреживание и частичное обучение тем

Формулы M-шага:

$$\phi_{wt} = \operatorname{norm}_w(n_{wt} + \beta_{wt}), \quad \theta_{td} = \operatorname{norm}_t(n_{td} + \alpha_{td}).$$

Разреживание и сглаживание описывается общей формулой:

- разреживание — максимизация KL,  $\beta_{wt} < 0$ ,  $\alpha_{td} < 0$
- сглаживание — минимизация KL,  $\beta_{wt} > 0$ ,  $\alpha_{td} > 0$

*Частичное обучение* темы  $t$ :

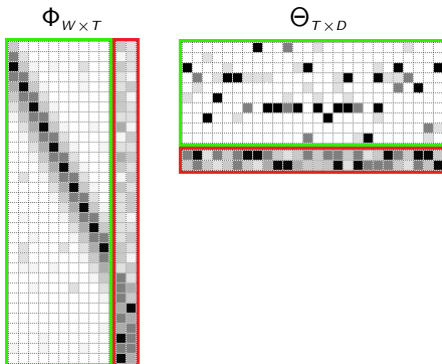
- $\beta_{wt} = +\tau_{6T}[w \in W_t]$  — «белый список» терминов
- $\beta_{wt} = -\tau_{чT}[w \in W_t]$  — «чёрный список» терминов
- $\alpha_{td} = +\tau_{6D}[d \in D_t]$  — «белый список» документов
- $\alpha_{td} = -\tau_{чD}[d \in D_t]$  — «чёрный список» документов

## Разделение тем на предметные и фоновые

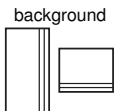
$T = S \sqcup B$  — множество всех тем

$S$  — разреженные *предметные* темы, специальная лексика

$B$  — сглаженные *фоновые* темы, общая лексика языка

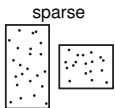


## Регуляризаторы для улучшения интерпретируемости тем



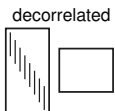
Сглаживание фоновых тем  $B \subset T$ :

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_w \beta_w \ln \phi_{wt} + \alpha_0 \sum_d \sum_{t \in B} \alpha_t \ln \theta_{td}$$



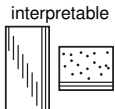
Разреживание предметных тем  $S = T \setminus B$ :

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_w \beta_w \ln \phi_{wt} - \alpha_0 \sum_d \sum_{t \in S} \alpha_t \ln \theta_{td}$$



Декоррелирование для повышения различности тем:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_w \phi_{wt} \phi_{ws}$$



Сглаживание + разреживание + декоррелирование  
 для улучшения интерпретируемости тем

## Иерархические, темпоральные, регрессионные модели

hierarchy



Связь родительских тем  $t$  с дочерними подтемами  $s$ :

$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \phi_{ws} \psi_{st}.$$

temporal



Темпоральные модели с модальностью времени  $i$ :

$$R(\Phi) = -\tau \sum_{i \in I} \sum_{t \in T} |\phi_{it} - \phi_{i-1,t}|.$$

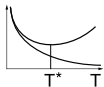
regression



Линейная модель регрессии  $\hat{y}_d = \langle v, \theta_d \rangle$  документов:

$$R(\Theta, v) = -\tau \sum_{d \in D} \left( y_d - \sum_{t \in T} v_t \theta_{td} \right)^2.$$

n of topics



Разреживание  $p(t)$  для отбора тем:

$$R(\Theta) = -\tau \sum_{t \in T} \frac{1}{|T|} \ln p(t), \quad p(t) = \sum_{d \in D} p(d) \theta_{td}.$$

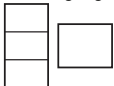
## Специальные случаи мультимодальных тематических моделей

supervised



Модальности меток классов или категорий для задач классификации и категоризации текстов.

multilanguage



Модальность языков и регуляризация со словарём  $\pi_{uwt} = p(u|w, t)$  переводов с языка  $k$  на  $\ell$ :

$$R(\Phi, \Pi) = \tau \sum_{u \in W^k} \sum_{t \in T} n_{ut} \ln \sum_{w \in W^\ell} \pi_{uwt} \phi_{wt}$$

graph



Модальность вершин графа  $v$ , содержащих  $D_v$ :

$$R(\Phi) = -\frac{\tau}{2} \sum_{(u,v) \in E} S_{uv} \sum_{t \in T} n_t^2 \left( \frac{\phi_{vt}}{|D_v|} - \frac{\phi_{ut}}{|D_u|} \right)^2.$$

geospatial



Модальность геолокаций  $g$  с близостью  $S_{gg'}$ :

$$R(\Phi) = -\frac{\tau}{2} \sum_{g, g' \in G} S_{gg'} \sum_{t \in T} n_t^2 \left( \frac{\phi_{gt}}{n_g} - \frac{\phi_{g't}}{n_{g'}} \right)^2$$

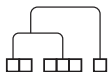
## В обход гипотезы «мешка слов» (beyond bag-of-words)

n-gram



Модальности  $n$ -грамм, коллокаций, именованных сущностей

syntax



Модальность  $n$ -грамм после применения SyntaxNet

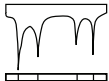
coherence



Совстречаемость слов  $n_{uv}$  в битермах  $(u, v)$ :

$$R(\Phi) = \tau \sum_{u,v} n_{uv} \ln \sum_t n_t \phi_{ut} \phi_{vt}$$

segmentation



Регуляризация E-шага, постобработка распределений  $p(t|d, w)$  для тематической сегментации



## Регуляризация E-шага

Максимизация log-правдоподобия с регуляризаторами  $R$  и  $\tilde{R}$ :

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

где  $\Pi = (p_{tdw})_{T \times D \times W}$  — матрица распределений  $p_{tdw} = p(t|d, w)$ .

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & \left\{ \begin{array}{l} p_{tdw} = \mathop{\text{norm}}_{t \in T}(\phi_{wt} \theta_{td}) \\ \tilde{p}_{tdw} = p_{tdw} \left( 1 + \frac{1}{n_{dw}} \left( \frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z \in T} p_{zdw} \frac{\partial R(\Pi)}{\partial p_{zdw}} \right) \right) \end{array} \right. \\ \text{M-шаг:} & \left\{ \begin{array}{l} \phi_{wt} = \mathop{\text{norm}}_{w \in W} \left( \sum_{d \in D} n_{dw} \tilde{p}_{tdw} + \phi_{wt} \frac{\partial \tilde{R}}{\partial \phi_{wt}} \right) \\ \theta_{td} = \mathop{\text{norm}}_{t \in T} \left( \sum_{w \in d} n_{dw} \tilde{p}_{tdw} + \theta_{td} \frac{\partial \tilde{R}}{\partial \theta_{td}} \right) \end{array} \right. \end{cases}$$

## Доказательство

**Лемма 1.** Для функции  $p_{tdw}(\Phi, \Theta) = \frac{\phi_{wt}\theta_{td}}{\sum_z \phi_{wz}\theta_{zd}}$  и любого  $z \in T$

$$\phi_{wt} \frac{\partial p_{zdw}}{\partial \phi_{wt}} = \theta_{td} \frac{\partial p_{zdw}}{\partial \theta_{td}} = p_{tdw} ([z=t] - p_{zdw}).$$

Введём функцию от вспомогательных переменных  $\Pi$ :

$$Q_{tdw}(\Pi) = \frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z \in T} p_{zdw} \frac{\partial R(\Pi)}{\partial p_{zdw}}.$$

**Лемма 2.** Если  $R(\Pi)$  не зависит от  $p_{tdw}$  при  $w \notin d$ , то

$$\phi_{wt} \frac{\partial R(\Pi)}{\partial \phi_{wt}} = \sum_{d \in D} p_{tdw} Q_{tdw}(\Pi); \quad \theta_{td} \frac{\partial R(\Pi)}{\partial \theta_{td}} = \sum_{w \in d} p_{tdw} Q_{tdw}(\Pi).$$

**Лемма 3.** Формулы M-шага:

$$\phi_{wt} = \operatorname{norm}_{w \in W} \left( \sum_{d \in D} n_{dw} p_{tdw} + \sum_{d \in D} Q_{tdw} p_{tdw} + \phi_{wt} \frac{\partial \tilde{R}}{\partial \phi_{wt}} \right);$$

$$\theta_{td} = \operatorname{norm}_{t \in T} \left( \sum_{w \in d} n_{dw} p_{tdw} + \sum_{w \in d} Q_{tdw} p_{tdw} + \theta_{td} \frac{\partial \tilde{R}}{\partial \theta_{td}} \right).$$

## Поиск и классификация этнического дискурса в соцсетях

**Задача:** найти все этно-релевантные темы для мониторинга межнациональных отношений.

Используем словарь из 300 этнонимов для обучения тем.

Мешок регуляризаторов:

$$\mathcal{L} \left( \begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array} \right) + R \left( \begin{array}{c} \text{interpretable} \\ \text{bar chart} \quad \text{scatter plot} \end{array} \right) + R \left( \begin{array}{c} \text{multimodal} \\ \text{stacked bar} \quad \text{table} \end{array} \right) \\
 + R \left( \begin{array}{c} \text{temporal} \\ \text{line graph} \end{array} \right) + R \left( \begin{array}{c} \text{geospatial} \\ \text{map} \end{array} \right) + R \left( \begin{array}{c} \text{sentiment} \\ \text{sentiment scale} \end{array} \right) \rightarrow \max$$

**Результаты:** число релевантных тем выросло с 45 для LDA до 83 для ARTM.

*M. Apishev, S. Koltcov, O. Koltsova, S. Nikolenko, K. Vorontsov. Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI, 2016.*

## Разведочный поиск в коллективных блогах

**Задача:** поиск документов по длинному запросу.

Мешок регуляризаторов:

$$\mathcal{L} \left( \begin{array}{c} \text{PLSA} \\ \left( \begin{array}{|c|} \hline \Phi \\ \hline \end{array} \begin{array}{|c|} \hline \Theta \\ \hline \end{array} \right) \end{array} \right) + R \left( \begin{array}{c} \text{interpretable} \\ \left( \begin{array}{|c|} \hline \text{bar chart} \\ \hline \end{array} \begin{array}{|c|} \hline \text{matrix} \\ \hline \end{array} \right) \end{array} \right) + R \left( \begin{array}{c} \text{multimodal} \\ \left( \begin{array}{|c|} \hline \text{stacked boxes} \\ \hline \end{array} \begin{array}{|c|} \hline \text{empty box} \\ \hline \end{array} \right) \end{array} \right) + R \left( \begin{array}{c} \text{n-gram} \\ \left( \begin{array}{|c|} \hline \text{grid of boxes} \\ \hline \end{array} \right) \end{array} \right) \rightarrow \max$$

**Результаты:**

- Точность и полнота увеличились с (65%, 73%) для LDA до (85%, 92%) для ARTM на данных Habrahabr.ru и TechCrunch.com.
- Точность и полнота сравнимы с результатами ассессоров.
- Тематический поиск даёт результат мгновенно, ассессоры тратят на эту же работу в среднем 30 минут.

---

A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

## Иерархическая темпоральная модель новостного потока

### Задачи:

- наращивать 3х-уровневую иерархию динамически
- обеспечить интерпретируемость и именованность всех тем
- управлять медиакомпаниями и творческими заданиями

### Мешок регуляризаторов:

$$\begin{aligned}
 & \mathcal{L} \left( \begin{array}{|c|c|} \hline \text{PLSA} & \\ \hline \Phi & \Theta \\ \hline \end{array} \right) + R \left( \begin{array}{|c|c|} \hline \text{interpretable} & \\ \hline \text{[Bar Chart]} & \text{[Scatter Plot]} \\ \hline \end{array} \right) + R \left( \begin{array}{|c|} \hline \text{hierarchy} \\ \hline \text{[Tree Diagram]} \\ \hline \end{array} \right) + R \left( \begin{array}{|c|} \hline \text{temporal} \\ \hline \text{[Line Graph]} \\ \hline \end{array} \right) \\
 & + R \left( \begin{array}{|c|c|} \hline \text{multimodal} & \\ \hline \text{[Grid]} & \text{[Box]} \\ \hline \end{array} \right) + R \left( \begin{array}{|c|} \hline \text{n-gram} \\ \hline \text{[Grid]} \\ \hline \end{array} \right) + R \left( \begin{array}{|c|c|} \hline \text{multilanguage} & \\ \hline \text{[Grid]} & \text{[Box]} \\ \hline \end{array} \right) + R \left( \begin{array}{|c|} \hline \text{sentiment} \\ \hline \text{[Sentiment Diagram]} \\ \hline \end{array} \right) \rightarrow \max
 \end{aligned}$$

Результат: ... (исследование продолжается)

## Сценарный анализ записей разговоров контакт-центра

### Задачи:

- выделить сценарии диалогов оператор–клиент
- автоматизировать оценивание качества работы операторов
- выработать онлайн-подсказки для оператора

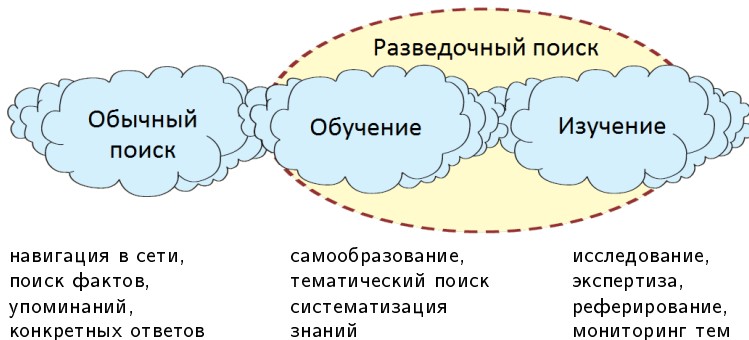
### Мешок регуляризаторов:

$$\begin{aligned}
 \mathcal{L} & \left( \begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array} \right) + R \left( \begin{array}{c} \text{interpretable} \\ \text{[bar chart]} \quad \text{[scatter plot]} \end{array} \right) + R \left( \begin{array}{c} \text{segmentation} \\ \text{[waveform]} \end{array} \right) + R \left( \begin{array}{c} \text{n-gram} \\ \text{[grid]} \end{array} \right) \\
 & + R \left( \begin{array}{c} \text{syntax} \\ \text{[tree diagram]} \end{array} \right) + R \left( \begin{array}{c} \text{sentence} \\ \text{[horizontal bars]} \end{array} \right) + R \left( \begin{array}{c} \text{dialog} \\ \text{[stacked bars]} \end{array} \right) \rightarrow \max
 \end{aligned}$$

**Результат:** качество сегментации выросло с 40% у базового решения до 75% у ARTM

## Концепция разведочного поиска (exploratory search)

- пользователь может не знать ключевых терминов,
- запросом может быть текст произвольной длины,
- информационной потребностью — систематизация знаний



Gary Marchionini. Exploratory Search: from finding to understanding. 2006.

## Разведочный тематический поиск

$q = (w_1, \dots, w_{n_q})$  — текст запроса произвольной длины  $n_q$

$\theta_{tq} = p(t|q)$  — тематический профиль запроса  $q$

$\theta_{td} = p(t|d)$  — тематические профили документов  $d \in D$

Косинусная мера близости документа  $d$  и запроса  $q$ :

$$\text{sim}(q, d) = \frac{\sum_t \theta_{tq} \theta_{td}}{(\sum_t \theta_{tq}^2)^{1/2} (\sum_t \theta_{td}^2)^{1/2}}.$$

Ранжируем документы коллекции  $d \in D$  по убыванию  $\text{sim}(q, d)$

Выдача тематического поиска —  $k$  первых документов.

Реализация: *инвертированный индекс* для быстрого поиска документов  $d$  по каждой из тем  $t$  запроса



## Данные коллективного блога Habrahabr.ru

### Данные

- 175 143 статей
- Модальности:
  - 10552 слов (униграмм), 742 000 биграмм
  - 524 авторов статей
  - 10 000 комментаторов (авторов комментариев к статьям)
  - 2546 тегов
  - 123 хаба (категории)

### Предобработка текстов

- отброшены 5% наиболее частотных слов (общая лексика)
- удаление пунктуации
- нижний регистр, ё→е
- лемматизация rymorphy2

# Методика оценивания качества разведочного поиска

## Поисковый запрос

набор ключевых слов или фрагментов текста, около одной страницы A4

## Поисковая выдача

документы  $d$  с распределением  $p(t|d)$ , близким к распределению  $p(t|q)$  запроса

## Два задания ассессорам

- 1 найти как можно больше статей, пользуясь любыми средствами поиска (и засечь время)
- 2 оценить релевантность поисковой выдачи на том же запросе

### Поиск MapReduce

**Поиск MapReduce** – программа поиска (библиотека) написанная распределенно вычислений для больших объемов данных в рамках параллельных вычислений, представляющая собой набор Java-классов и исполняемых утилит для создания и обработки данных на параллельную обработку.

**Основные возможности Поиск MapReduce** можно сформулировать как:

- обработка вычислений больших объемов данных;
- масштабируемость;
- автоматическое распределение заданий;
- работа на невидимых обрабатывающих;
- автоматическая обработка отказов вычислений заданий.

**Поиск** – популярная программная платформа (библиотека, библиотека) построена распределенных приложений для массово-параллельной обработки (разделов, разделов, регионов, МР) данных.

**Поиск** включает в себе следующие компоненты:

1. HDFS – распределенная файловая система;

2. **Поиск MapReduce** – программная модель (библиотека) написанная распределенно вычислений для больших объемов данных в рамках параллельных вычислений;

**Ключевые, объектные** и архитектура **Поиск MapReduce** и структура HDFS, стали примером того, как можно сделать в рамках вычислений, в том числе и с помощью точки зрения. Это, в конечном итоге, определило направление платформ **Поиск** в целом. К последним можно отнести:

Ограничение масштабируемости кластера **Поиск** – это вычислительный узел, – это параллельные задания.

Сильная связность **Поиск** распределенно вычислений и клиентских вычислений, реализованных распределенно алгоритмом. Как следствие:

Отсутствие поддержки альтернативной программной модели вычислений распределенно вычислений в **Поиск v1.0** поддерживается только модель вычислений **MapReduce**.

Модель вычислений, точки зрения и как следствие, необходимость использования в среде с высоким требованиями к надежности;

Проблема **вычислений** совместности: требование по единственному объектно-ориентированному использованию всех вычислительных узлов кластера при обращении платформ **Поиск** (установка новой версии или пакета обновлений).

Пример запроса для разведочного поиска

## Пример: фрагмент запроса «Система IBM Watson»

IBM Watson — суперкомпьютер фирмы IBM, оснащённый вопросно-ответной системой искусственного интеллекта, созданный группой исследователей под руководством Дэвида Феруччи. Его создание — часть проекта DeepQA. Основная задача Уотсона — понимать вопросы, сформулированные на естественном языке, и находить на них ответы в базе данных. Назван в честь основателя IBM Томаса Уотсона.

IBM Watson представляет собой когнитивную систему, которая способна понимать, делать выводы и обучаться. Она также позволяет преобразовывать целые отрасли, различные направления науки и техники. Например, предсказывать появление эпидемий или возникновения очагов природных катастроф в различных регионах, вести мониторинг состояния атмосферы больших городов, оптимизировать бизнес-процессы, узнавать, какие товары будут в тренде в ближайшее время.

... ..

**Релевантные тексты:** примеры сервисов и приложений, основа которых — когнитивная платформа IBM Watson, используемые в IBM Watson технологии, вопрос-ответные системы, сопоставление IBM Watson с Wolfram-Alpha.

**Нерелевантные тексты:** общие вопросы искусственного интеллекта, другие коммерческие решения на рынке бизнес-аналитики.

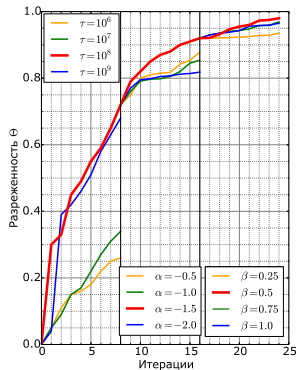
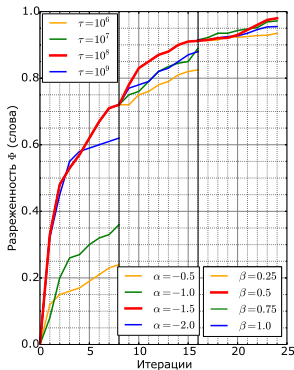
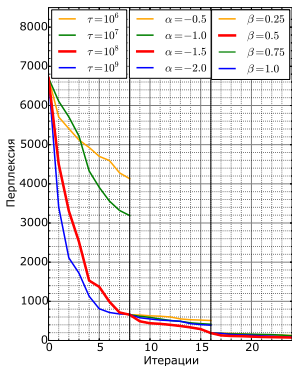
## Тематика запросов разведочного поиска

Примеры заголовков разведочных запросов к Хабру  
(объём каждого запроса — около одной страницы А4):

Алгоритмы раскраски графов	Система IBM Watson
Рекомендательная система Netflix	3D-принтеры
Методики быстрого набора текста	CERN-кластер
Космические проекты Илона Маска	АВ-тестирование
Технологии Hadoop MapReduce	Облачные сервисы
Беспилотный автомобиль Google car	Контекстная реклама
Криптосистемы с открытым ключом	Марсоход Curiosity
Обзор платформ онлайн-курсов	Видеокарты NVIDIA
Data Science Meetups в Москве	Распознавание образов
Образовательные проекты mail.ru	Сервисы Google scholar
Межпланетная станция New horizons	MIT MediaLab Research
Языковая модель word2vec	Платформа Microsoft Azure

## Последовательный подбор коэффициентов регуляризации

- декоррелирование распределений терминов в темах ( $\tau$ ),
- разреживание распределений тем в документах ( $\alpha$ ),
- сглаживание распределений терминов в темах ( $\beta$ ).



## Оценки качества поиска

Precision — доля релевантных среди найденных

Recall — доля найденных среди релевантных

$$P = \frac{TP}{TP + FP} \text{ — точность (precision)}$$

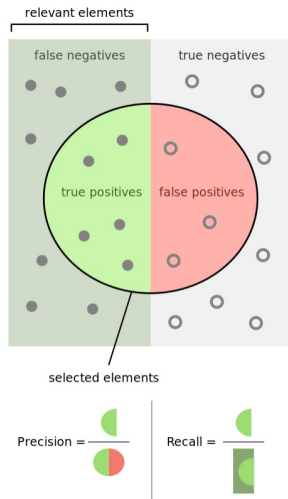
$$R = \frac{TP}{TP + FN} \text{ — полнота, (recall)}$$

$$F_1 = \frac{P + R}{2PR} \text{ — F1-мера}$$

TP (true positive) — найденные релевантные

FP (false positive) — найденные нерелевантные

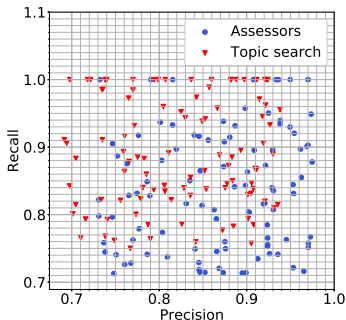
FN (false negative) — ненайденные релевантные



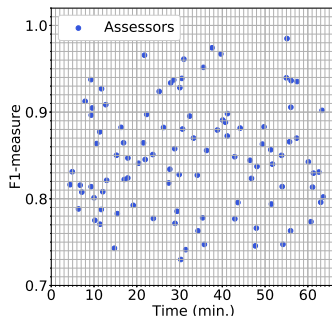
## Результаты измерения точности и полноты по запросам

100 запросов, 3 ассессора на запрос

точность и полнота поиска



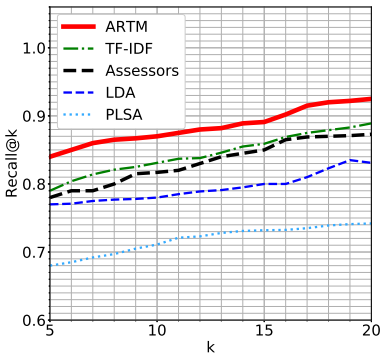
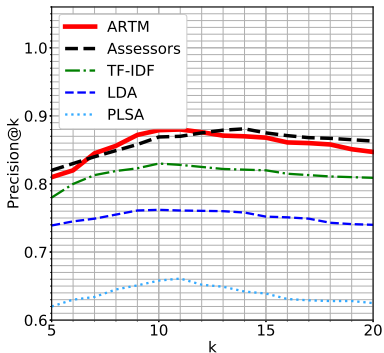
время и  $F_1$ -мера (ассессоры)



- среднее время обработки запроса ассессором — 30 минут
- точность выше у ассессоров, полнота — у поисковика

## Сравнение с ассессорами по качеству поиска

Точность и полнота по первым  $k$  позициям поисковой выдачи (коллекция Habrhabr.ru)

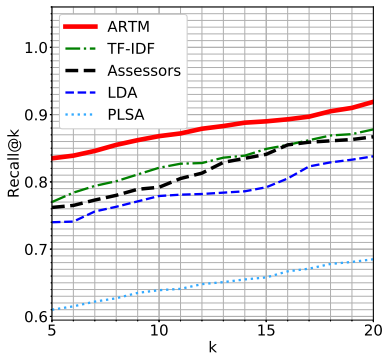
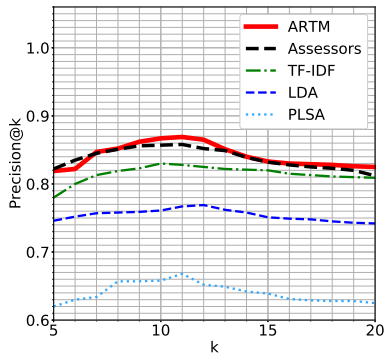


A. Ianina, K. Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.



## Сравнение с ассессорами по качеству поиска

Точность и полнота по первым  $k$  позициям поисковой выдачи (коллекция TechCrunch.com)



A. Ianina, K. Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

## Влияние комбинаций регуляризаторов на качество поиска

Декоррелирование, Θ-разреживание, Φ-сглаживание

	Habrahabr				TechCrunch			
	$R = 0$	Д	ДΘ	ДΘΦ	$R = 0$	Д	ДΘ	ДΘΦ
Prec@5	0.628	0.748	0.771	<b>0.810</b>	0.652	0.775	0.779	<b>0.819</b>
Prec@10	0.653	0.776	0.812	<b>0.879</b>	0.679	0.787	0.819	<b>0.867</b>
Prec@15	0.642	0.765	0.792	<b>0.868</b>	0.669	0.773	0.798	<b>0.833</b>
Prec@20	0.643	0.759	0.783	<b>0.847</b>	0.673	0.777	0.792	<b>0.825</b>
Recall@5	0.692	0.784	0.805	<b>0.840</b>	0.673	0.812	0.812	<b>0.835</b>
Recall@10	0.714	0.814	0.834	<b>0.870</b>	0.685	0.821	0.845	<b>0.868</b>
Recall@15	0.725	0.835	0.867	<b>0.891</b>	0.712	0.859	0.869	<b>0.890</b>
Recall@20	0.735	0.862	0.891	<b>0.925</b>	0.723	0.882	0.895	<b>0.919</b>

- комбинирование регуляризаторов улучшает качество поиска
- хотя исходно все регуляризаторы нацелены на улучшение интерпретируемости тем и не оптимизируют поиск явно

- Тематическое моделирование — это восстановление латентных тем по коллекции текстовых документов
- Задача сводится к стохастическому матричному разложению
- Задача является некорректно поставленной, так как множество её решений в общем случае бесконечно
- Стандартные методы PLSA и LDA не решают эту проблему
- Аддитивная регуляризация (ARTM) доопределяет задачу и позволяет строить модели с заданными свойствами
- Онлайнный EM-алгоритм хорошо распараллеливается и тематизирует большие коллекции за один проход
- Разведочный тематический поиск против ассессоров: точность та же, полнота на 5% выше, 1 сек. вместо 30 мин.