Многокритериальный тематический анализ текстовых коллекций

Воронцов Константин Вячеславович

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

Коллоквиум ФКН НИУ ВШЭ • 5 октября 2017

Содержание

Что такое «тематическое моделирование» (Topic Modeling)

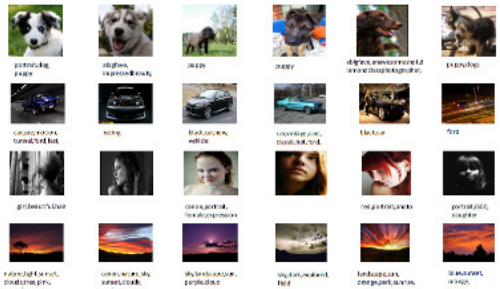
- Одно из направлений обработки естественного языка
- Разновидность статистического анализа текстов
- Технология поиска информации не по словам, а по смыслу
- Выявление скрытых интересов по наблюдаемым данным
- «Мягкая кластеризация» текстовых документов
- Би-кластеризация слов и документов по кластерам-темам
- 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

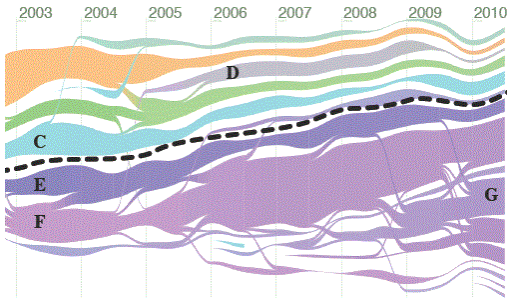
разведочный поиск в электронных библиотеках

персонализированный поиск в соцсетях

мультимодальный поиск текстов и изображений

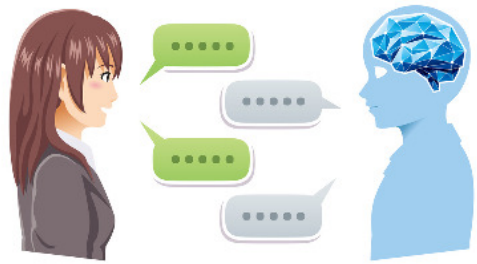


детектирование и трекинг новостных сюжетов



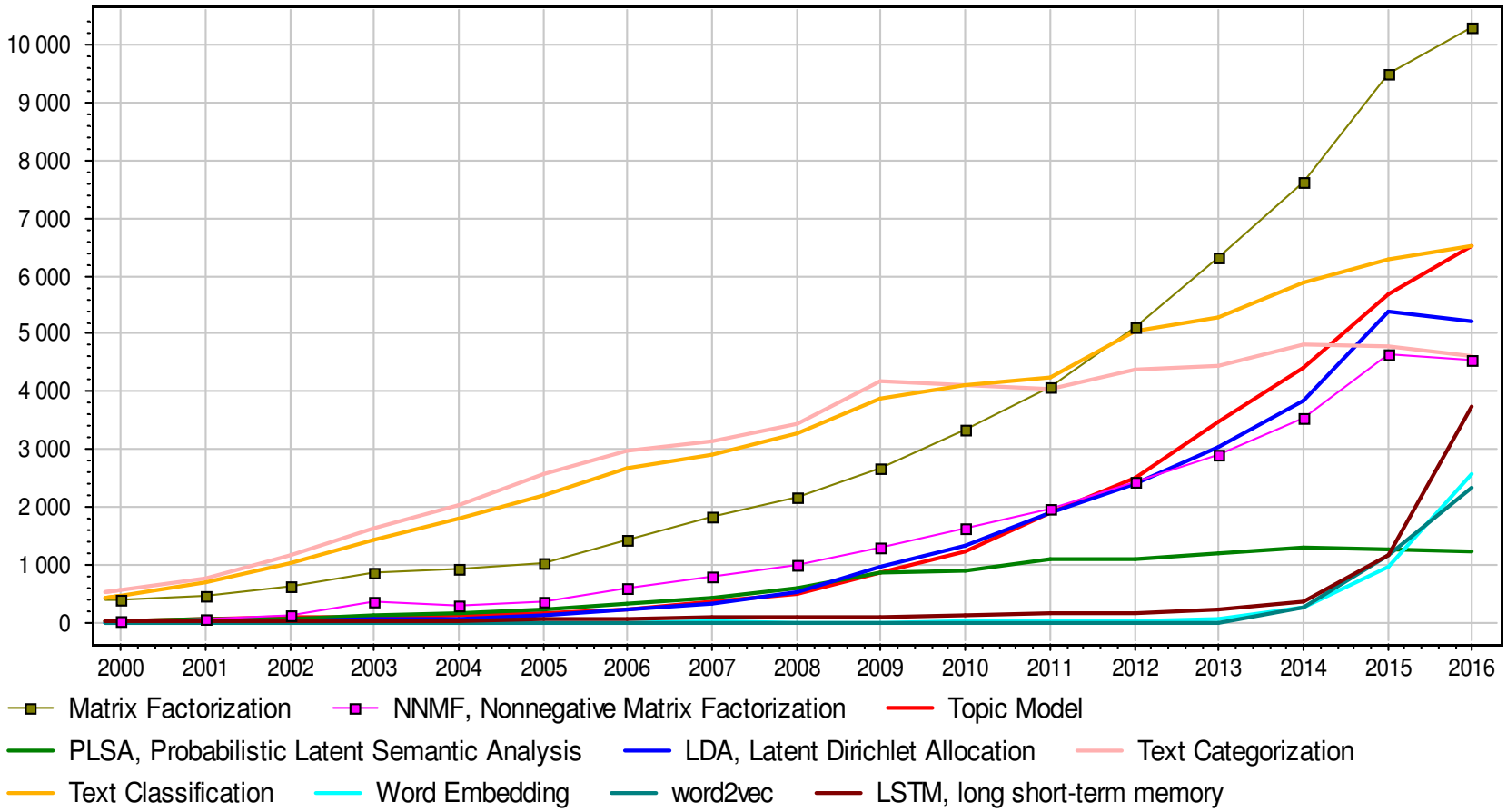
навигация по большим текстовым коллекциям

управлением 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

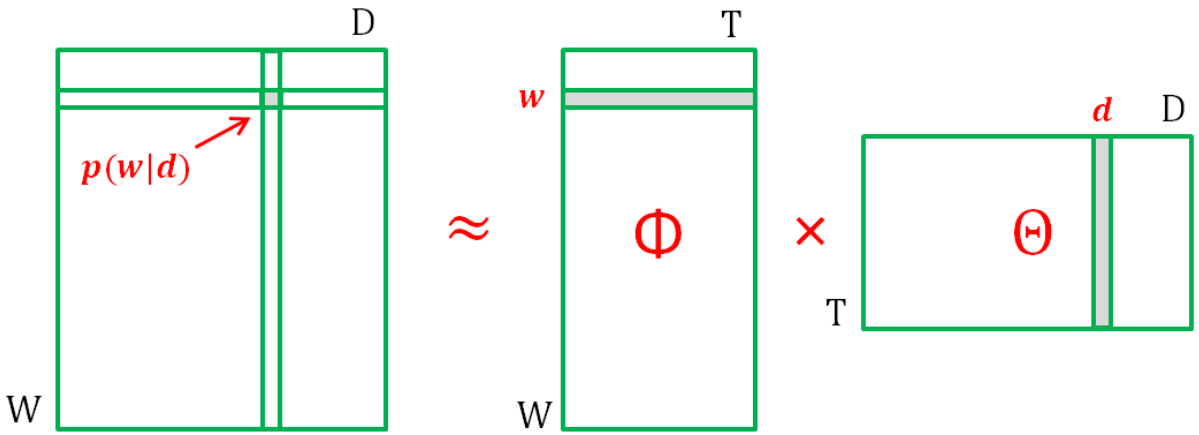
Дано: коллекция текстовых документов

- 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

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

эквивалентна максимизации функционала

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

при ограничениях неотрицательности и нормировки

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

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 Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

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

Регуляризация — стандартный приём доопределения решения с помощью дополнительных критериев.

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} \equiv p(t|d,w) = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}} \right), & n_{wt} = \sum\limits_{d\in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}} \right), & n_{td} = \sum\limits_{w\in d} n_{dw} p_{tdw} \end{cases}$$

где $\underset{t\in T}{\text{norm}}(x_t) = \dfrac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ — 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 ϕ_{wt}, θ_{td} по формуле Байеса:

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

M-шаг: при $R=0$ частотные оценки условных вероятностей вычисляются суммированием счётчика $n_{tdw} = n_{dw}p(t|d,w)$:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_{d\in D} n_{tdw}, \qquad n_t = \sum_{w\in W} n_{wt};$$

$$\theta_{td} = \frac{n_{td}}{n_d}, \qquad n_{td} = \sum_{w\in d} n_{tdw}, \qquad n_d = \sum_{t\in T} n_{td}.$$

Условия вырожденности модели для тем и документов

Решение может быть вырожденным для некоторых тем (столбцов матриц Φ) и документов (столбцов матрицы Θ).

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

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

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 $t \in T$

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

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:

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

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

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

Вывод системы уравнений из условий Каруша–Куна–Таккера

1. Условия ККТ для ϕ_{wt} (для θ_{td} всё аналогично):

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

2. Умножим обе части равенства на ϕ_{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 $\lambda_t \leqslant 0$, 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 λ_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 Θ и тем документов Φ;

инициализация ϕ_{wt}, θ_{td} для всех $d \in D$, $w \in W$, $t \in T$;
для всех итераций $i = 1, \ldots, i_{\max}$
  $n_{wt}, n_{td} := 0$ для всех $d \in D$, $w \in W$, $t \in T$;
  для всех документов $d \in D$ и всех слов $w \in d$
    $n_{tdw} := n_{dw} \operatorname*{norm}_{t \in T}\left(\phi_{wt}\theta_{td}\right)$ для всех $t \in T$;
    $n_{wt} \mathrel{+}= n_{tdw}$; $n_{td} \mathrel{+}= n_{tdw}$ 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 $d \in D$, $t \in T$;

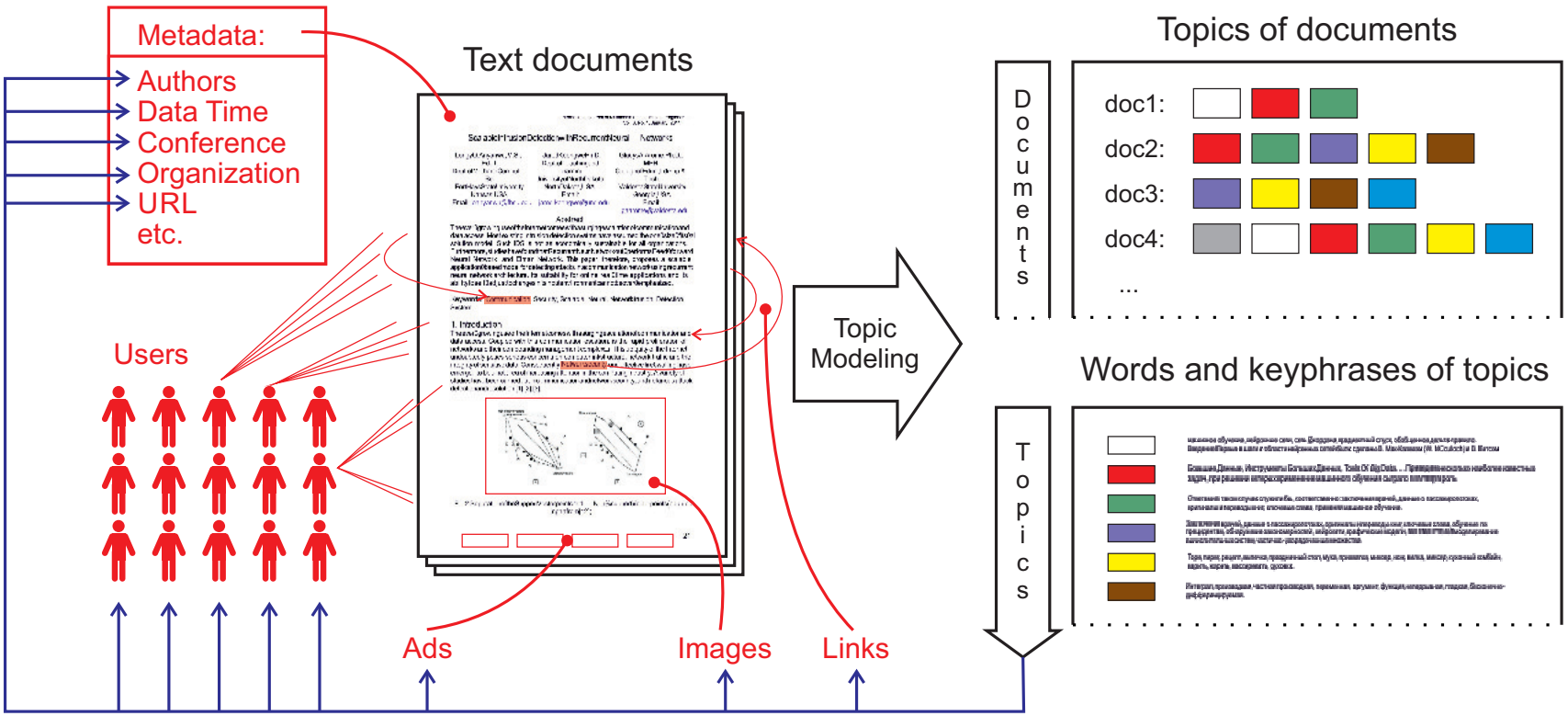
Онлайновый EM-алгоритм (реализован в BigARTM)

Вход: коллекция D, число тем $|T|$, параметры $i_{\max}$, $j_{\max}$, γ;
Выход: матрицы терминов тем Θ и тем документов Φ;

инициализировать $n_{wt} := 0$ и ϕ_{wt};
для всех $i = 1, \ldots, i_{\max}$ (для больших коллекций $i_{\max} = 1$)
  для всех документов $d \in D$
    инициализировать $\theta_{td} := \frac{1}{|T|}$;
    для всех $j = 1, \ldots, j_{\max}$ (итерации по документу)
      $n_{tdw} := n_{dw} \operatorname*{norm}_{t \in T}\left(\phi_{wt}\theta_{td}\right)$ 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 Φ **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*: $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{элемент\_изображения}|t)$, $p(\text{пользователь}|\text{t}), \ldots$



Мультимодальная ARTM

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

Максимизация суммы log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt} \theta_{td}\right)$$

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

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновый 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	
	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	

-- нестандартизуемые 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:

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

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 β_w, α_t:

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

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet allocation. 2003.

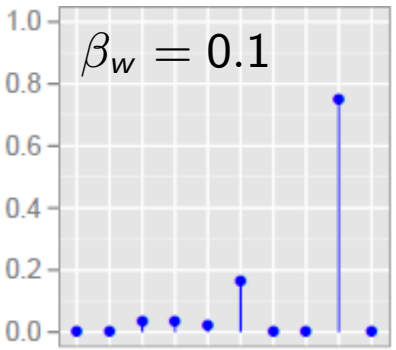
Вероятностная байесовская интерпретация LDA [Blei, 2003]

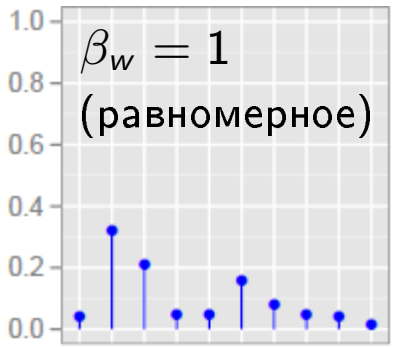
Гипотеза. Вектор-столбцы $\phi_t = (\phi_{wt})_{w \in W}$ и $\theta_d = (\theta_{td})_{t \in T}$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

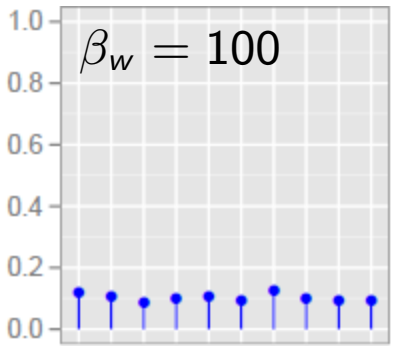
$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \phi_{wt} > 0; \quad \beta_0 = \sum_w \beta_w, \ \beta_t > 0;$$

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

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







Максимизация апостериорной вероятности для модели LDA

Совместное правдоподобие данных и модели:

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

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:

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

при $\beta_w > 1$, $\alpha_t > 1$ — сглаживание,
при $0 < \beta_w < 1$, $0 < \alpha_t < 0$ — 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 ϕ_{wt} и θ_{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 $\beta_w \leqslant 0$, $\alpha_t \leqslant 0$
- слабый регуляризатор: проблема неединственности остаётся

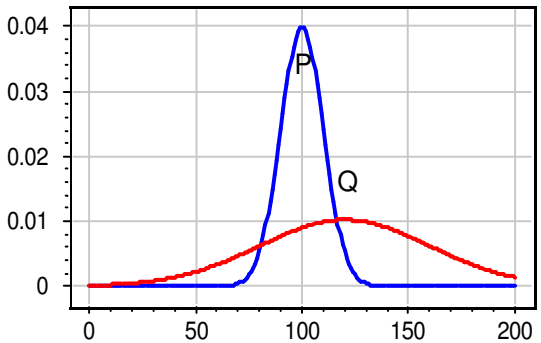
Напоминание. Дивергенция Кульбака–Лейблера

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

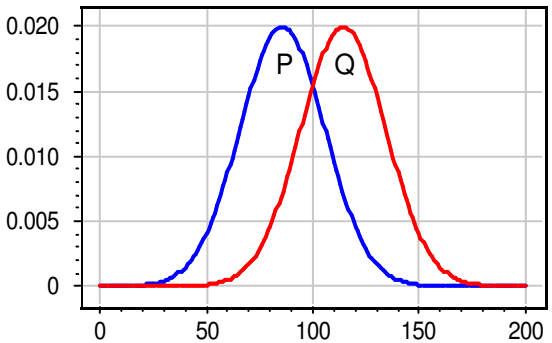
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

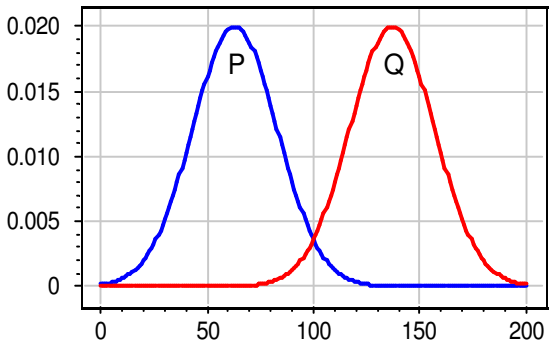
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P вложено в Q:



$\mathrm{KL}(P\|Q) = 0.44$
$\mathrm{KL}(Q\|P) = 2.97$



$\mathrm{KL}(P\|Q) = 0.44$
$\mathrm{KL}(Q\|P) = 0.44$



$\mathrm{KL}(P\|Q) = 2.97$
$\mathrm{KL}(Q\|P) = 2.97$

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 $\alpha_d(t)$:

$$\sum_{t\in T} \tau_t \mathrm{KL}\big(\beta_t(w)\|\phi_{wt}\big) \to \min_{\Phi}; \qquad \sum_{d\in D} \tau_d \mathrm{KL}\big(\alpha_d(t)\|\theta_{td}\big) \to \min_{\Theta}.$$

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:

$$\phi_{wt} = \underset{w}{\mathrm{norm}}\big(n_{wt} + \underbrace{\tau_t\beta_t(w)}_{\beta_{wt}}\big), \qquad \theta_{td} = \underset{t}{\mathrm{norm}}\big(n_{td} + \underbrace{\tau_d\alpha_d(t)}_{\alpha_{td}}\big).$$

Сглаживание, разреживание и частичное обучение тем

Формулы M-шага:

$$\phi_{wt} = \underset{w}{\text{norm}}\big(n_{wt} + \beta_{wt}\big), \qquad \theta_{td} = \underset{t}{\text{norm}}\big(n_{td} + \alpha_{td}\big).$$

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:

- $\beta_{wt} = +\tau_{\text{бт}}[w \in W_t]$ — «белый список» терминов
- $\beta_{wt} = -\tau_{\text{чт}}[w \in W_t]$ — «чёрный список» терминов
- $\alpha_{td} = +\tau_{\text{бд}}[d \in D_t]$ — «белый список» документов
- $\alpha_{td} = -\tau_{\text{чд}}[d \in D_t]$ — «чёрный список» документов

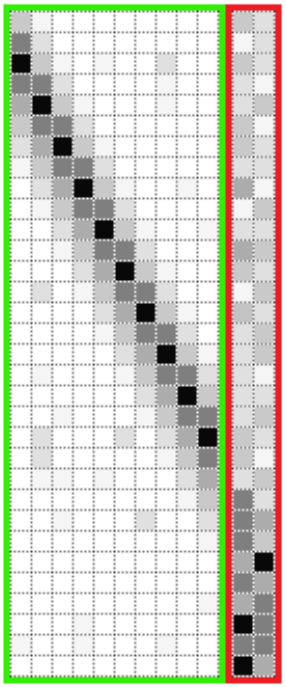
Разделение тем на предметные и фоновые

$T = S \sqcup B$ — множество всех тем

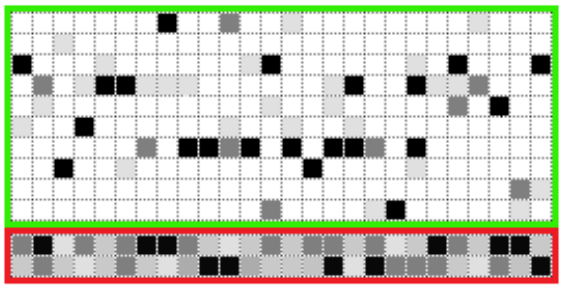
S — разреженные *предметные* темы, специальная лексика

B — сглаженные *фоновые* темы, общая лексика языка

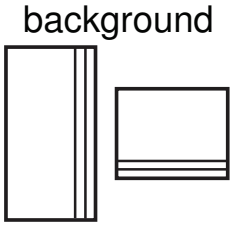
$\Phi_{W \times T}$



$\Theta_{T \times D}$

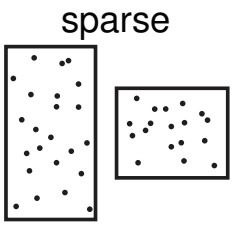


Регуляризаторы для улучшения интерпретируемости тем



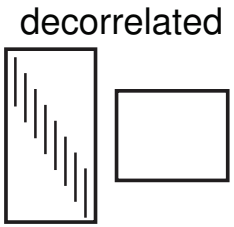
Сглаживание фоновых тем $B \subset T$:

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



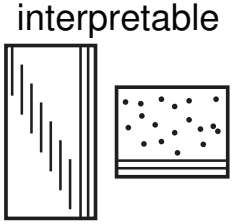
Разреживание предметных тем $S = T \setminus B$:

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



Декоррелирование для повышения различности тем:

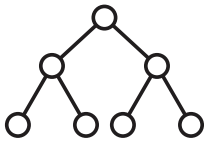
$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_{w} \phi_{wt} \phi_{ws}$$



Сглаживание + разреживание + декоррелирование для улучшения интерпретируемости тем

Иерархические, темпоральные, регрессионные модели

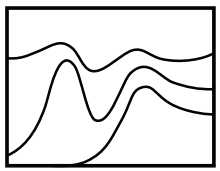
hierarchy



Связь родительских тем t с дочерними подтемами s:

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

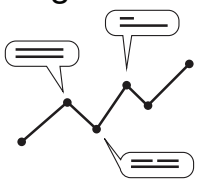
temporal



Темпоральные модели с модальностью времени i:

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

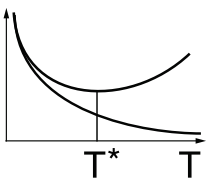
regression



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

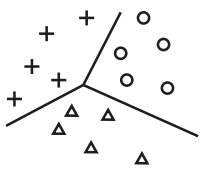


Разреживание $p(t)$ для отбора тем:

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

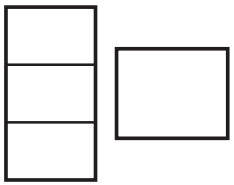
Специальные случаи мультимодальных тематических моделей

supervised



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

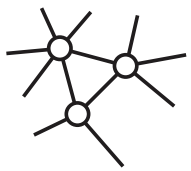
multilanguage



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

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

graph



Модальность вершин графа v, содержащих D_v:

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

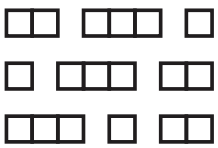
geospatial

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

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

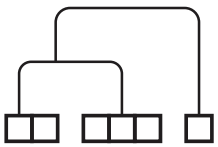
В обход гипотезы «мешка слов» (beyond bag-of-words)

n-gram



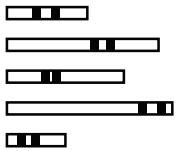
Модальности n-грамм, коллокаций, именованных сущностей

syntax



Модальность n-грамм после применения SyntaxNet

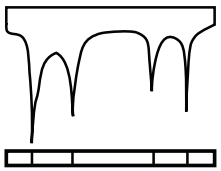
coherence



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 $\tilde{R}$:

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

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

$$\text{E-шаг:}\quad \begin{cases} p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi)}{\partial p_{zdw}}\right)\right) \end{cases}$$

$$\text{M-шаг:}\quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}}\right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}}\right) \end{cases}$$

Доказательство

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

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

Введём функцию от вспомогательных переменных Π:

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

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

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

Лемма 3. Формулы M-шага:

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

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

Поиск и классификация этничного дискурса в соцсетях

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

Используем словарь из 300 этнонимов для обучения тем.

Мешок регуляризаторов:

$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{temporal}\right) + R\left(\text{geospatial}\right) + R\left(\text{sentiment}\right) \to \max$$

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:

$$\mathscr{L}\left(\underset{\text{PLSA}}{\Phi\,\Theta}\right) + R\left(\underset{\text{interpretable}}{}\right) + R\left(\underset{\text{multimodal}}{}\right) + R\left(\underset{\text{n-gram}}{}\right) \to \max$$

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 асессоров.
- Тематический поиск даёт результат мгновенно,
  асессоры 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 именование всех тем
- управлять медиакомпаниями и творческими заданиями

Мешок регуляризаторов:

$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{interpretable}\right) + R\left(\text{hierarchy}\right) + R\left(\text{temporal}\right)$$
$$+ R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) + R\left(\text{multilanguage}\right) + R\left(\text{sentiment}\right) \to \max$$

Результат: ... (исследование продолжается)

Сценарный анализ записей разговоров контакт-центра

Задачи:

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

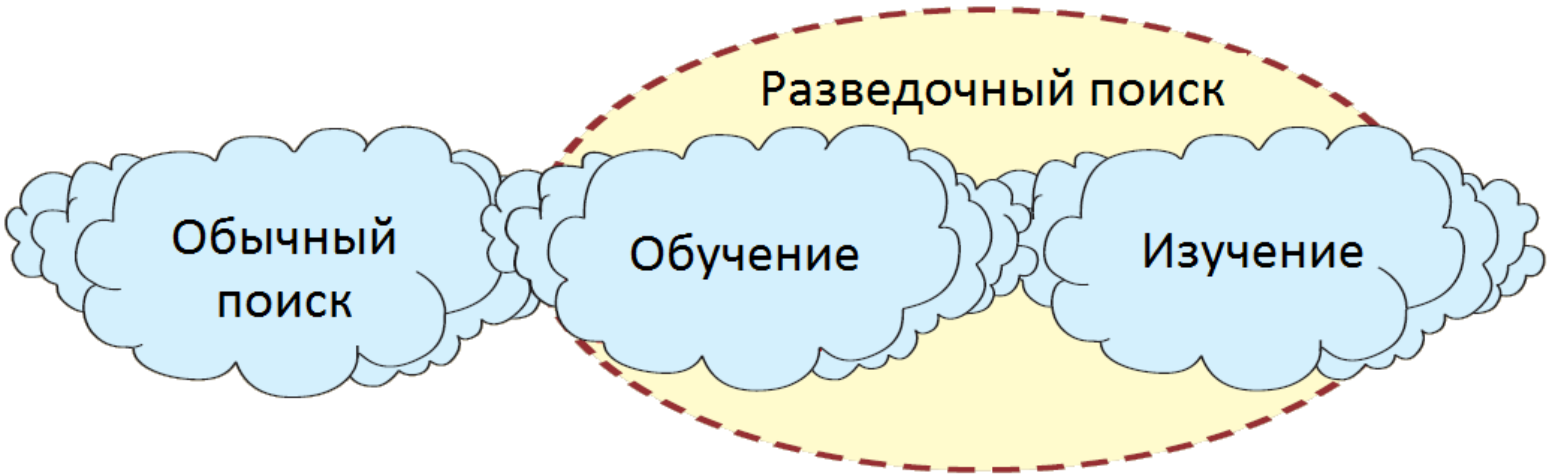
Мешок регуляризаторов:

$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{interpretable}\right) + R\left(\text{segmentation}\right) + R\left(\text{n-gram}\right) + R\left(\text{syntax}\right) + R\left(\text{sentence}\right) + R\left(\text{dialog}\right) \to \max$$

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



навигация в сети,	самообразование,	исследование,
поиск фактов,	тематический поиск	экспертиза,
упоминаний,	систематизация	реферирование,
конкретных ответов	знаний	мониторинг тем

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

Разведочный тематический поиск

$q = (w_1, \ldots, w_{n_q})$ — 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:

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

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

Две коллекции новостей про технологии

Habrahabr.ru

175 143 статей на русском
10 552 слов (униграмм)
742 000 биграмм
524 авторов статей
10 000 авторов комментариев
2546 тегов
123 хаба (категории)

TechCrunch.com

759 324 статей на английском
11 523 слов (униграмм)
1.2 млн. биграмм
605 авторов
184 категорий

Предобработка текстов

- отброшены 5% наиболее частотных слов (общая лексика)
- удаление пунктуации
- нижний регистр, ё→е
- лемматизация pymorphy2

Методика оценивания качества разведочного поиска

Поисковый запрос

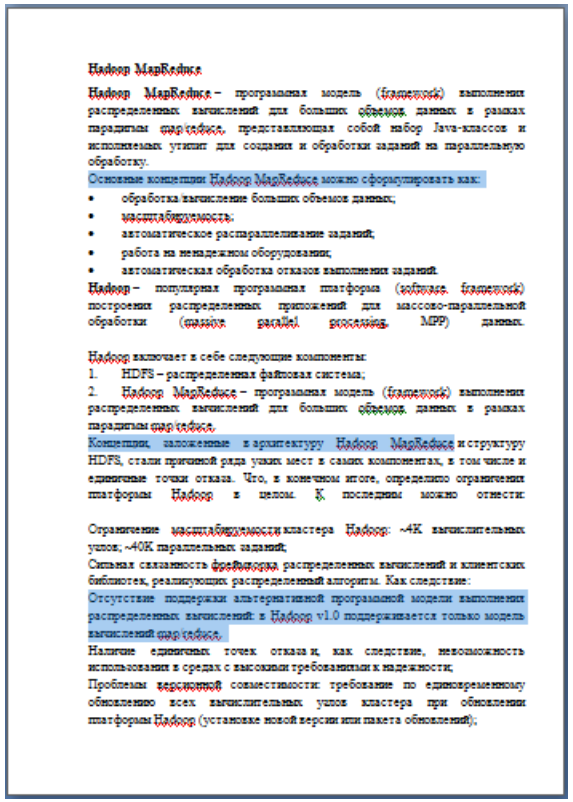
набор ключевых слов или фрагментов текста, около одной страницы A4

Поисковая выдача

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

Два задания асессорам

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



Hadoop MapReduce

Hadoop MapReduce – программная модель (framework) выполнения распределенных вычислений для больших объемов данных в рамках парадигмы map/reduce, представляющая собой набор Java-классов и исполняемых утилит для создания и обработки заданий на параллельную обработку.

Основные концепции Hadoop MapReduce можно сформулировать как:

- обработка/вычисление больших объемов данных;
- масштабируемость;
- автоматическое распараллеливание заданий;
- работа на ненадежном оборудовании;
- автоматическая обработка отказов выполнения заданий.

Hadoop – популярная программная платформа (software framework) построения распределенных приложений для массово-параллельной обработки (massive parallel processing, MPP) данных.

Hadoop включает в себе следующие компоненты:

1. HDFS – распределенная файловая система;
2. Hadoop MapReduce – программная модель (framework) выполнения распределенных вычислений для больших объемов данных в рамках парадигмы map/reduce.

Концепции, заложенные в архитектуру Hadoop MapReduce и структуру HDFS, стали причиной ряда узких мест в самих компонентах, в том числе и единичные точки отказа. Что, в конечном итоге, определило ограничения платформы Hadoop в целом. К последним можно отнести:

Ограничение масштабируемости кластера Hadoop: ~4K вычислительных узлов; ~40K параллельных заданий;

Сильная связанность фреймворка распределенных вычислений и клиентских библиотек, реализующих распределенный алгоритм. Как следствие:

Отсутствие поддержки альтернативной программной модели выполнения распределенных вычислений: в Hadoop v1.0 поддерживается только модель вычислений map/reduce.

Наличие единичных точек отказа и, как следствие, невозможность использования в средах с высокими требованиями к надежности;

Проблемы версионной совместимости: требование по единовременному обновлению всех вычислительных узлов кластера при обновлении платформы Hadoop (установке 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.

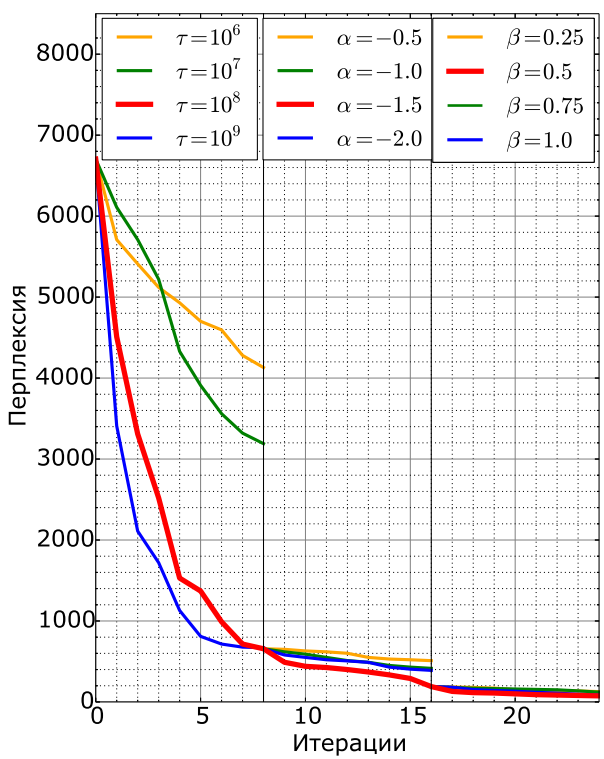
Тематика запросов разведочного поиска

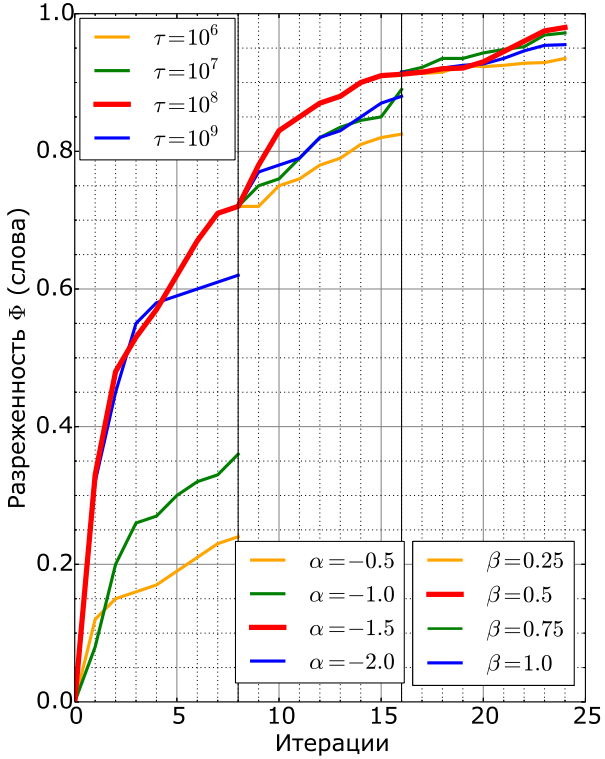
Примеры заголовков разведочных запросов к Хабру (объём каждого запроса — около одной страницы A4):

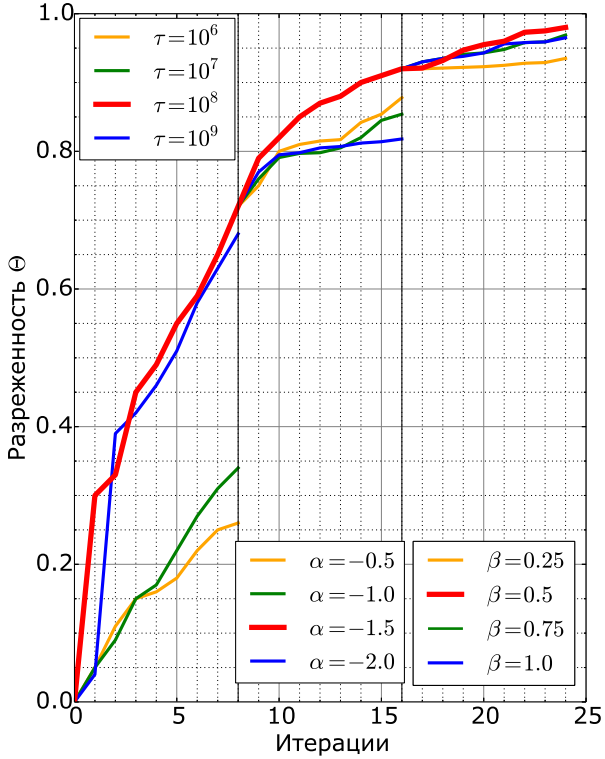
Алгоритмы раскраски графов	Система IBM Watson
Рекомендательная система Netflix	3D-принтеры
Методики быстрого набора текста	CERN-кластер
Космические проекты Илона Маска	AB-тестирование
Технологии Hadoop MapReduce	Облачные сервисы
Беспилотный автомобиль Google car	Контекстная реклама
Криптосистемы с открытым 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 (τ),
- разреживание распределений тем в документах (α),
- сглаживание распределений терминов в темах (β).







Оценки качества поиска

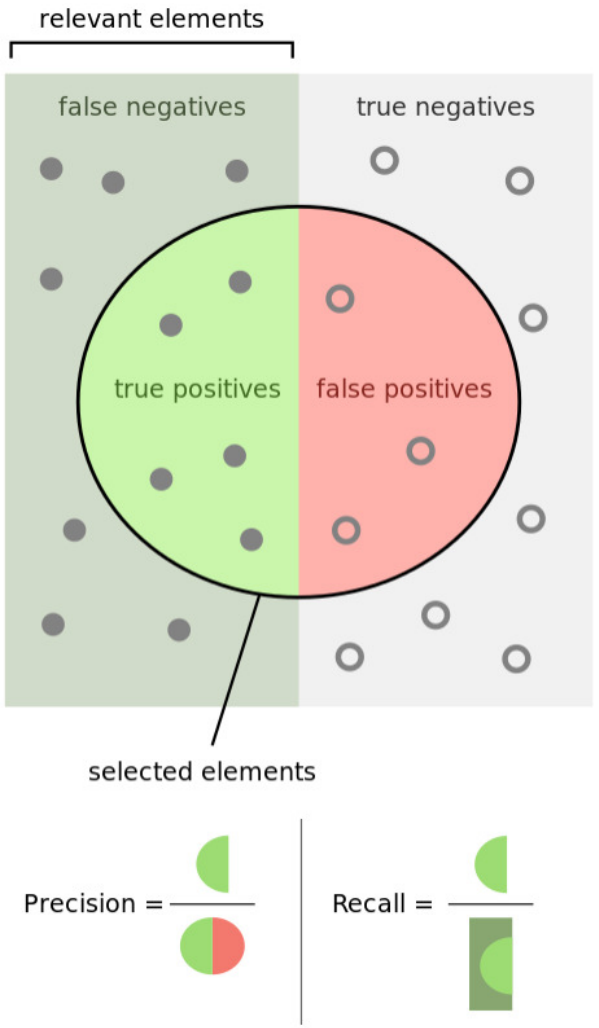
Precision — доля релевантных среди найденных
Recall — доля найденных среди релевантных

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

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

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

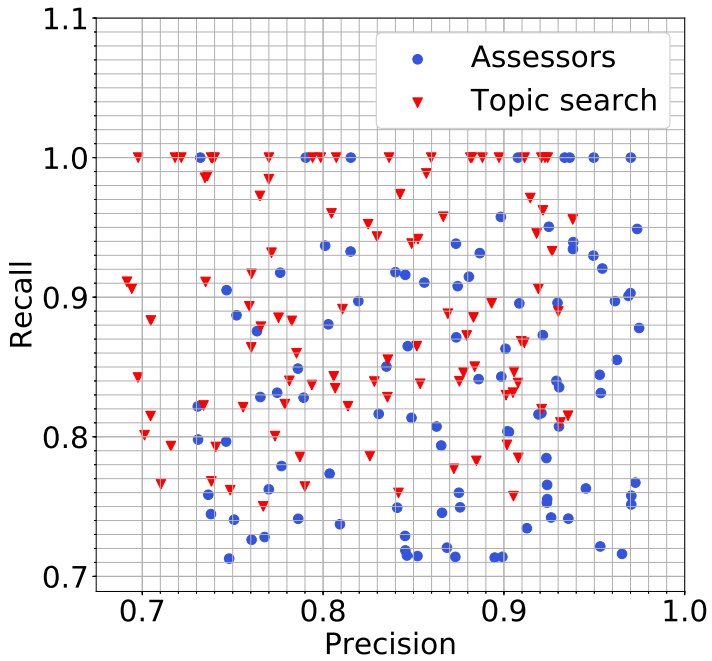
TP (true positive) — найденные релевантные
FP (false positive) — найденные нерелевантные
FN (false negative) — ненайденные релевантные



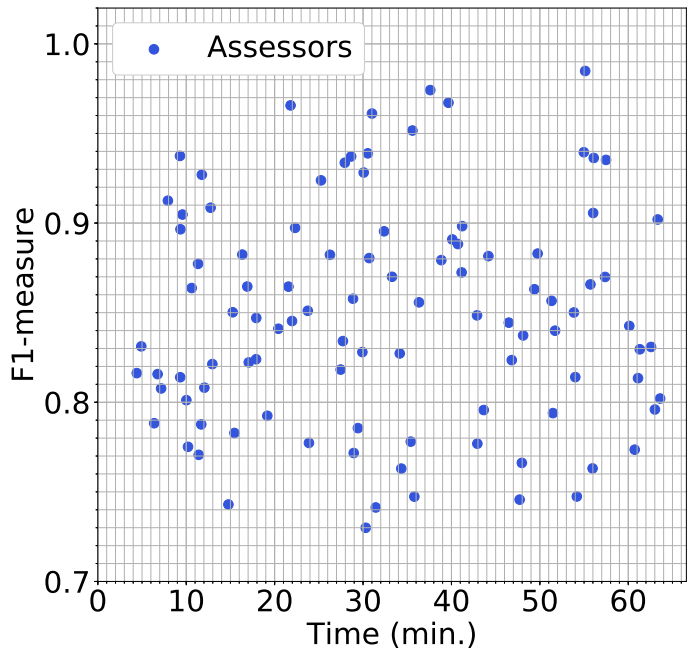
Результаты измерения точности и полноты по запросам

100 запросов, 3 асессора на запрос

точность и полнота поиска



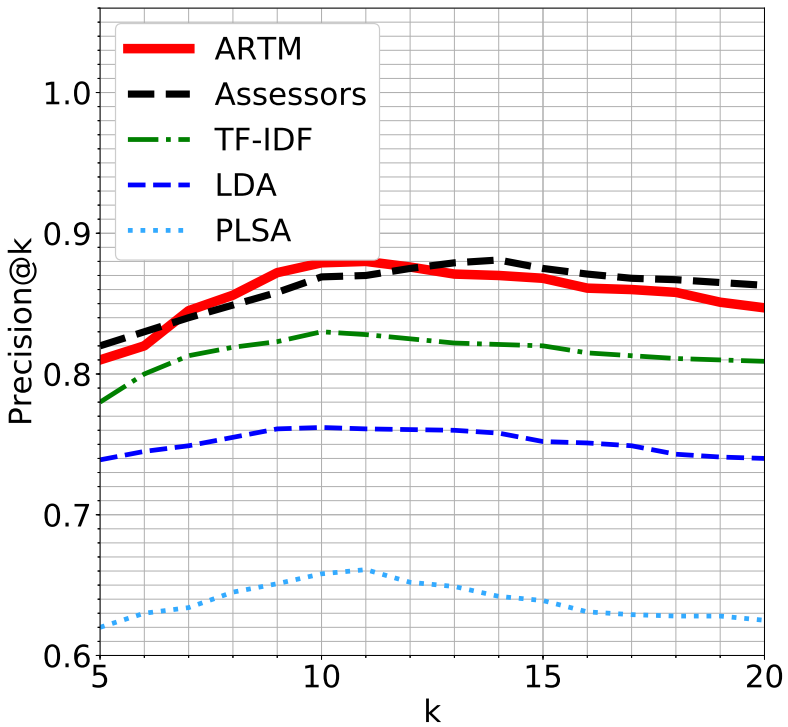
время и F_1-мера (асессоры)

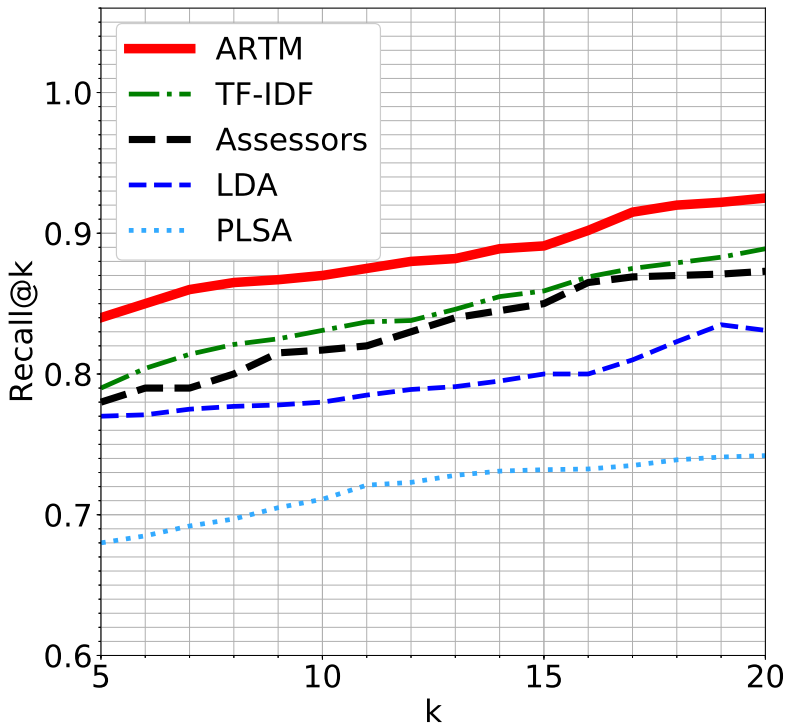


- среднее время обработки запроса асессором — 30 минут
- точность выше у асессоров, полнота — у поисковика

Сравнение с асессорами по качеству поиска

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

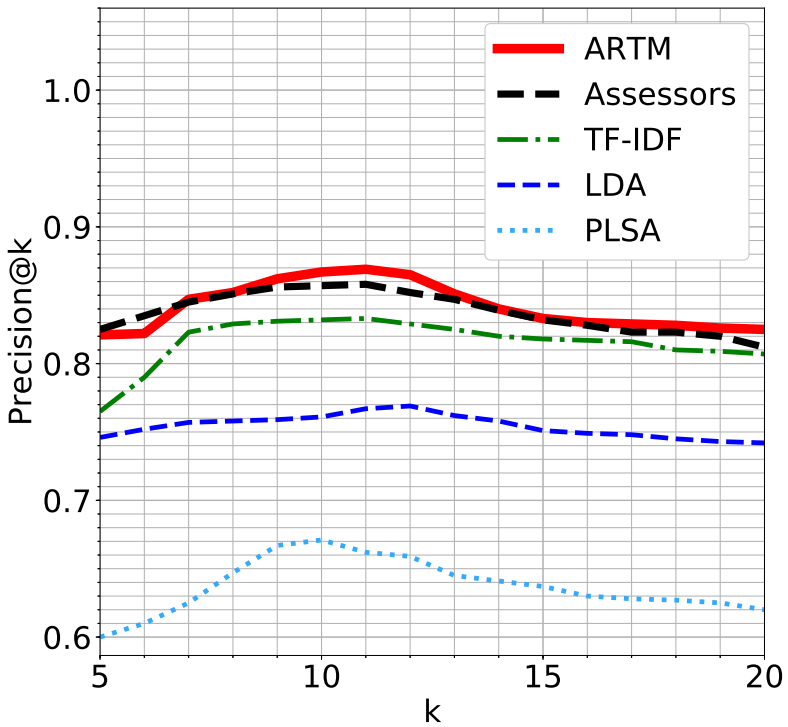


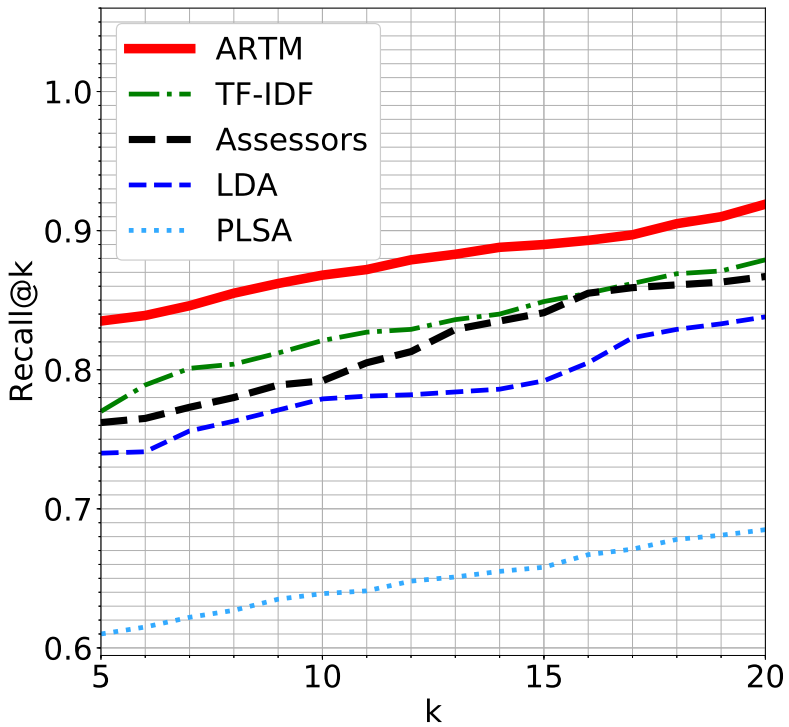


A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Сравнение с асессорами 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 меры близости документа и запроса на качество поиска

Меры близости распределений:
Euclidean, Cosine, Manhattan, Hellinger, Kullback–Leibler

	Коллекция Habrahabr.ru					Коллекция TechCrunch.com				
	Eu	**cos**	Ma	He	KL	Eu	**cos**	Ma	He	KL
Prec@5	0.612	**0.810**	0.682	0.709	0.721	0.635	**0.819**	0.673	0.732	0.715
Prec@10	0.657	**0.879**	0.697	0.735	0.749	0.665	**0.867**	0.683	0.752	0.732
Prec@15	0.627	**0.868**	0.635	0.727	0.711	0.643	**0.833**	0.642	0.742	0.724
Prec@20	0.619	**0.847**	0.627	0.728	0.707	0.638	**0.825**	0.638	0.729	0.708
Recall@5	0.672	**0.840**	0.692	0.721	0.803	0.658	**0.835**	0.669	0.733	0.775
Recall@10	0.682	**0.870**	0.707	0.775	0.856	0.671	**0.868**	0.682	0.753	0.787
Recall@15	0.705	**0.891**	0.725	0.791	0.878	0.715	**0.890**	0.708	0.785	0.809
Recall@20	0.703	**0.925**	0.732	0.812	0.888	0.712	**0.919**	0.715	0.808	0.812

- Наилучшее качество поиска — при косинусной мере
- Одни и те же асессорские оценки можно использовать для оценивания новых моделей и поисковых движков

Влияние комбинаций регуляризаторов на качество поиска

Декоррелирование, Θ-разреживание, Φ-сглаживание

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

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

Влияние сочетания модальностей на качество поиска

Коллекция Habrahabr.ru. Число тем $|T| = 200$. Модальности: Слова, Биграммы, Теги, Хабы, Комментаторы, Авторы.

	асессоры	С	К	СБ	СБТХ	все
Prec@5	0.821	0.612	0.549	0.654	0.737	**0.810**
Prec@10	0.869	0.635	0.568	0.701	0.752	**0.879**
Prec@15	0.875	0.625	0.532	0.685	0.682	**0.868**
Prec@20	0.863	0.616	0.533	0.682	0.687	**0.847**
Recall@5	0.780	0.722	0.636	0.797	0.827	**0.840**
Recall@10	0.817	0.744	0.648	0.812	0.875	**0.870**
Recall@15	0.850	0.778	0.677	0.842	0.893	**0.891**
Recall@20	0.873	0.803	0.685	0.852	0.898	**0.925**

- Наилучшее качество поиска — по всем модальностям
- Наиболее полезные модальности — слова и теги

Влияние сочетания модальностей на качество поиска

Коллекция TechCrunch.com. Число тем $|T| = 450$.
Модальности: Слова, Категории, Биграммы, Авторы.

	асессоры	С	К	СБ	СБК	все
Prec@5	0.822	0.711	0.557	0.767	0.808	**0.819**
Prec@10	0.851	0.721	0.581	0.783	0.818	**0.867**
Prec@15	0.835	0.733	0.594	0.793	0.833	**0.833**
Prec@20	0.813	0.727	0.566	0.772	0.822	**0.825**
Recall@5	0.762	0.752	0.657	0.775	0.825	**0.835**
Recall@10	0.792	0.776	0.669	0.808	0.855	**0.868**
Recall@15	0.835	0.782	0.684	0.825	0.877	**0.890**
Recall@20	0.867	0.825	0.702	0.837	0.901	**0.919**

- Наилучшее качество поиска — по всем модальностям
- Наиболее полезные модальности — слова и категории

Влияние числа тем на качество поиска

Коллекция Habrahabr.ru

Используем все 5 модальностей, меняем $|T|$

	асессоры	100	150	**200**	250	400
Prec@5	0.821	0.662	0.721	**0.810**	0.761	0.693
Prec@10	0.869	0.761	0.812	**0.879**	0.825	0.673
Prec@15	0.875	0.733	0.795	**0.868**	0.791	0.651
Prec@20	0.863	0.724	0.795	**0.847**	0.792	0.642
Recall@5	0.780	0.732	0.807	**0.840**	0.821	0.721
Recall@10	0.817	0.771	0.843	**0.870**	0.851	0.751
Recall@15	0.850	0.824	0.895	**0.891**	0.871	0.773
Recall@20	0.873	0.857	0.905	**0.925**	0.892	0.771

- Наилучшее качество поиска — при 200 темах
- Тематический поиск превосходит асессоров по полноте

Влияние числа тем на качество поиска

Коллекция TechCrunch.com

Используем все 4 модальности, меняем $|T|$

	асессоры	350	400	450	**475**	500
Prec@5	0.822	0.653	0.725	0.752	**0.819**	0.777
Prec@10	0.851	0.663	0.732	0.762	**0.867**	0.811
Prec@15	0.835	0.682	0.743	0.787	**0.833**	0.793
Prec@20	0.813	0.650	0.743	0.773	**0.825**	0.793
Recall@5	0.762	0.731	0.762	0.793	**0.835**	0.817
Recall@10	0.792	0.763	0.793	0.812	**0.868**	0.855
Recall@15	0.835	0.782	0.807	0.855	**0.890**	0.882
Recall@20	0.867	0.792	0.823	0.862	**0.919**	0.903

- Наилучшее качество поиска — при 475 темах
- Тематический поиск превосходит асессоров по полноте

Резюме

- 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
- Онлайновый EM-алгоритм хорошо распараллеливается и тематизирует большие коллекции за один проход
- Разведочный тематический поиск против асессоров: точность та же, полнота на 5% выше, 1 сек. вместо 30 мин.