

Регуляризация тематических моделей: библиотека с открытым кодом BigARTM и разведочный информационный поиск

Воронцов Константин Вячеславович,
Фрей А. И., Потапенко А. А., Апишев М. А., Ромов П. А.,
Дойков Н. В., Суворова М. А., Царьков С. В.

ФИЦ ИУ РАН • МФТИ • МГУ



Светлогорск • 19–25 сентября 2015

- 1 Вероятностное тематическое моделирование**
 - Задача стохастического матричного разложения
 - Разведочный информационный поиск
 - PLSA, LDA и байесовские тематические модели
- 2 Аддитивная регуляризация тематических моделей**
 - Регуляризованные и мультимодальные модели
 - Проект BigARTM
 - Эксперименты с BigARTM
- 3 Дальнейшие обобщения ARTM**
 - Тематические модели на гиперграфах
 - Лингвистическая регуляризация
 - Выводы. Направления дальнейших исследований

Что такое «тема»?

- *Тема* — специальная терминология предметной области.
- *Тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах.

Более формально,

- *тема* — условное распределение на множестве терминов, $p(w|t)$ — вероятность термина w в теме t ;
- *тематический профиль* документа — условное распределение $p(t|d)$ — вероятность темы t в документе d .

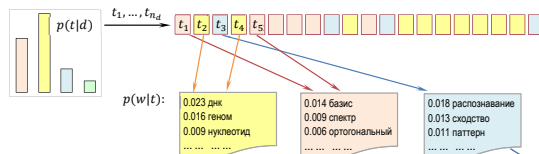
Когда автор писал термин w в документе d , он думал о теме t , и мы хотели бы выявить, о какой именно.

Тематическая модель выявляет латентные темы по наблюдаемым распределениям слов $p(w|d)$ в документах.

Прямая задача — порождение коллекции по $p(w|t)$ и $p(t|d)$

Вероятностная тематическая модель коллекции документов D описывает появление терминов w в документах d темами t :

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



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).

Обратная задача — восстановление $p(w|t)$ и $p(t|d)$ по коллекции

Дано: W — словарь терминов

D — коллекция текстовых документов $d = \{w_1 \dots w_{n_d}\}$

n_{dw} = сколько раз термин w встречается в документе d

Найти параметры модели $\frac{n_{dw}}{n_d} \equiv p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$:

$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t

$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Задача стохастического матричного разложения $\left(\frac{n_{dw}}{n_d}\right)_{W \times D} = \Phi_{W \times T} \cdot \Theta_{T \times D}$

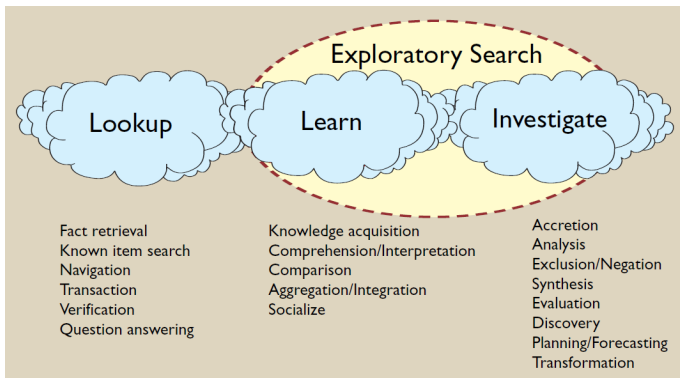
некорректно поставлена — решение не единственно:

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

для невырожденных $S_{T \times T}$ таких, что Φ', Θ' — стохастические.

Разведочный поиск — знания «на кончиках пальцев»

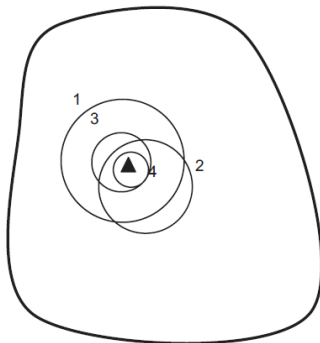
- пользователь может не знать ключевых терминов
- пользователя может интересовать множество ответов



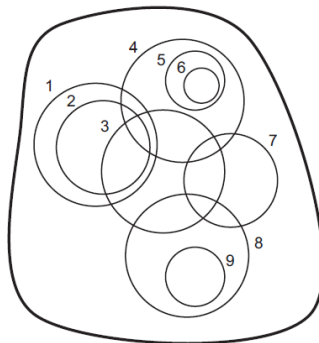
Gary Marchionini. Exploratory Search: from finding to understanding.
Communications of the ACM. 2006, 49(4), p. 41–46.

От поиска «query-browse-refine» к разведочному поиску

Iterative Search



Exploratory Search



- ▲ Search target ◊ Information space
○ Result sets (larger = more results, intersection = overlap, # = iteration)

R.W.White, R.A.Roth. Exploratory Search: beyond the Query-Response paradigm. San Rafael, CA: Morgan and Claypool, 2009.

Возможный сценарий разведочного поиска

Поисковый запрос:

- документ любой длины или даже коллекция документов

Цели поиска:

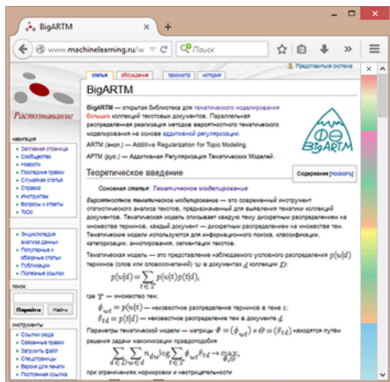
- к каким темам относится мой запрос?
- что ещё известно по этим темам?
- какова тематическая структура этой предметной области?
- что ещё есть понятного, обзорного, важного, свежего?

Сценарий поиска:

- 1 имея любой текст под рукой, в любом приложении,
- 2 хотим получить картину содержащихся в нём тем-подтем,
- 3 и «дорожную карту» предметной области в целом

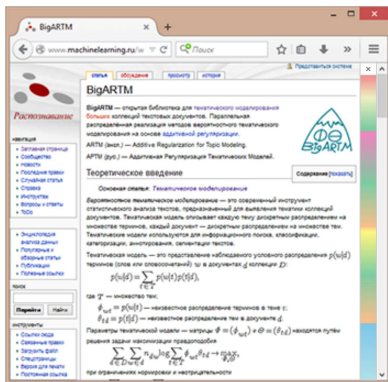
Разведочный поиск: прототип интерфейса

Радужная полоса напоминает, что знания всегда под рукой



Разведочный поиск: прототип интерфейса

Клик по **радужной полосе** — тематический поисковый запрос



Разведочный поиск: прототип интерфейса

Темы-подтемы выбранного фрагмента текста

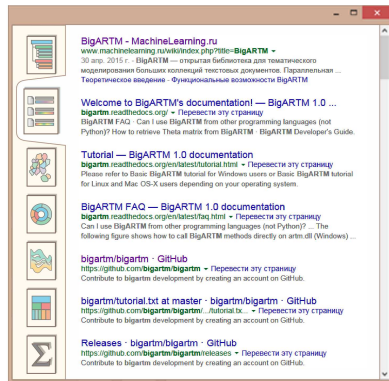
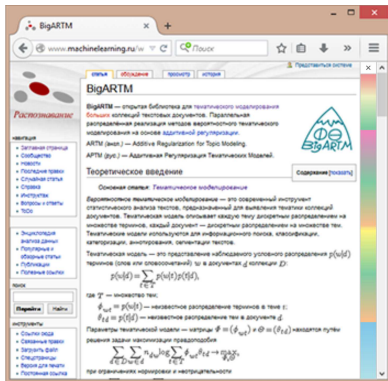
The screenshot shows the BigARTM website. The main content area includes a header with navigation tabs (глава, обзоры, проекты, история), a logo, and introductory text about the library. A sidebar on the left contains a menu with categories like 'Новости', 'События', and 'Экспозиция'. The main text area contains a section titled 'Теоретическое введение' with mathematical formulas and text describing the model.

The screenshot shows a window titled 'Topics in «BigARTM»' with a language selector for English and Russian. It displays a hierarchical list of topics and sub-topics:

- Natural language processing
 - Statistical text analysis
 - Probabilistic topic modeling
- Probability theory
 - Likelihood maximization
- Mathematical programming
 - Nonconvex optimization
 - Constrained nonconvex optimization
- Machine Learning
 - Topic Modeling
 - Probabilistic Topic Modeling
- Matrix Factorization
 - Nonnegative Matrix Factorization
 - Probabilistic Topic Modeling
- Parallel computing
- Big Data

Разведочный поиск: прототип интерфейса

Документы и иные объекты, ранжированные по релевантности



Разведочный поиск: прототип интерфейса

Дорожная карта: кластеризация релевантных документов

The screenshot shows the BigARTM web interface. At the top, there are navigation tabs: "глава", "обсуждение", "протокол", "история". The main content area is titled "BigARTM" and contains introductory text about the software. A sidebar on the left lists various categories like "Новости", "События", "Справочник". The main text includes a section "Теоретическое введение" with a sub-section "Основная идея: Тематическое моделирование".

Теоретическое введение

Основная идея: Тематическое моделирование

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выявления тематик коллекций документов. Тематическая модель связывает каждую тему с определенным набором терминов, каждый документ — с определенным распределением на множестве тем. Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, семантики текстов.

Тематическая модель — это представление наблюдаемого условного распределения $p(w|d)$ термов (слов или словосочетаний) w в документе d коллекции D :

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

где T — множество тем;

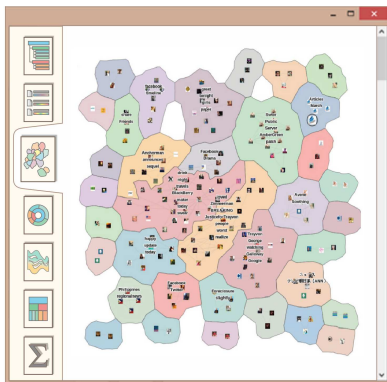
$$\phi_{wt} = p(w|t) \text{ — неизвестное распределение термов в теме } t;$$

$$\theta_{dt} = p(t|d) \text{ — неизвестное распределение тем в документе } d.$$

Параметры тематической модели — матрицы $\Phi = (\phi_{wt})$ и $\Theta = (\theta_{dt})$ находят путем решения задачи максимизации правдоподобия

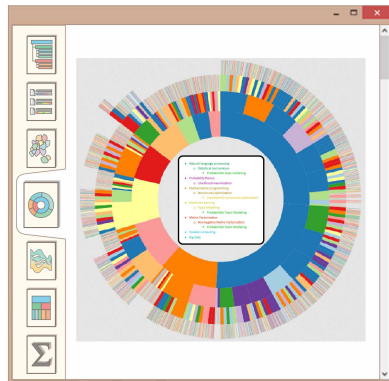
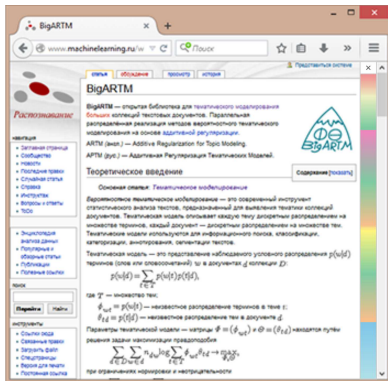
$$\sum_{d \in D} \sum_{w \in V} p_w \log \sum_{t \in T} \phi_{wt} \theta_{dt} \rightarrow \max_{\Phi, \Theta},$$

при ограничениях нормировки и неотрицательности



Разведочный поиск: прототип интерфейса

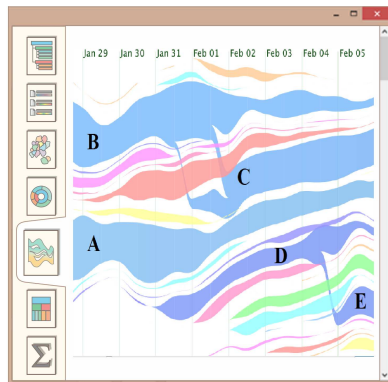
Тематическая иерархия: структура предметной области



Разведочный поиск: прототип интерфейса

Динамика тем: эволюция предметной области

The screenshot shows the BigARTM web interface. The browser address bar displays 'www.machinelearning.ru'. The page title is 'BigARTM'. The main content area includes a description of the software as a library for topic modeling, a 'Теоретическое введение' (Theoretical Introduction) section, and a 'Содержание' (Table of Contents) button. The interface is designed with a clean, professional look, featuring a sidebar with navigation options and a main content area with clear typography and a logo.



Разведочный поиск: прототип интерфейса

Тематическая сегментация документа запроса

BigARTM — открытая библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределенная реализация метода вероятностного тематического моделирования на основе аддитивной регуляризации.

ARTM (англ.) — Additive Regularization for Topic Modeling.
 ARTM (рус.) — Аддитивная Регуляризация Тематических Моделей.

Теоретическое введение

Основная идея: Тематическое моделирование

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выявления тематик коллекций документов. Тематическая модель использует каждую тему (дискретное распределение на множестве термов), каждый документ — дискретное распределение на множестве тем. Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, сегментации текстов.

Тематическая модель — это представление наблюдаемого условного распределения $p(w|d)$ термов (или для ассоциативной) z в документе d коллекции D :

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

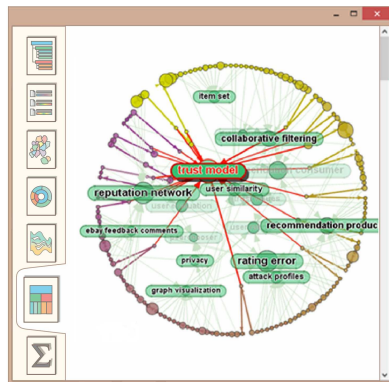
где T — множество тем;

$\phi_{wt} = p(w|t)$ — известное распределение термов в теме t ;
 $\theta_{dt} = p(t|d)$ — известное распределение тем в документе d .

Параметры тематической модели — матрицы $\Phi = (\phi_{wt})$ и $\Theta = (\theta_{dt})$ находят путем решения задачи максимизации правдоподобия

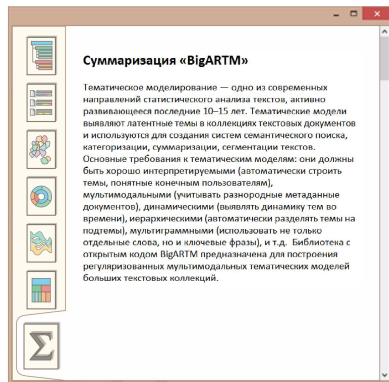
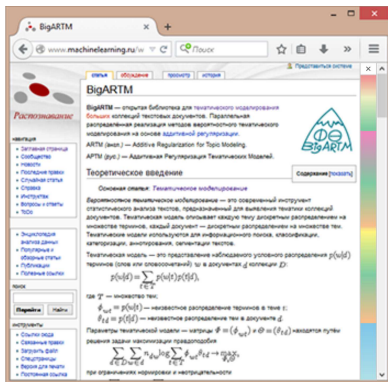
$$\sum_{d \in D} \sum_{w \in V} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{dt} \rightarrow \max_{\Phi, \Theta},$$

при ограничениях неотрицательности и нормированности



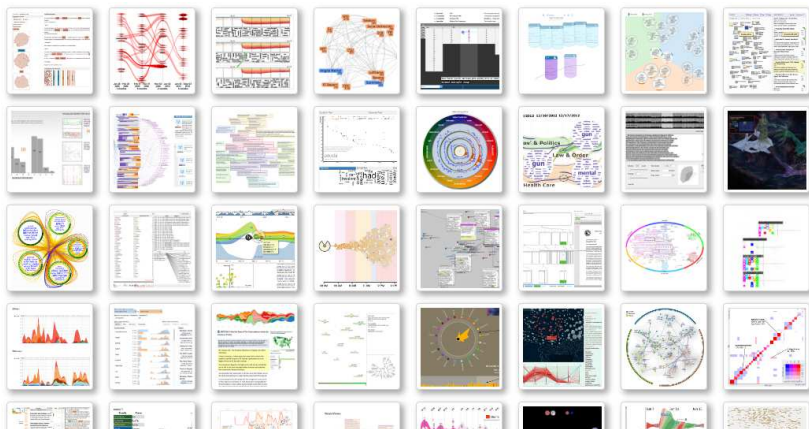
Разведочный поиск: прототип интерфейса

Суммаризация документа запроса



<http://textvis.lnu.se>

Интерактивный обзор ~~170~~ 220 средств визуализации текстов



Технологические элементы разведочного поиска

- 1 Интернет-краулинг имеются готовые решения
- 2 Фильтрация контента имеются готовые решения
- 3 Тематическое моделирование **математика здесь**
- 4 Инвертированный индекс имеются готовые решения
- 5 Ранжирование имеются готовые решения
- 6 Визуализация имеются готовые решения

Тематическая модель для разведочного поиска должна быть...

- 1 Интерпретируемая: каждая тема понятна людям
- 2 Мультиграммная: термины-словосочетания неразрывны
- 3 Мультимодальная: авторы, связи, тэги, пользователи, ...
- 4 Мультиязычная: для кросс- и много-языкового поиска
- 5 Динамическая: выявление истории развития тем
- 6 Иерархическая: выявление иерархических связей тем
- 7 Сегментирующая: выделение тем внутри документа
- 8 Обучаемая по оценкам ассессоров и пользователей
- 9 Определяющая число тем автоматически
- 10 Создающая новые темы автоматически
- 11 Именуемая новые темы автоматически
- 12 Онлайновая: обрабатывающая коллекцию за 1 проход

Обратная задача — восстановление $p(w|t)$ и $p(t|d)$ по коллекции

Дано: W — словарь терминов

D — коллекция текстовых документов $d = \{w_1 \dots w_{n_d}\}$

n_{dw} = сколько раз термин w встречается в документе d

Найти параметры модели $p(w|d) = \sum_t \phi_{wt} \theta_{td}$:

$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t

$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Задача максимизации логарифма правдоподобия:

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

при ограничениях нормировки и неотрицательности

$$\phi_{wt} \geq 0; \quad \sum_w \phi_{wt} = 1; \quad \theta_{td} \geq 0; \quad \sum_t \theta_{td} = 1$$

PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Теорема

Решение данной задачи удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} \equiv p(t|d, w)$:

$$\begin{array}{l} \text{E-шаг:} \\ \text{M-шаг:} \end{array} \left\{ \begin{array}{l} p_{tdw} = \mathop{\text{norm}}_{t \in T}(\phi_{wt} \theta_{td}) \\ \phi_{wt} = \mathop{\text{norm}}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} \right); \\ \theta_{td} = \mathop{\text{norm}}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} \right); \end{array} \right.$$

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

EM-алгоритм = метод простых итераций для системы уравнений

Тематическое моделирование на основе байесовского обучения

LDA и другие вероятностные модели порождения текста.

Байесовский вывод — нестандартная задача для каждой модели.

$$\begin{aligned}
 p(Z, W | \alpha, \beta) &= p(Z | W, \beta) p(Z | \alpha) \\
 p(W | Z, \beta) &= \int p(W | Z, \Phi) p(\Phi) d\Phi \\
 p(\Phi | \beta) &= \prod_{k=1}^K p(\phi_k | \beta) = \prod_{k=1}^K \frac{1}{B(\beta)} \prod_{i=1}^V \phi_{k,i}^{\beta_i - 1} \\
 p(W | Z, \Phi) &= \prod_{i=1}^N \theta_{i, \alpha_i} = \prod_{i=1}^N \prod_{k=1}^K \Phi_{k,i}^{w_i^{(k)}} \\
 \Phi(k, v) &= \sum_{i=1}^N I\{w_i = v \wedge z_i = k\} \\
 p(W | Z, \beta) &= \int \prod_{i=1}^N \frac{1}{B(\beta)} \prod_{k=1}^K \Phi_{k,i}^{w_i^{(k)} + \beta_k - 1} d\phi_k \\
 &= \int \prod_{k=1}^K f_k(\phi_k) \dots d\phi_K = \prod_{k=1}^K \int f_k(\phi_k) d\phi_k \\
 p(W | Z, \beta) &= \prod_{k=1}^K \left(\int \frac{1}{B(\beta)} \prod_{i=1}^V \Phi_{k,i}^{w_i^{(k)} + \beta_k - 1} d\phi_k \right) \\
 &= \prod_{k=1}^K \left(\frac{1}{B(\beta)} \int \prod_{i=1}^V \Phi_{k,i}^{w_i^{(k)} + \beta_k - 1} d\phi_k \right) \\
 p(W | Z, \beta) &= \prod_{k=1}^K \frac{B(\Psi_k + \beta)}{B(\beta)}
 \end{aligned}$$

$$\begin{aligned}
 p(\Theta | \alpha) &= \prod_{d=1}^D p(\theta_{d, \alpha_d}) = \prod_{d=1}^D \frac{1}{B(\alpha)} \prod_{k=1}^K \theta_{d,k}^{\alpha_k - 1} \\
 p(Z | \Theta) &= \prod_{d=1}^D \prod_{k=1}^K \theta_{d,k}^{\alpha_k} = \prod_{d=1}^D \prod_{k=1}^K \theta_{d,k}^{d(\alpha_k)} \\
 p(Z | \alpha) &= \int p(Z | \Theta) p(\Theta | \alpha) d\Theta \\
 &= \prod_{d=1}^D \left(\int \frac{1}{B(\alpha)} \prod_{k=1}^K \theta_{d,k}^{d(\alpha_k) + \alpha_k - 1} d\theta_k \right) \\
 &= \prod_{d=1}^D \frac{B(\Omega_d + \alpha)}{B(\alpha)}, \\
 \Omega(d, k) &= \sum_{i=1}^V I\{d_i = m \wedge z_i = k\} \\
 p(Z, W | \alpha, \beta) &= p(W | Z, \beta) p(Z | \alpha) \\
 &= \prod_{k=1}^K \frac{B(\Psi_k + \beta)}{B(\beta)} \cdot \prod_{d=1}^D \frac{B(\Omega_d + \alpha)}{B(\alpha)} \\
 p(z_i = k | Z_{-i}, W, \alpha, \beta) &= \frac{p(z_i = k, Z_{-i}, W | \alpha, \beta)}{p(Z_{-i}, W | \alpha, \beta)} \\
 p(z_i | Z_{-i}, W, \alpha, \beta) &= \frac{p(Z, W | \alpha, \beta)}{p(Z_{-i}, W | \alpha, \beta)} \\
 p(Z, W | \alpha, \beta) &= p(W | Z, \beta) p(Z | \alpha) \\
 &= \prod_{k=1}^K \frac{B(\Psi_k + \beta)}{B(\beta)} \cdot \prod_{d=1}^D \frac{B(\Omega_d + \alpha)}{B(\alpha)}
 \end{aligned}$$

$$\begin{aligned}
 \Psi^{-1}(k, v) &= \sum_{i=1}^V I\{w_i = v \wedge z_i = k\} \\
 \Omega^{-1}(d, k) &= \sum_{i=1}^V I\{d_i = d \wedge z_i = k\} \\
 \Phi(k, v) &= \begin{cases} \Psi^{-1}(k, v) + 1 & \text{if } v = w_i \text{ and } k = z_i \\ \Psi^{-1}(k, v) & \text{all other cases.} \end{cases} \\
 \Omega(d, k) &= \begin{cases} \Omega^{-1}(d, k) + 1 & \text{if } d = d_i \text{ and } k = z_i \\ \Omega^{-1}(d, k) & \text{all other cases.} \end{cases} \\
 \sum_{i=1}^V n(v; z_i) + 1 &= \sum_{i=1}^V n_{-i}(v; z_i) \\
 \sum_{k=1}^K n(z_i; d_k) + 1 &= \sum_{k=1}^K n_{-i}(z_i; d_k) \\
 p(z_i | Z_{-i}, W, \alpha, \beta) &= \frac{B(n(z_i; z_i) + \beta)}{B(n(z_i; z_i) + \beta)} \cdot \frac{B(n(z_i; m_1) + \alpha)}{B(n(z_i; (m_1) + \alpha)} \\
 &= \frac{\prod_{k=1}^K \Gamma(n(z_i; z_i) + \beta_k)}{\Gamma(\sum_{k=1}^K n(z_i; z_i) + \beta)} \cdot \frac{\prod_{k=1}^K \Gamma(n(z_i; m_k) + \alpha_k)}{\Gamma(\sum_{k=1}^K n(z_i; m_k) + \alpha)} \\
 &= \frac{\prod_{k=1}^K \Gamma(n_{-i}(z_i; z_i) + \beta_k)}{\Gamma(\sum_{k=1}^K n_{-i}(z_i; z_i) + \beta)} \cdot \frac{\prod_{k=1}^K \Gamma(n_{-i}(z_i; m_k) + \alpha_k)}{\Gamma(\sum_{k=1}^K n_{-i}(z_i; m_k) + \alpha)}
 \end{aligned}$$

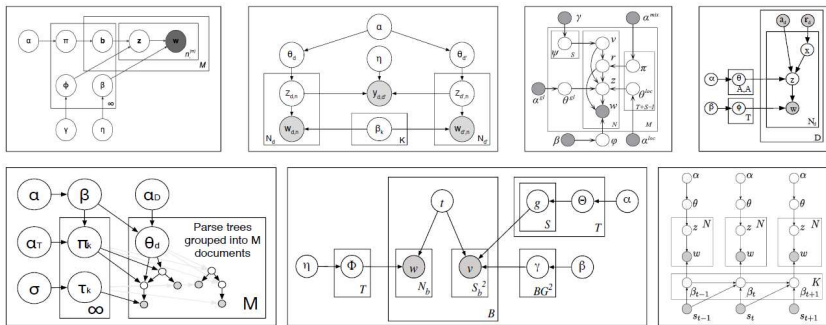
$$\begin{aligned}
 p(z_i | Z_{-i}, W, \alpha, \beta) &= \frac{n(z_i; z_i) + \beta_{z_i} - 1}{\sum_{k=1}^K n(z_i; z_i) + \beta} \cdot \frac{n(z_i; d_i) + \alpha_{z_i} - 1}{\sum_{k=1}^K n(z_i; d_i) + \alpha_{z_i} - 1} \\
 p(z_i | Z_{-i}, W, \alpha, \beta) &= \frac{n(z_i; z_i) + \beta_{z_i} - 1}{\sum_{k=1}^K n(z_i; z_i) + \beta} \cdot [n(z_i; d_i) + \alpha_{z_i} - 1] \\
 \phi_{k,i} &= p(w = i | z = k, W, Z, \beta) \\
 \theta_{d,k} &= p(z = k | Z, \alpha) \\
 \phi_{k,i} \cdot \theta_{d,k} &= p(w = i | z = k, W, Z, \beta) \cdot p(z = k | Z, \alpha) \\
 &= p(w = i, z = k | W, Z, \alpha, \beta) \\
 &= \frac{p(W, Z | \alpha, \beta)}{p(W, Z | \alpha, \beta)} \\
 \phi_{k,i} \cdot \theta_{d,k} &= \frac{\Gamma(n(z_i; i) + \beta_i)}{\Gamma(\sum_{k=1}^K n(z_i; i) + \beta)} \cdot \frac{\Gamma(n(z_i; d_i) + \alpha_{z_i})}{\Gamma(\sum_{k=1}^K n(z_i; d_i) + \alpha_{z_i})} \\
 &= \frac{\Gamma(n_{-i}(z_i; i) + \beta_i)}{\Gamma(\sum_{k=1}^K n_{-i}(z_i; i) + \beta)} \cdot \frac{\Gamma(n_{-i}(z_i; d_i) + \alpha_{z_i})}{\Gamma(\sum_{k=1}^K n_{-i}(z_i; d_i) + \alpha_{z_i})} \\
 \phi_{k,i} \cdot \theta_{d,k} &= \frac{n(z_i; i) + \beta_i}{(\sum_{k=1}^K n(z_i; i) + \beta)} \cdot \frac{n(z_i; d_i) + \alpha_{z_i}}{(\sum_{k=1}^K n(z_i; d_i) + \alpha_{z_i})} \\
 \phi_{k,i} &= \frac{n(z_i; i) + \beta_i}{(\sum_{k=1}^K n(z_i; i) + \beta)} \\
 \theta_{d,k} &= \frac{n(z_i; m) + \alpha_{z_i}}{(\sum_{k=1}^K n(z_i; m) + \alpha_{z_i})}
 \end{aligned}$$

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.

Yi Wang. Distributed Gibbs Sampling of Latent Dirichlet Allocation: The Gritty Details. 2008.

Тематическое моделирование на основе байесовского обучения

Графические модели упрощают понимание, но байесовский вывод остаётся нестандартной задачей для каждой модели.



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

Кризис байесовского обучения в тематическом моделировании

- сотни тематических моделей, начиная с LDA (Blei, 2003),
- создаются скорее ради теории, а не ради приложений,
- часто не имеют достаточных лингвистических обоснований,
- слишком сложны для понимания, вывода, сравнения,
- реализуют требуемые функции лишь по-отдельности,
- не комбинируются и взаимно не заменяются,
- не имеют полнофункциональных библиотек в открытом коде,
- что создаёт барьеры вхождения для прикладников,
- которые предпочитают устаревшие но понятные PLSA и LDA

ARTM — аддитивная регуляризация тематической модели

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

Метод многокритериальной оптимизации — скаляризация.

Задача максимизации регуляризованного правдоподобия:

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

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.

EM-алгоритм с регуляризацией M-шага

Теорема

Решение данной задачи удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t|d, w)$:

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

EM-алгоритм = метод простых итераций для системы уравнений

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

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

ARTM: зоопарк регуляризаторов

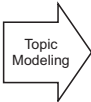
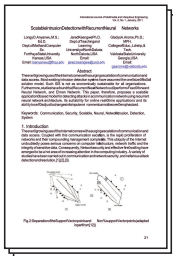
- разреживание и декоррелирование предметных тем
- сглаживание фоновых тем общей лексики (LDA)
- энтропийное разреживание для отбора тем
- сглаживание и разреживание тем во времени
- выявление иерархических связей между темами
- многоязычное тематическое моделирование
- выявление внутренней тематической структуры текста
- обучение с учителем для классификации и регрессии
- частичное (semi-supervised) обучение
- и др.

Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning. Special Issue "Data Analysis and Intelligent Optimization with Applications". Springer, 2015.

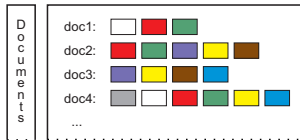
Мультиязычная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$, ...

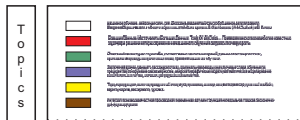
Text documents



Topics of documents



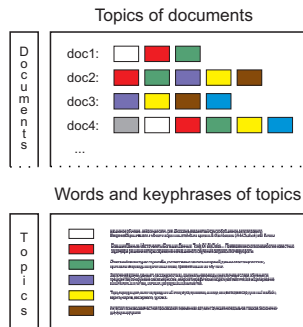
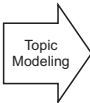
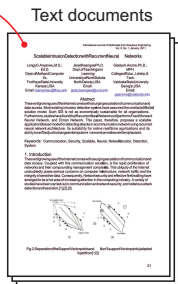
Words and keyphrases of topics



Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$,
 авторов $p(t|a)$, времени $p(t|t)$,...

Metadata:
 Authors
 Data Time
 Conference
 Organization
 URL
 etc.



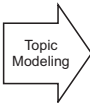
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$,
 авторов $p(t|a)$, времени $p(t|t)$, элементов изображений $p(t|e)$,...

Metadata:
 Authors
 Data Time
 Conference
 Organization
 URL
 etc.

Text documents

Images



Topics of documents

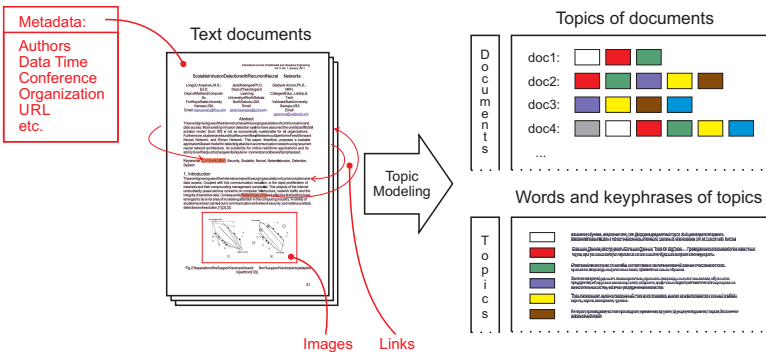
| | | | | | | | |
|---|-------|--------|--------|--------|--------|--------|------|
| D o c u m e n t s | doc1: | White | Red | Green | | | |
| | doc2: | Red | Green | Purple | Yellow | Brown | |
| | doc3: | Purple | Yellow | Brown | Blue | | |
| | doc4: | Grey | White | Red | Green | Yellow | Blue |
| | ... | | | | | | |

Words and keyphrases of topics

| | | |
|----------------------------|--------|-------------|
| T o p i c s | White | Тема 1: ... |
| | Red | Тема 2: ... |
| | Green | Тема 3: ... |
| | Purple | Тема 4: ... |
| | Yellow | Тема 5: ... |

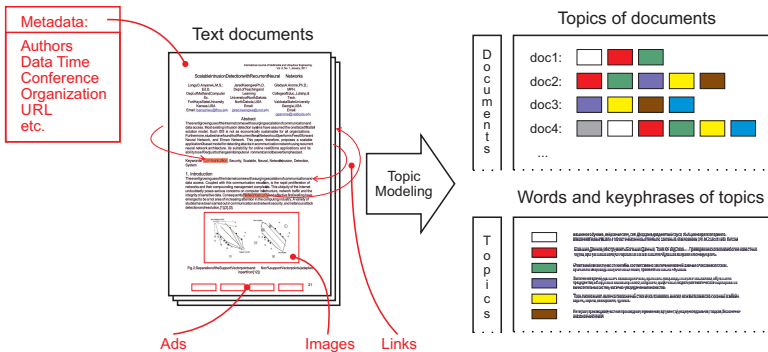
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$,
 авторов $p(t|a)$, времени $p(t|t)$, элементов изображений $p(t|e)$,
 ссылок $p(d'|r)$, ...



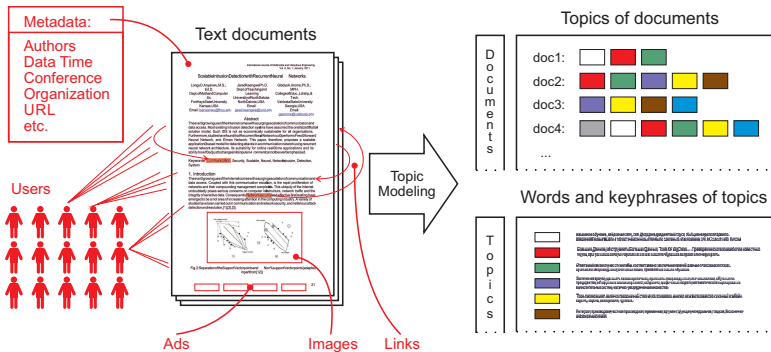
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$,
 авторов $p(t|a)$, времени $p(t|t)$, элементов изображений $p(t|e)$,
 ссылок $p(d'|r)$, **баннеров** $p(t|b)$,...



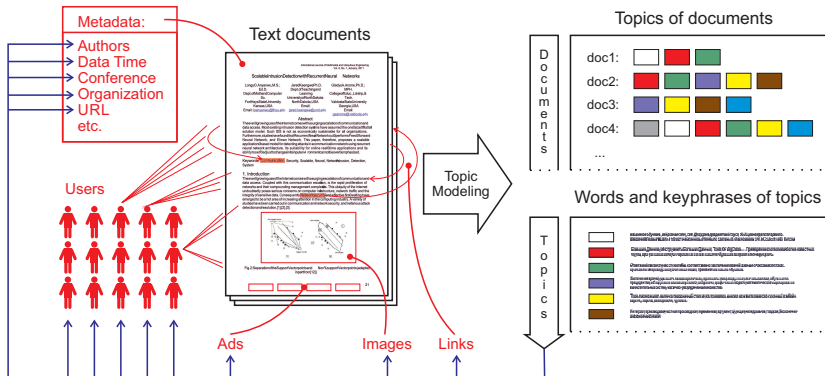
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$, авторов $p(t|a)$, времени $p(t|t)$, элементов изображений $p(t|e)$, ссылок $p(d'|r)$, баннеров $p(t|b)$, **пользователей $p(t|u)$, ...**



Мультимодальная тематическая модель

Каждая модальность $t \in M$ описывается своим словарём W^m , документы могут содержать токены разных модальностей, каждая тема имеет своё распределение $p(w|t)$, $w \in W^m$



Мультимодальная ARTM

Каждая модальность $m \in M$ описывается своим словарём W^m , документы могут содержать токены разных модальностей, каждая тема имеет своё распределение $p(w|t)$, $w \in W^m$

Задача максимизации регуляризованного правдоподобия:

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

при ограничениях неотрицательности и нормировки

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

где $\tau_m > 0$, $\tau_i > 0$ — коэффициенты регуляризации.

EM-алгоритм для мультимодальной ARTM

Теорема

Решение данной задачи удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t|d, w)$:

$$p_{tdw} = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td});$$

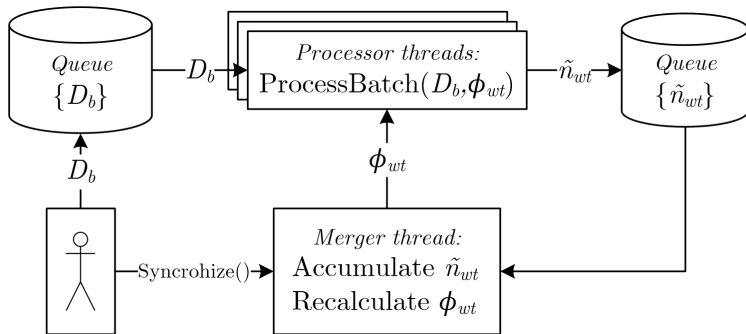
$$\phi_{wt} = \operatorname{norm}_{w \in W^m} \left(\sum_{d \in D} \tau_{m(w)} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right);$$

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

где $m(w)$ — модальность токена w .

EM-алгоритм = метод простых итераций для системы уравнений

Параллельная архитектура



- коллекция разбивается на пакеты $D = D_1 \sqcup \dots \sqcup D_B$
- простой однопоточный `ProcessBatch`
- пользователь определяет моменты обновлений модели
- гарантируется воспроизводимость от запуска к запуску

Онлайновый параллельный EM-алгоритм для ARTM

Вход: коллекция D_b , коэффициент дисконтирования $\rho \in (0, 1]$;

Выход: матрица Φ ;

- 1 инициализировать ϕ_{wt} для всех $w \in W$, $t \in T$;
- 2 $n_{wt} := 0$, $\tilde{n}_{wt} := 0$ для всех $w \in W$, $t \in T$;
- 3 **для всех** пакетов D_b , $b = 1, \dots, B$
- 4 $(\tilde{n}_{wt}) := (\tilde{n}_{wt}) + \mathbf{ProcessBatch}(D_b, \Phi)$;
- 5 **если** пора выполнить синхронизацию, **то**
- 6 $n_{wt} := \rho n_{wt} + \tilde{n}_{wt}$ для всех $w \in W$, $t \in T$;
- 7 $\phi_{wt} := \mathop{\text{norm}}_{w \in W^m} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ для всех $w \in W$, $t \in T$;
- 8 $\tilde{n}_{wt} := 0$ для всех $w \in W$, $t \in T$;

Онлайновый параллельный EM-алгоритм для ARTM

ProcessBatch обрабатывает пакет D_b при фиксированной Φ .

Вход: пакет D_b , матрица $\Phi = (\phi_{wt})$;

Выход: матрица (\tilde{n}_{wt}) ;

1 $\tilde{n}_{wt} := 0$ для всех $w \in W$, $t \in T$;

2 **для всех** $d \in D_b$

3 инициализировать $\theta_{td} := \frac{1}{|T|}$ для всех $t \in T$;

4 **повторять**

5 $p_{tdw} := \mathop{\text{norm}}_{t \in T}(\phi_{wt}\theta_{td})$ для всех $w \in d$, $t \in T$;

6 $\theta_{td} := \mathop{\text{norm}}_{t \in T} \left(\sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$ для всех $t \in T$;

7 **пока** θ_d не сойдётся;

8 $\tilde{n}_{wt} := \tilde{n}_{wt} + \tau_{m(w)} n_{dw} p_{tdw}$ для всех $w \in d$, $t \in T$;

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайн-овая параллельная мультимодальная ARTM
- Большие данные: коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов

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

От теории ARTM к технологии BigARTM


Разработка тематической модели с заданными свойствами:


Этапы моделирования

Bayesian TM

ARTM

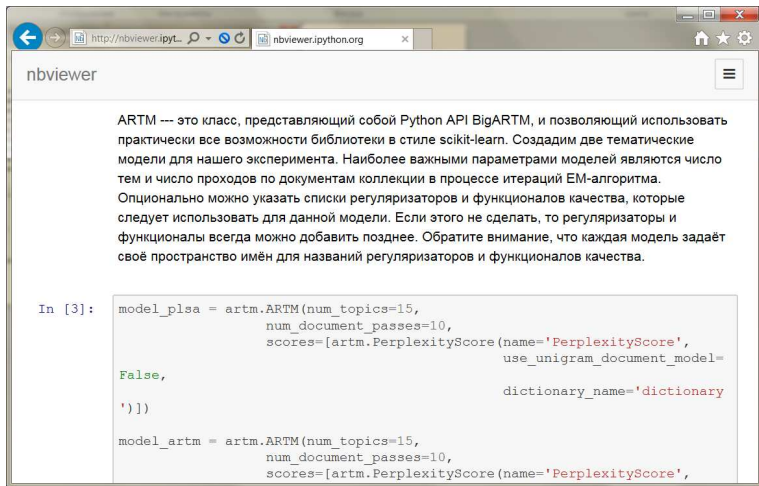
| | | | |
|-----------------|--|--|---------------|
| | Анализ требований | Анализ требований | |
| Формализация: | Вероятностная порождающая модель данных | Стандартные критерии | Свои критерии |
| Алгоритмизация: | Байесовский вывод для данной порождающей модели (VI, GS, EP) | Общий регуляризованный 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 | |

 -- нестандартизируемые этапы, уникальная разработка для каждой задачи

 -- стандартизуемые этапы

Разработка тематических моделей в среде IPython Notebook

http://nbviewer.ipython.org/github/bigartm/bigartm-book/blob/master/BigARTM_example_RU.ipynb



The screenshot shows a web browser window with the URL `http://nbviewer.ipython.org`. The page content includes a paragraph of Russian text explaining the ARTM class and its parameters, followed by a code cell. The code cell contains two Python snippets for creating ARTM models with specific regularization parameters.

ARTM --- это класс, представляющий собой Python API BigARTM, и позволяющий использовать практически все возможности библиотеки в стиле scikit-learn. Создадим две тематические модели для нашего эксперимента. Наиболее важными параметрами моделей являются число тем и число проходов по документам коллекции в процессе итераций EM-алгоритма. Опционально можно указать списки регуляризаторов и функционалов качества, которые следует использовать для данной модели. Если этого не сделать, то регуляризаторы и функционалы всегда можно добавить позднее. Обратите внимание, что каждая модель задаёт своё пространство имён для названий регуляризаторов и функционалов качества.

```
In [3]: model_plsa = artm.ARTM(num_topics=15,
                             num_document_passes=10,
                             scores=[artm.PerplexityScore(name='PerplexityScore',
                                                           use_unigram_document_model=
False,
                                                           dictionary_name='dictionary'
                             ')]])

model_artm = artm.ARTM(num_topics=15,
                       num_document_passes=10,
                       scores=[artm.PerplexityScore(name='PerplexityScore',
```

Эксперимент 1. Обгоняем конкурентов по скорости

- 3.7M статей английской Вики, 100K уникальных слов

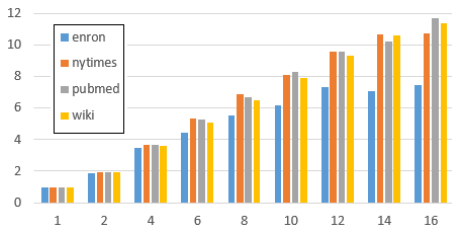
| | procs | train | inference | perplexity |
|---------------------|-------|---------|-----------|------------|
| BigARTM | 1 | 35 min | 72 sec | 4000 |
| Gensim.LdaModel | 1 | 369 min | 395 sec | 4161 |
| VowpalWabbit.LDA | 1 | 73 min | 120 sec | 4108 |
| BigARTM | 4 | 9 min | 20 sec | 4061 |
| Gensim.LdaMulticore | 4 | 60 min | 222 sec | 4111 |
| BigARTM | 8 | 4.5 min | 14 sec | 4304 |
| Gensim.LdaMulticore | 8 | 57 min | 224 sec | 4455 |

- *procs* = число параллельных потоков
- *inference* = время тематизации 100K тестовых документов
- *perplexity* вычислена на тестовой выборке документов

Эксперимент 2. Масштабируемость по числу потоков

| коллекция | $ W , 10^3$ | $ D , 10^6$ | $n, 10^6$ | размер, Гб |
|-----------|-------------|-------------|-----------|------------|
| enron | 28 | 0.04 | 6.4 | 0.07 |
| nytimes | 103 | 0.3 | 100 | 0.13 |
| pubmed | 141 | 8.2 | 738 | 1.0 |
| wiki | 100 | 3.7 | 1009 | 1.2 |

ускорение



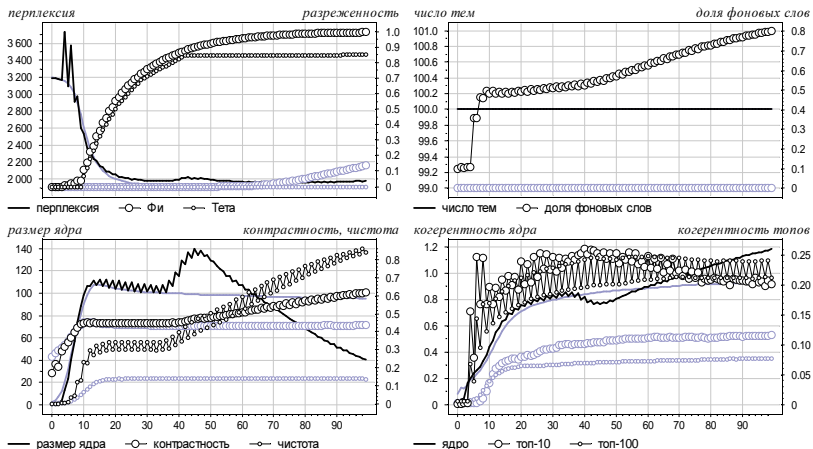
число ядер

Amazon EC2 cc2.8xlarge instance:

16 cores + hyperthreading, Intel[®] Xeon[®] CPU E5-2670 2.6GHz.

Эксперимент 3. Комбинирование регуляризаторов

Сравнение PLSA (серый) и ARTM со сглаживанием, разреживанием и декоррелированием (чёрный)



Эксперимент 4. Мультязычная модель

Модальности — это разные языки.

216 175 русско-английских пар статей Вики.

Первые 10 слов и их вероятностями $p(w|t)$ в %:

| Topic 68 | | | | Topic 79 | | | |
|-------------|------|--------------|------|----------|------|-----------|------|
| research | 4.56 | институт | 6.03 | goals | 4.48 | матч | 6.02 |
| technology | 3.14 | университет | 3.35 | league | 3.99 | игрок | 5.56 |
| engineering | 2.63 | программа | 3.17 | club | 3.76 | сборная | 4.51 |
| institute | 2.37 | учебный | 2.75 | season | 3.49 | фк | 3.25 |
| science | 1.97 | технический | 2.70 | scored | 2.72 | против | 3.20 |
| program | 1.60 | технология | 2.30 | cup | 2.57 | клуб | 3.14 |
| education | 1.44 | научный | 1.76 | goal | 2.48 | футболист | 2.67 |
| campus | 1.43 | исследование | 1.67 | apps | 1.74 | гол | 2.65 |
| management | 1.38 | наука | 1.64 | debut | 1.69 | забивать | 2.53 |
| programs | 1.36 | образование | 1.47 | match | 1.67 | команда | 2.14 |

Эксперимент 4. Мультиязычная модель

216 175 русско-английских пар статей Вики.

Первые 10 слов и их вероятностями $p(w|t)$ в %:

| Topic 88 | | | | Topic 251 | | | |
|-------------|------|---------|------|------------|------|--------------|------|
| opera | 7.36 | опера | 7.82 | windows | 8.00 | windows | 6.05 |
| conductor | 1.69 | оперный | 3.13 | microsoft | 4.03 | microsoft | 3.76 |
| orchestra | 1.14 | дирижер | 2.82 | server | 2.93 | версия | 1.86 |
| wagner | 0.97 | певец | 1.65 | software | 1.38 | приложение | 1.86 |
| soprano | 0.78 | певица | 1.51 | user | 1.03 | сервер | 1.63 |
| performance | 0.78 | театр | 1.14 | security | 0.92 | server | 1.54 |
| mozart | 0.74 | партия | 1.05 | mitchell | 0.82 | программный | 1.08 |
| sang | 0.70 | сопрано | 0.97 | oracle | 0.82 | пользователь | 1.04 |
| singing | 0.69 | вагнер | 0.90 | enterprise | 0.78 | обеспечение | 1.02 |
| operas | 0.68 | оркестр | 0.82 | users | 0.78 | система | 0.96 |

Независимый ассессор оценил 396 тем из $|T| = 400$ как хорошо интерпретируемые.

Эксперимент 5. Интерпретируемость мультиграммной модели

Две модальности — униграммы и биграммы.

Коллекция 1000 статей конференций ММРО, ИОИ на русском

| распознавание образов в биоинформатике | | теория вычислительной сложности | |
|--|-------------------------|---------------------------------|----------------------|
| unigrams | bigrams | unigrams | bigrams |
| объект | задача распознавания | задача | разделять множества |
| задача | множество мотивов | множество | конечное множество |
| множество | система масок | подмножество | условие задачи |
| мотив | вторичная структура | условие | задача о покрытии |
| разрешимость | структура белка | класс | покрытие множества |
| выборка | распознавание вторичной | решение | сильный смысл |
| маска | состояние объекта | конечный | разделяющий комитет |
| распознавание | обучающая выборка | число | минимальный аффинный |
| информативность | оценка информативности | аффинный | аффинный комитет |
| состояние | множество объектов | случай | аффинный разделяющий |
| закономерность | разрешимость задачи | покрытие | общее положение |
| система | критерий разрешимости | общий | множество точек |
| структура | информативность мотива | пространство | случай задачи |
| значение | первичная структура | схема | общий случай |
| регулярность | тупиковое множество | комитет | задача MASC |

Мотивации

Выборка может содержать не только пары (d, w) , но также тройки, \dots , n -ки элементов разных модальностей.

Примеры:

- **Данные социальной сети:**
 (d, u, w) — в блоге d пользователь u записал слово w
- **Данные сети интернет-рекламы:**
 (u, d, b) — пользователь u кликнул рекламное объявление b на веб-странице d
- **Данные рекомендательной системы:**
 (u, f, s) — пользователь u оценил фильм f в ситуативном контексте s

Хотим объяснить наблюдаемую выборку рёбер гиперграфа латентными тематическими профилями его вершин.

Тематическая модель гиперграфа: определения и обозначения

$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.

V^m — множество вершин модальности m

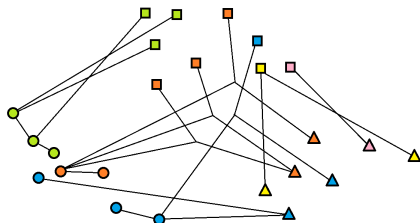
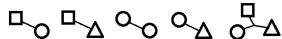
M — множество модальностей:



T — множество тем:



K — множество типов рёбер:



Каждое ребро $e = (d, x = (v_1, \dots, v_h))$ типа k имеет:

d — вершину-контейнер модальности m_{0k} ,

v_j — вершины модальностей m_{jk} .

X^k — наблюдаемая выборка рёбер-транзакций типа k

n_{dx} — число вхождений ребра $e = (d, x)$ в выборку X^k

$p_k(d, x)$ — неизвестное распределение на рёбрах типа k

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k :

$$p_k(x|d) = \sum_{t \in T} \theta_{td} \prod_{v \in X} \phi_{kvt},$$

$\theta_{td} = p(t|d)$ — тематика контейнера не зависит от типа ребра k
 $\phi_{kvt} = p_k(v|t)$ — распределение модальности V^m в теме t
на рёбрах типа k

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in X^k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in X} \phi_{kvt} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta},$$
$$\phi_{kvt} \geq 0, \quad \sum_{v \in V^m} \phi_{kvt} = 1, \quad k \in K; \quad \theta_{td} \geq 0, \quad \sum_{t \in T} \theta_{td} = 1,$$

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Теорема

Точка максимума удовлетворяет системе уравнений со вспомогательными переменными $p_{tdx} = p(t|d, x)$:

$$p_{tdx} = \operatorname{norm}_{t \in T} \left(\theta_{td} \prod_{v \in X} \phi_{kvt} \right);$$

$$\phi_{kvt} = \operatorname{norm}_{v \in V^m} \left(\sum_{(d,x) \in X^k} [v \in X] \tau_k n_{dx} p_{tdx} + \phi_{kvt} \frac{\partial R}{\partial \phi_{kvt}} \right);$$

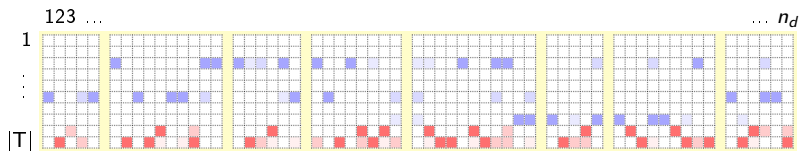
$$\theta_{td} = \operatorname{norm}_{t \in T} \left(\sum_{k \in K} \sum_{x: (d,x) \in X^k} \tau_k n_{dx} p_{tdx} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right);$$

EM-алгоритм = метод простых итераций для системы уравнений

Тематическое моделирование последовательного текста

Документ $d = \{w_1, \dots, w_{n_d}\}$, n_d — длина документа d

Матрица тематических профилей слов $p(t|d, w_i)$ размера $T \times n_d$:



Предположения разреженности и непрерывности тематики:

- каждое предложение относится к 1–2 предметным темам
- соседние предложения часто имеют одинаковые темы
- слова общей лексики не влияют на тематику предложений
- между абзацами вероятность смены темы выше
- между секциями она ещё выше

EM-алгоритм с регуляризацией E-шага

Теорема

Если регуляризатор зависит от Φ, Θ через $p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$,

$$\mathcal{L}(\Phi, \Theta) + \sum_{d \in D} \sum_{i=1}^{n_d} R_{di}(p_{1dw_i}, \dots, p_{Tdw_i}) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

то точка максимума удовлетворяет системе уравнений

$$\tilde{p}_{tdw} = p_{tdw} \frac{1}{n_{dw}} \sum_{\substack{i=1 \\ w_i=w}}^{n_d} \left(1 + \frac{\partial R_{di}}{\partial p_{tdw}} - \sum_{s \in T} p_{sdw} \frac{\partial R_{di}}{\partial p_{sdw}} \right);$$

$$\phi_{wt} = \text{norm}_w \left(\sum_{d \in D} n_{dw} \tilde{p}_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right);$$

$$\theta_{td} = \text{norm}_t \left(\sum_{w \in D} n_{dw} \tilde{p}_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right).$$

Регуляризатор разреживания распределений $p(t|d, w)$

Гипотеза разреженности распределений $p_{tdw} = p(t|d, w)$:
в документе слово может относиться только к одной теме.

Максимизируем KL-дивергенции между $\hat{p}(t) = \frac{1}{|T|}$ и $p(t|d, w)$:

$$R(\Phi, \Theta) = -\tau \sum_{d \in D} \sum_{w \in d} n_{dw} \frac{1}{|T|} \sum_{t \in T} \ln p_{tdw}.$$

Подставляем, получаем формулу модифицированного E-шага:

$$\tilde{p}_{tdw} = \frac{\phi_{wt} \theta_{td}}{\sum_{s \in T} \phi_{ws} \theta_{sd}} (1 + \tau) - \frac{\tau}{|T|}.$$

Эффект:

если $p(t|w) < \frac{1}{|T|}$, то ϕ_{wt} уменьшается;

если $p(t|d) < \frac{1}{|T|}$, то θ_{td} уменьшается.

Регуляризатор сглаживания распределений $p(t|d, w)$ по контексту

Контекст слова w_i — множество слов w_j недалеко от слова w_i
 \hat{p}_{tdi} — эмпирическая оценка $p_{tdw_i} = p(t|d, w_i)$ по контексту,

$$\hat{p}_{tdi} = \sum_j K_{ij} p_{tdw_j},$$

где K_{ij} — оценка важности слова w_j в контексте w_i .

Минимизируем KL-дивергенции между \hat{p}_{tdi} и p_{tdw_i} :

$$R(\Phi, \Theta) = \sum_{d \in D} \sum_{i=1}^{n_d} \hat{p}_{tdi} \ln p_{tdw_i}.$$

Подставляем, получаем формулу модифицированного E-шага:

$$\tilde{p}_{tdw} = \frac{\phi_{wt} \theta_{td}}{\sum_{s \in T} \phi_{ws} \theta_{sd}} (1 - \tau) + \tau \hat{p}_{tdi}.$$

Выводы

- Для создания систем разведочного информационного поиска нужны многофункциональные тематические модели
- Разработка и комбинирование сложных моделей сильно затруднены в рамках байесовского подхода
- **Аддитивная регуляризация тематических моделей (ARTM)** радикально упрощает реализацию, понимание, комбинирование, сравнение тематических моделей
- ARTM допускает дальнейшие обобщения: мультимодальные, мультязычные, иерархические, динамические, гиперграфовые, лингвистические модели.

Направления дальнейших исследований

- научиться строить 50 тысяч хорошо интерпретируемых тем
- научиться автоматически создавать и именовать темы
- соединить лингвистическую регуляризацию и word2vec
- применять гиперграфовые модели к данным соцсетей
- разработать визуальные средства систематизации знаний
- создать систему тематического разведочного поиска



<http://bigartm.org>

Join BigARTM community!