Введение в вероятностное тематическое моделирование

Воронцов Константин Вячеславович

ФИЦ ИУ РАН • МФТИ • МГУ • Яндекс • Форексис

ЛИНИС НИУ ВШЭ • 23 октября 2015

Содержание

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.

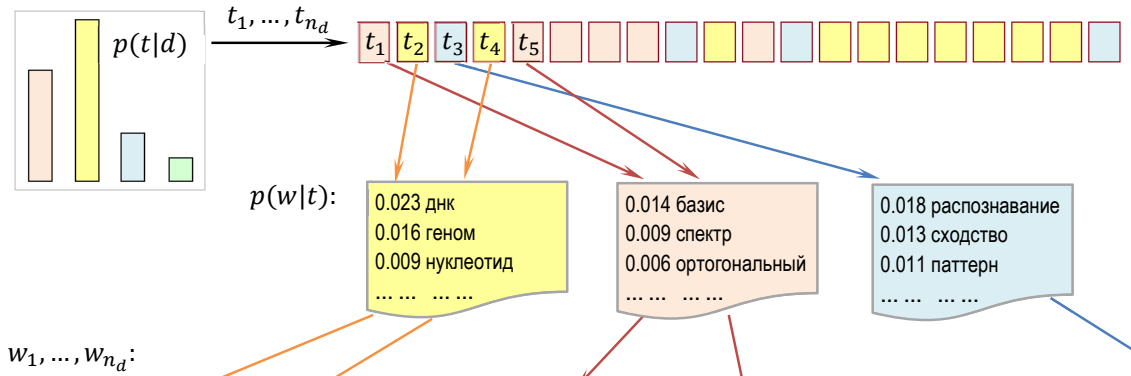
Когда автор писал термин w в документе d, он думал о теме t, и мы хотели бы выявить, о какой именно.

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:

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



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 встретился в документе d

n_d — длина документа d

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

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

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

Эта задача стохастического матричного разложения является *некорректно поставленной*, т. к. её решение не единственно:

$$\underset{W\times D}{\left(\frac{n_{dw}}{n_d}\right)} \approx \underset{W\times T}{\Phi} \cdot \underset{T\times D}{\Theta} = (\Phi S)(S^{-1}\Theta) = \underset{W\times T}{\Phi'} \cdot \underset{T\times D}{\Theta'}$$

для невырожденных $\underset{T\times T}{S}$ таких, что Φ', Θ' тоже стохастические.

PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Задача максимизации логарифма правдоподобия:

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

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdw} = p(t|d,w)$:

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

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.

EM-алгоритм. Элементарная интерпретация

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: частотные оценки условных вероятностей вычисляются путём суммирования счётчика $n_{dwt} = n_{dw}p(t|d,w)$:

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

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

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

Максимизация апостериорной вероятности (Dirichlet prior):

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td}}_{\text{log правдоподобия } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}}_{\text{критерий регуляризации } R(\Phi,\Theta)} \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

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

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 тематических моделей

Максимизация log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014. Т. 455., № 3. 268–271.

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.

Комбинирование регуляризованных тематических моделей

Максимизация log правдоподобия с n регуляризаторами R_i:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta) \to \max_{\Phi,\Theta},$$

где τ_i — коэффициенты регуляризации.

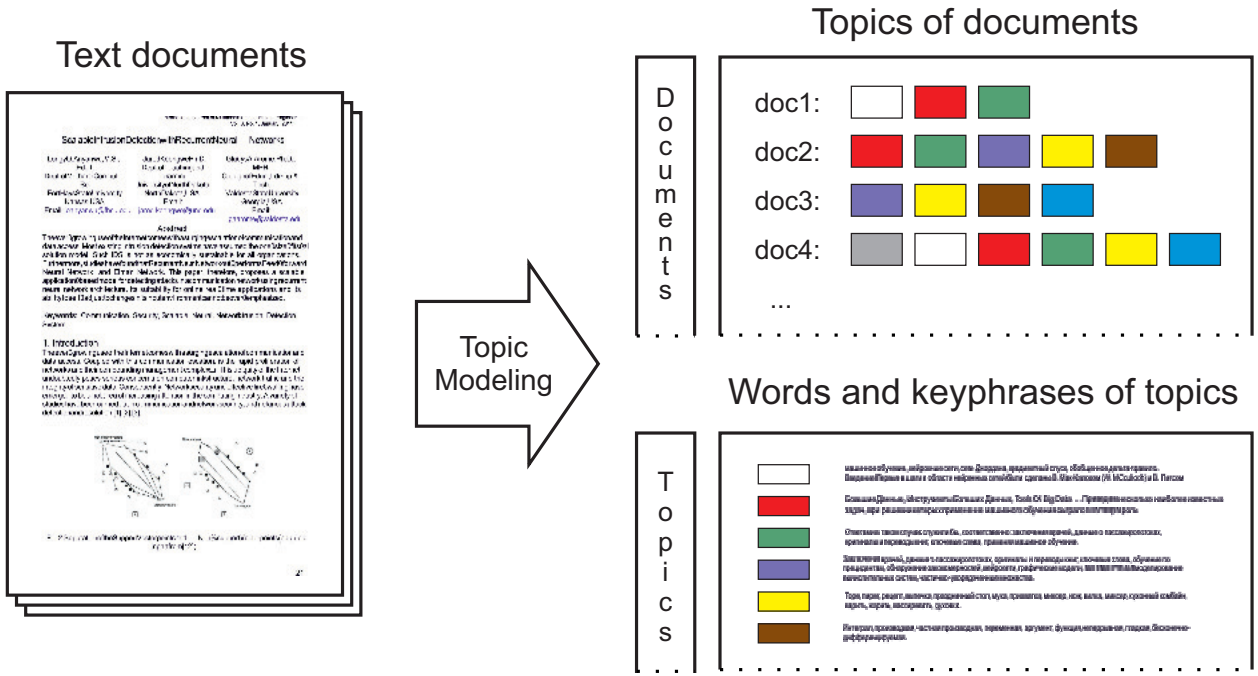
EM-алгоритм: метод простой итерации для системы уравнений

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

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw} p_{tdw} + \phi_{wt} \sum_{i=1}^{n} \tau_i \frac{\partial R_i}{\partial \phi_{wt}} \right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw} p_{tdw} + \theta_{td} \sum_{i=1}^{n} \tau_i \frac{\partial R_i}{\partial \theta_{td}} \right) \end{cases}$$

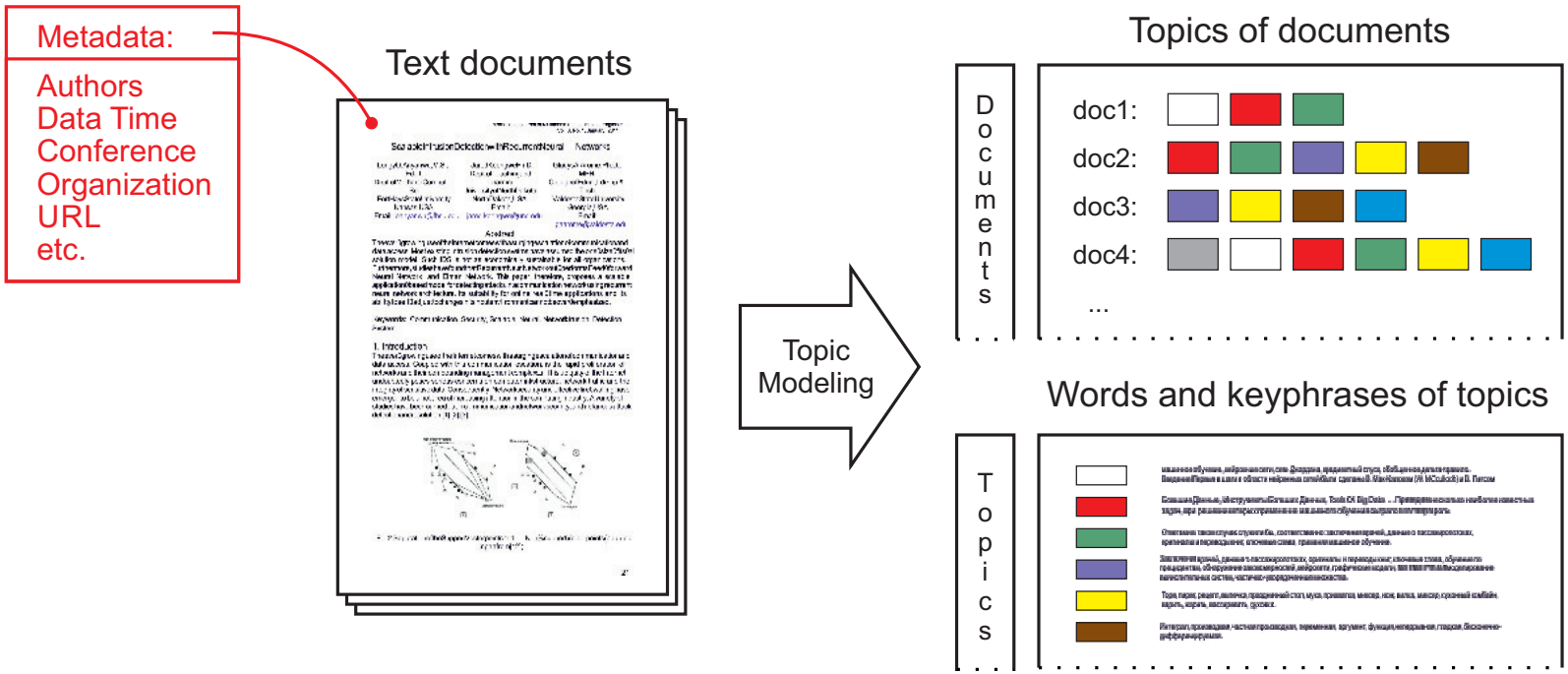
Мультимодальная тематическая модель

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



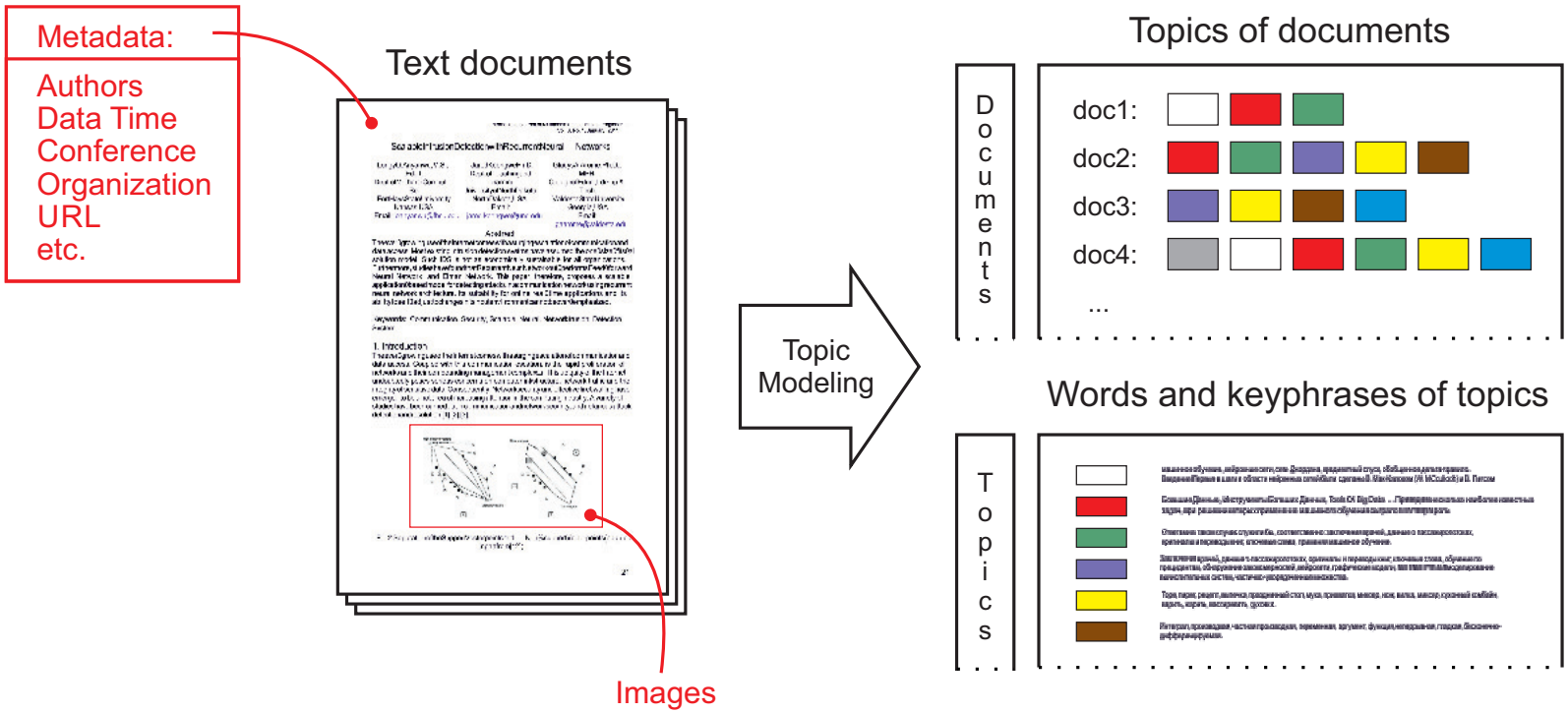
Мультимодальная тематическая модель

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



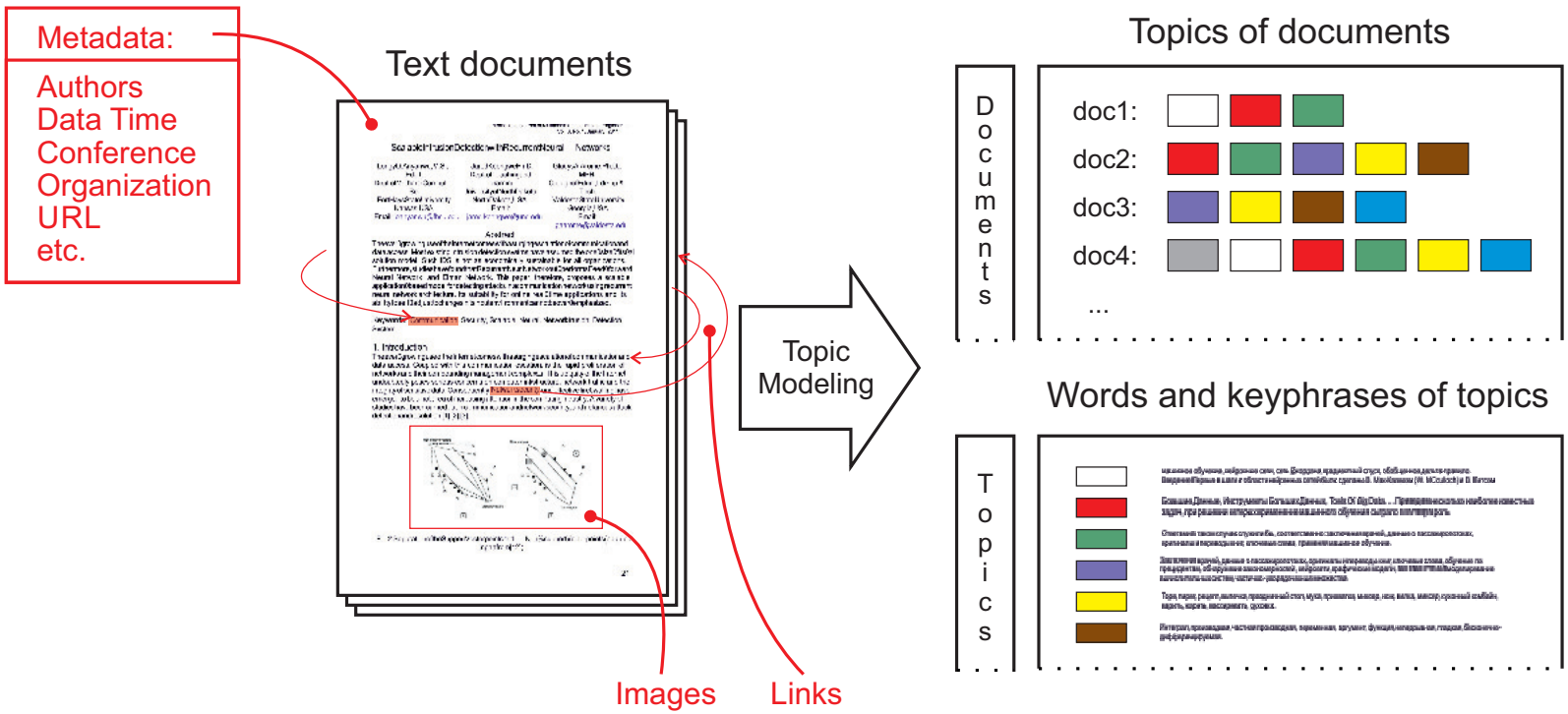
Мультимодальная тематическая модель

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



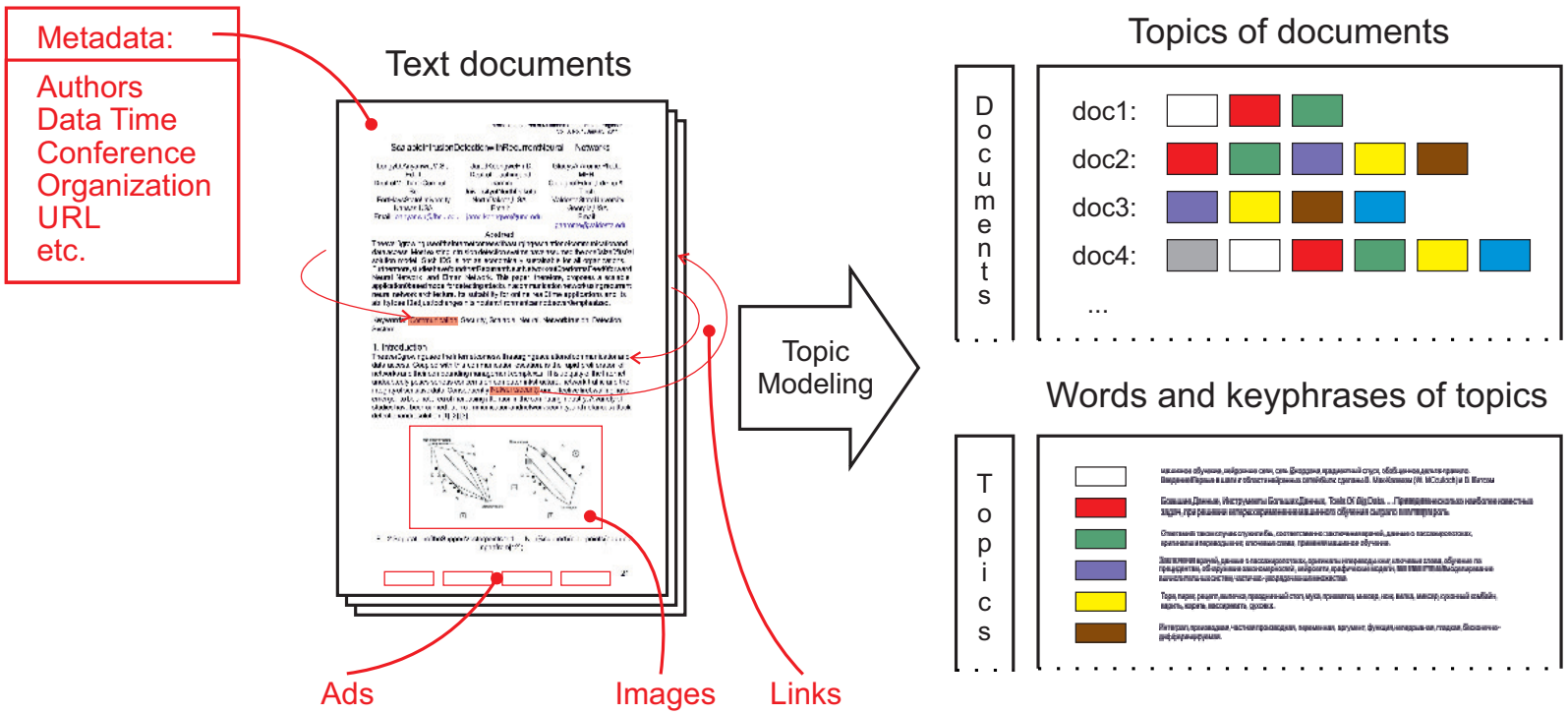
Мультимодальная тематическая модель

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



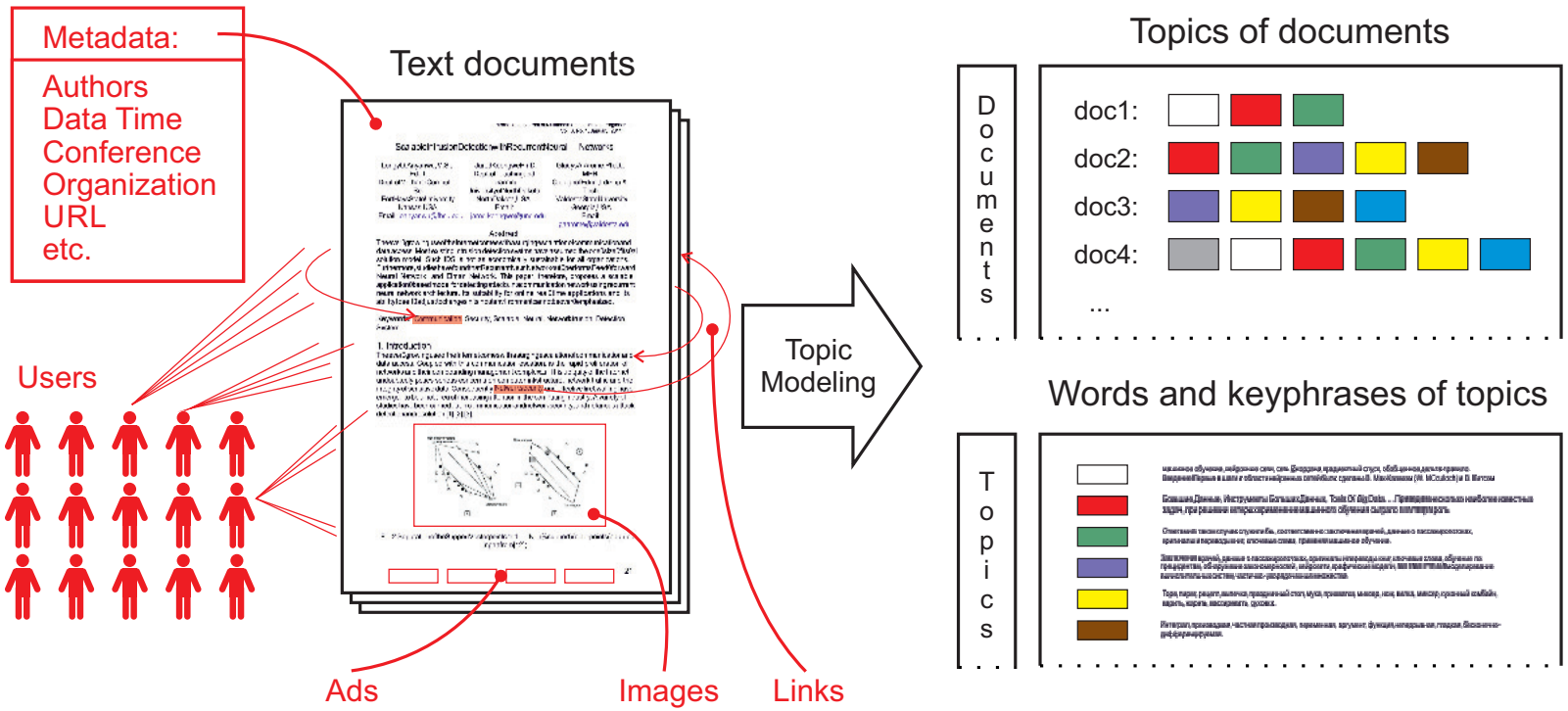
Мультимодальная тематическая модель

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



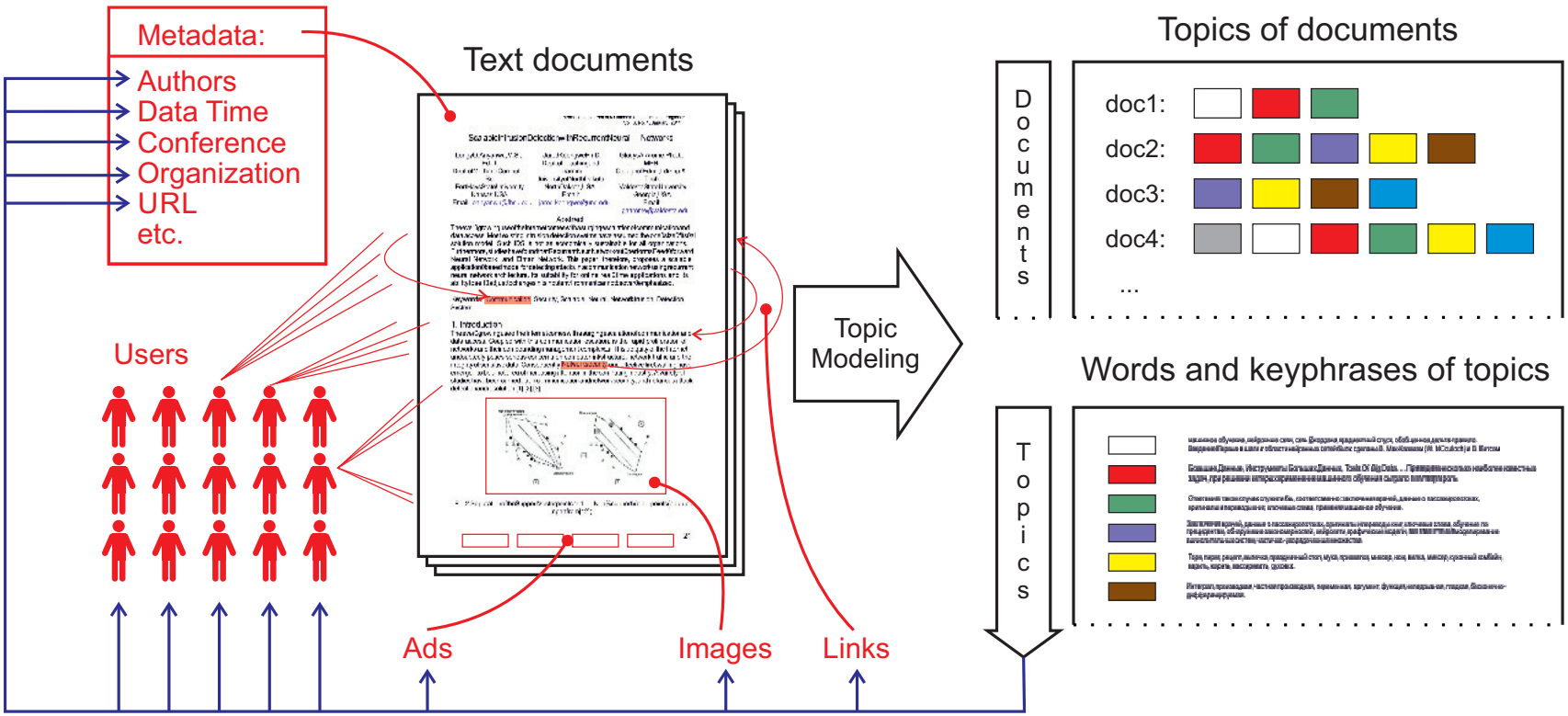
Мультимодальная тематическая модель

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



Мультимодальная тематическая модель

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



Мультимодальная ARTM [Vorontsov et al, 2015]

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

$W = W^1 \sqcup \cdots \sqcup W^M$ — объединённый словарь всех модальностей

Максимизация суммы log правдоподобий с регуляризацией:

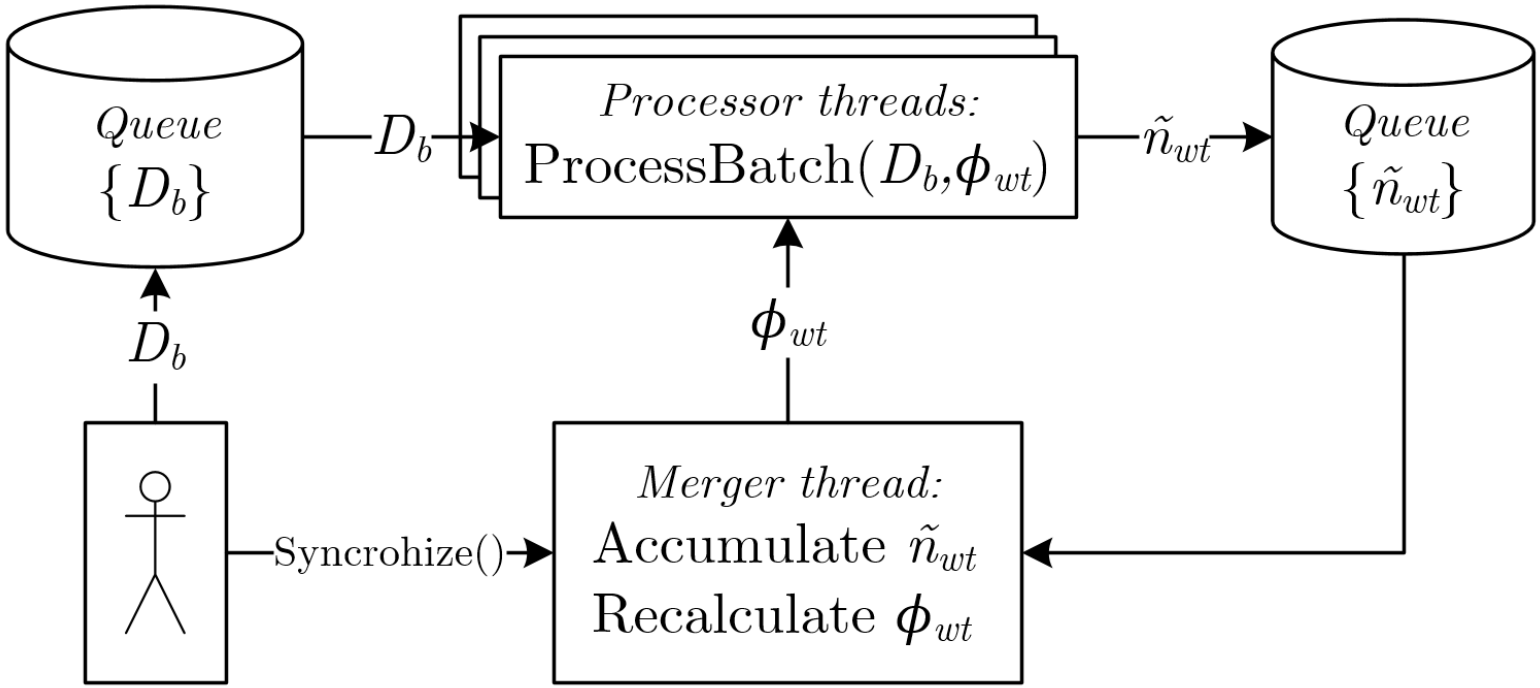
$$\sum_{m\in M} \lambda_m \sum_{d\in D} \sum_{w\in W^m} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Параллельная архитектура



- коллекция разбивается на пакеты $D = D_1 \sqcup \cdots \sqcup D_B$
- 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 $D_b,\ b = 1, \ldots, B$
4 $\quad (\tilde{n}_{wt}) := (\tilde{n}_{wt}) + \mathbf{ProcessBatch}(D_b, \Phi)$;
5 $\quad$ **если** пора выполнить синхронизацию, **то**
6 $\quad\quad n_{wt} := \rho n_{wt} + \tilde{n}_{wt}$ для всех $w \in W,\ t \in T$;
7 $\quad\quad \phi_{wt} := \underset{w \in W^m}{\operatorname{norm}} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ для всех $w \in W,\ t \in T$;
8 $\quad\quad \tilde{n}_{wt} := 0$ 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 $(\tilde{n}_{wt})$;

$\tilde{n}_{wt} := 0$ для всех $w \in W,\ t \in T$;
для всех $d \in D_b$
  инициализировать $\theta_{td} := \frac{1}{|T|}$ для всех $t \in T$;
  повторять
    $p_{tdw} := \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$ для всех $w \in d,\ t \in T$;
    $\theta_{td} := \underset{t \in T}{\text{norm}}\left(\sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$ для всех $t \in T$;
  пока θ_d не сойдётся;
  $\tilde{n}_{wt} := \tilde{n}_{wt} + \tau_{m(w)} n_{dw} p_{tdw}$ для всех $w \in d,\ t \in T$;

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновая 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	

-- *нестандартизуемые этапы, уникальная разработка для каждой задачи*

-- *стандартизуемые этапы*

Разработка тематических моделей в среде IPython Notebook

http://nbviewer.ipython.org/github/bigartm/bigartm-book/tree/master/

Коллекция:

Воспользуемся небольшой коллекцией 'kos', доступной в репозитории UCI https://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/. Параметры коллекции следующие:

- 3430 документов;
- 6906 слов в словаре;
- 467714 слов в коллекции.

Для начала подключим все необходимые модули (убедитесь, что путь к Python API BigARTM находится в вашей переменной PATH):

In [1]:
```
%matplotlib inline
import glob
import matplotlib.pyplot as plt

import artm
```

Прежде всего необходимо подготовить входные данные. BigARTM имеет собственный формат документов для обработки, называемый батчами. В библиотеки присутствуют средства по созданию батчей из файлов Bag-Of-Words в форматах UCI и Vowpal Wabbit (подробности можно найти в http://docs.bigartm.org/en/latest/formats.html).

В Python API, по аналогии с алгоритмами из scikit-learn, входные данные представлены одним классом BatchVectorizer. Объект этого класса принимает на вход батчи или файлы с Bag-Of-Words и подаётся на вход всем методам. В случае, если входные данные не являются батчами, он создаёт их и сохраняет на диск для последующего быстрого использования.

Итак, создадим объект BatchVectorizer:

In [2]:
```
batch_vectorizer = None
if len(glob.glob('kos' + "/*.batch")) < 1:
    batch_vectorizer = artm.BatchVectorizer(data_path='', data_format='bow
_uci', collection_name='kos', target_folder='kos')
else:
    batch_vectorizer = artm.BatchVectorizer(data_path='kos', data_format='
batches')
```

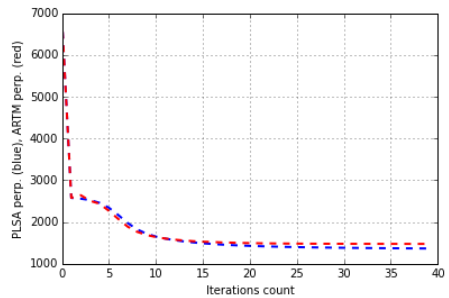
ARTM --- это класс, представляющий собой Python API BigARTM, и позволяющий использовать практически все возможности библиотеки в стиле scikit-learn. Создадим две тематические модели для нашего эксперимента. Наиболее важным параметром модели является число тем. Опционально можно указать списки регуляризаторов 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

Продолжим обучение моделей, инициировав 25 проходов по коллекции, после чего снова посмотрим на значения функционалов качества:

In [11]:
```
model_plsa.fit_offline(batch_vectorizer=batch_vectorizer, num_collection_p
asses=25, num_document_passes=1)
model_artm.fit_offline(batch_vectorizer=batch_vectorizer, num_collection_p
asses=25, num_document_passes=1)
```

In [12]:
```
print_measures(model_plsa, model_artm)
```

```
Sparsity Phi: 0.332 (PLSA) vs. 0.740 (ARTM)
Sparsity Theta: 0.082 (PLSA) vs. 0.602 (ARTM)
Kernel contrast: 0.530 (PLSA) vs. 0.568 (ARTM)
Kernel purity: 0.396 (PLSA) vs. 0.531 (ARTM)
Perplexity: 1365.804 (PLSA) vs. 1475.455 (ARTM)
```



Кроме того, для наглядности построим графики изменения разреженностей матриц по итерациям:

In [13]:
```
plt.plot(xrange(model_plsa.num_phi_updates), model_plsa.score_tracker['Spa
rsityPhiScore'].value, 'b--',
                 xrange(model_artm.num_phi_updates), model_artm.score_trac
ker['SparsityPhiScore'].value, 'r--', linewidth=2)
plt.xlabel('Iterations count')
plt.ylabel('PLSA Phi sp. (blue), ARTM Phi sp. (red)')
plt.grid(True)
plt.show()

plt.plot(xrange(model_plsa.num_phi_updates), model_plsa.score_tracker['Spa
rsityThetaScore'].value, 'b--',
                 xrange(model_artm.num_phi_updates), model_artm.score_trac
ker['SparsityThetaScore'].value, 'r--', linewidth=2)
```

Эксперимент 1. Обгоняем конкурентов по скорости

- 3.7M статей английской Вики, 100K уникальных слов

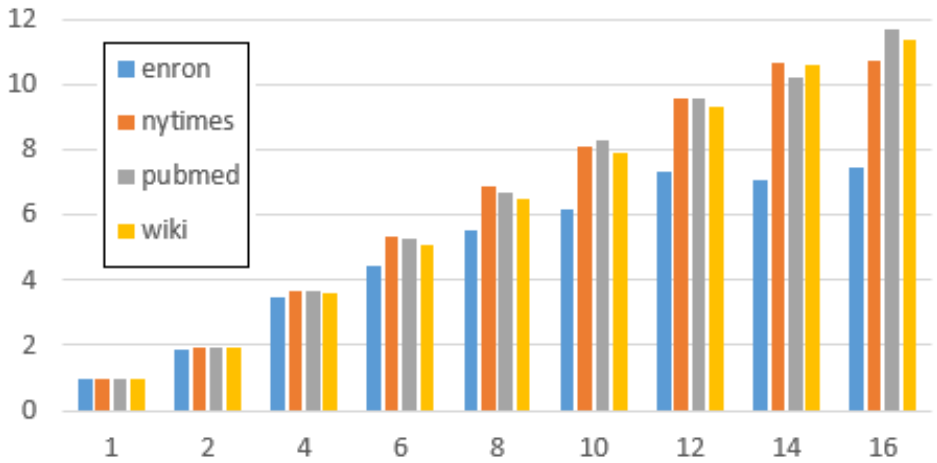
	procs	train	inference	perplexity
BigARTM	1	35 min	72 sec	4000
Gensim.LdaModel	1	369 min	395 sec	4161
VowpalWabbit.LDA	1	73 min	120 sec	4108
BigARTM	4	9 min	20 sec	4061
Gensim.LdaMulticore	4	60 min	222 sec	4111
BigARTM	8	4.5 min	14 sec	4304
Gensim.LdaMulticore	8	57 min	224 sec	4455

- *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

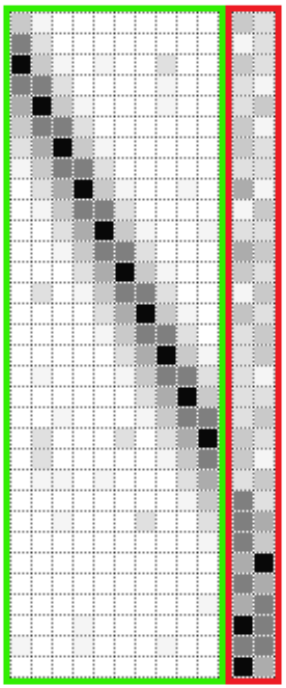
Amazon EC2 cc2.8xlarge instance:
16 cores + hyperthreading, Intel® Xeon® CPU E5-2670 2.6GHz.

Эксперимент 3. Комбинирование регуляризаторов

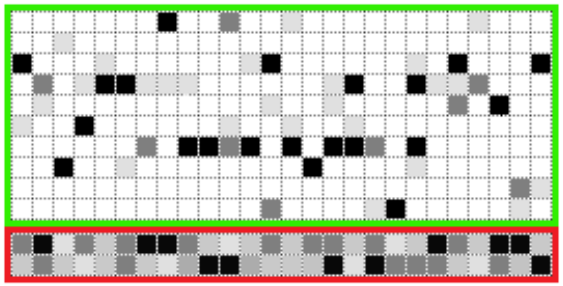
Предметные темы S содержат термины предметной области, $p(w|t)$ разреженные, существенно различные

Фоновые темы B содержат слова общей лексики, $p(w|t)$ и $p(t|d)$ не разреженные в этих темах

$\Phi_{W \times T}$



$\Theta_{T \times D}$



Напоминания. Дивергенция Кульбака–Лейблера

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

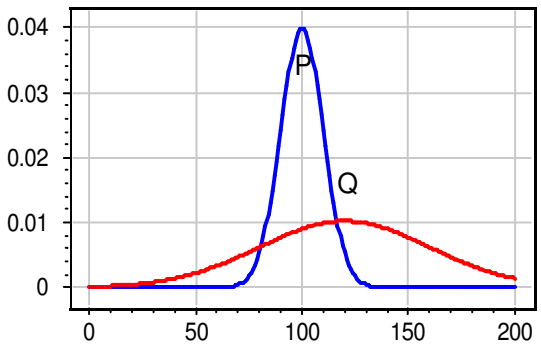
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

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

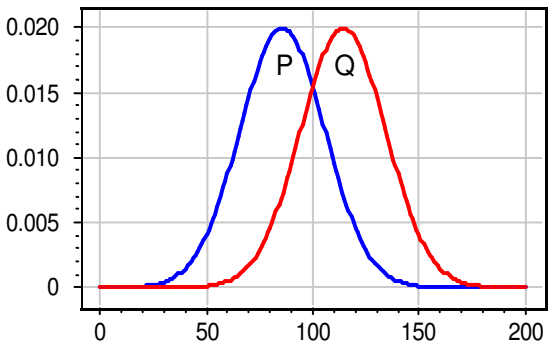
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

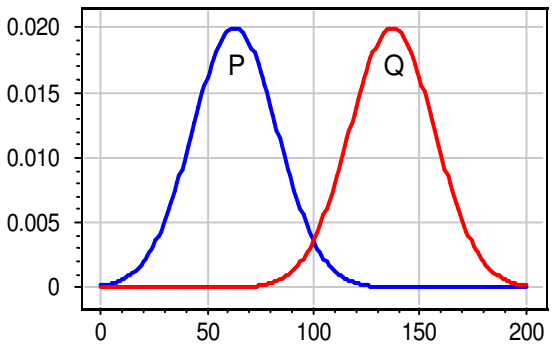
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



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



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



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

Регуляризатор сглаживания (почти совпадает с LDA)

Сглаживание фоновых тем $t \in B$:

распределения ϕ_{wt} близки к заданному распределению β_w

распределения θ_{td} близки к заданному распределению α_t

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

Максимизируем сумму регуляризаторов:

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

Подставляем, получаем формулы M-шага LDA:

$$\phi_{wt} = \underset{w}{\mathrm{norm}}(n_{wt} + \beta_0\beta_w), \qquad \theta_{td} = \underset{t}{\mathrm{norm}}(n_{td} + \alpha_0\alpha_t).$$

Этого вы не найдёте в *D.Blei, A.Ng, M.Jordan.* Latent Dirichlet allocation // Journal of Machine Learning Research, 2003. — Vol. 3. — Pp. 993–1022.

Регуляризатор разреживания (обобщение LDA)

Разреживание предметных тем $t \in S$:

распределения ϕ_{wt} далеки от заданного распределения β_w

распределения θ_{td} далеки от заданного распределения α_t

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

Максимизируем сумму регуляризаторов:

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

Подставляем, получаем «анти-LDA»:

$$\phi_{wt} = \underset{w}{\mathrm{norm}}\big(n_{wt} - \beta_0\beta_w\big), \qquad \theta_{td} = \underset{t}{\mathrm{norm}}\big(n_{td} - \alpha_0\alpha_t\big).$$

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010.

Регуляризатор декоррелирования тем

Цель — выделить *лексическое ядро* каждой темы, набор терминов, отличающий её от других тем.

Минимизируем ковариации между вектор-столбцами ϕ_t:

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

Подставляем, получаем ещё один вариант разреживания — постепенное контрастирование строк матрицы Φ:

$$\phi_{wt} = \underset{w}{\text{norm}} \left(n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \right).$$

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

Регуляризатор для максимизации когерентности тем

Гипотеза: тема лучше интерпретируется, если она содержит *когерентные* (часто встречающиеся рядом) слова $u, w \in W$.

Пусть C_{uw} — оценка когерентности, например $\hat{p}(w|u) = \frac{N_{uw}}{N_u}$.
Согласуем ϕ_{wt} с оценками $\hat{p}(w|t)$ по когерентным словам,

$$\hat{p}(w|t) = \sum_u p(w|u)p(u|t) = \frac{1}{n_t}\sum_u C_{uw} n_{ut};$$

$$R(\Phi,\Theta) = \tau \sum_{t\in T} n_t \sum_{w\in W} \hat{p}(w|t) \ln \phi_{wt} \to \max.$$

Подставляем, получаем ещё один вариант сглаживания:

$$\phi_{wt} = \underset{w}{\operatorname{norm}}\left(n_{wt} + \tau \sum_{u\in W\setminus w} C_{uw} n_{ut}\right).$$

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Empirical Methods in Natural Language Processing, EMNLP-2011. — Pp. 262–272.

Разреживание + Сглаживание + Декорреляция

M-шаг при комбинировании 5 регуляризаторов:

$$\phi_{wt} = \underset{w}{\text{norm}}\Big(n_{wt} + \tau_1 \underbrace{\beta_w[t \in B]}_{\substack{\text{сглаживание}\\ \text{фоновых}\\ \text{тем}}} - \tau_2 \underbrace{\beta_w[t \in S]}_{\substack{\text{разреживание}\\ \text{предметных}\\ \text{тем}}} - \tau_3 \phi_{wt} \underbrace{\sum_{s \in S \setminus t} \phi_{ws}}_{\text{декорреляция}} \Big)$$

$$\theta_{td} = \underset{t}{\text{norm}}\Big(n_{td} + \tau_4 \underbrace{\alpha_t[t \in B]}_{\substack{\text{сглаживание}\\ \text{фоновых}\\ \text{тем}}} - \tau_5 \underbrace{\alpha_t[t \in S]}_{\substack{\text{разреживание}\\ \text{предметных}\\ \text{тем}}} \Big)$$

Траектория регуляризации (*regularization path*) в пространстве $\tau = (\tau_1, \ldots, \tau_5)$ подбирается экспериментально.

Воронцов К. В., Потапенко А. А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Межд. конф. по компьютерной лингвистике Диалог-2014.

Эксперимент на коллекции NIPS

Данные: NIPS (Neural Information Processing System)

- $|D| = 1566$ статей конференции NIPS на английском языке;
- суммарной длины $n \approx 2.3 \cdot 10^6$,
- словарь $|W| \approx 1.3 \cdot 10^4$.
- контрольная коллекция: $|D'| = 174$.

Измерение качества модели:

- перплексия контрольной коллекции: $\mathcal{P} = \exp(-\frac{1}{n'}\mathscr{L}(D'))$
- разреженность — доля нулевых элементов в Φ и Θ
- чистота, контрастность, когерентность, размер ядра тем

Воронцов К. В., Потапенко А. А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Межд. конф. по компьютерной лингвистике Диалог-2014.

Критерии качества модели

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

- Перплексия контрольной коллекции: $\mathcal{P} = \exp\left(-\frac{1}{n'} L(D')\right)$
- Разреженность — доля нулевых элементов в Φ и Θ
- Характеристики интерпретируемости тем:
  - когерентность темы [Newman, 2010]
  - размер ядра темы: $|W_t|$, ядро $W_t = \{w : p(t|w) > 0.25\}$
  - чистота темы: $\sum_{w \in W_t} p(w|t)$
  - контрастность темы: $\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$
- Вырожденность тематической модели:
  - число тем: $|T|$
  - доля фоновых слов: $\frac{1}{n} \sum_{d \in D} \sum_{w \in d} \sum_{t \in B} p(t|d, w)$

Оценки интерпретируемости: когерентность

Когерентность темы t

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

где w_i — i-й термин в порядке убывания ϕ_{wt}.

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

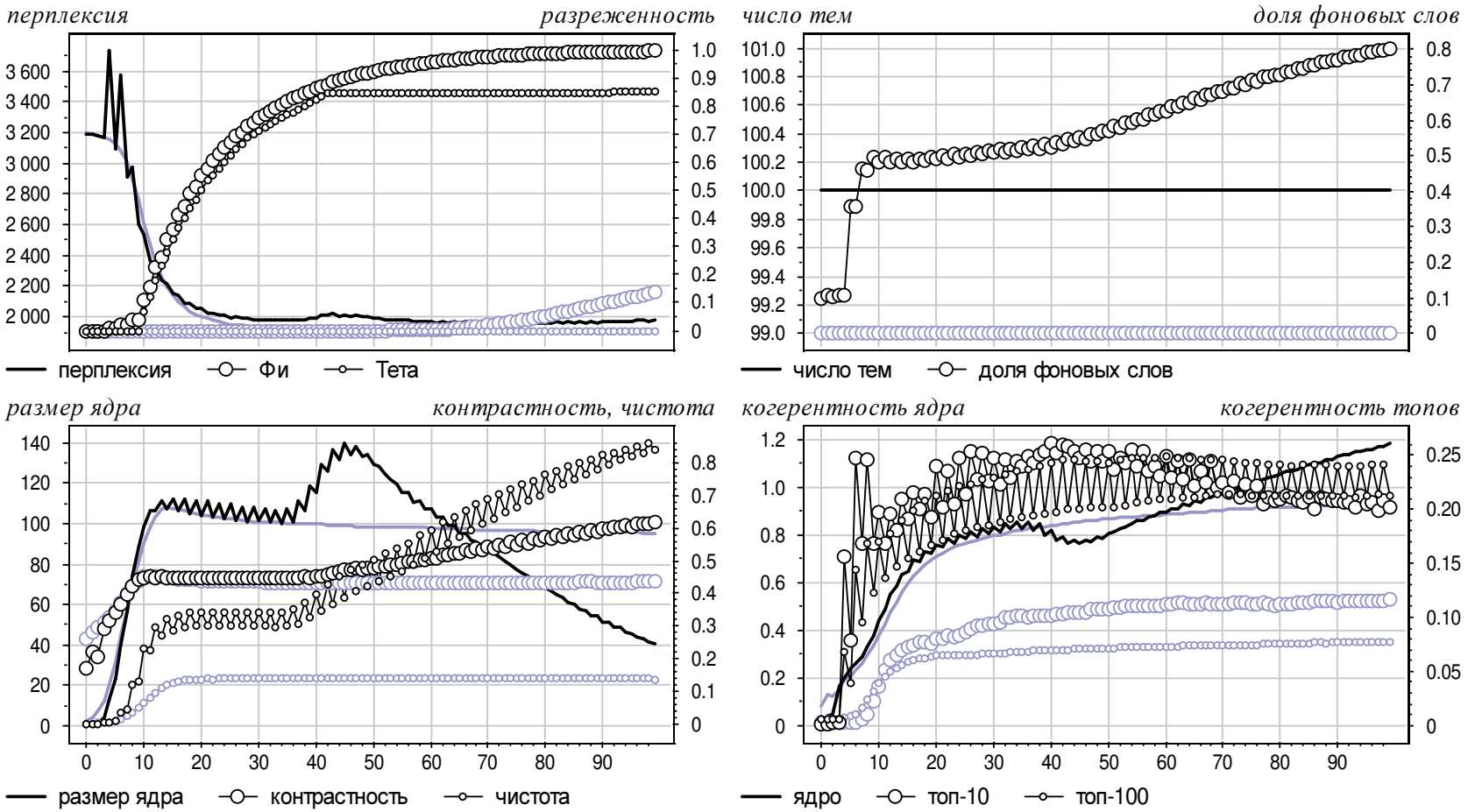
N_{uv} — число документов, в которых термины u, v хотя бы один раз встречаются рядом (в окне 10 слов),

N_u — число документов, в которых u встретился хотя бы 1 раз.

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Эксперимент 3. Комбинирование регуляризаторов

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



Эксперимент 4. 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

Эксперимент 5. Динамическая тематическая модель

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

Разреживание тем $p(t|y) = \sum_{d \in D_y} \theta_{td} p(d)$ в каждый момент y:

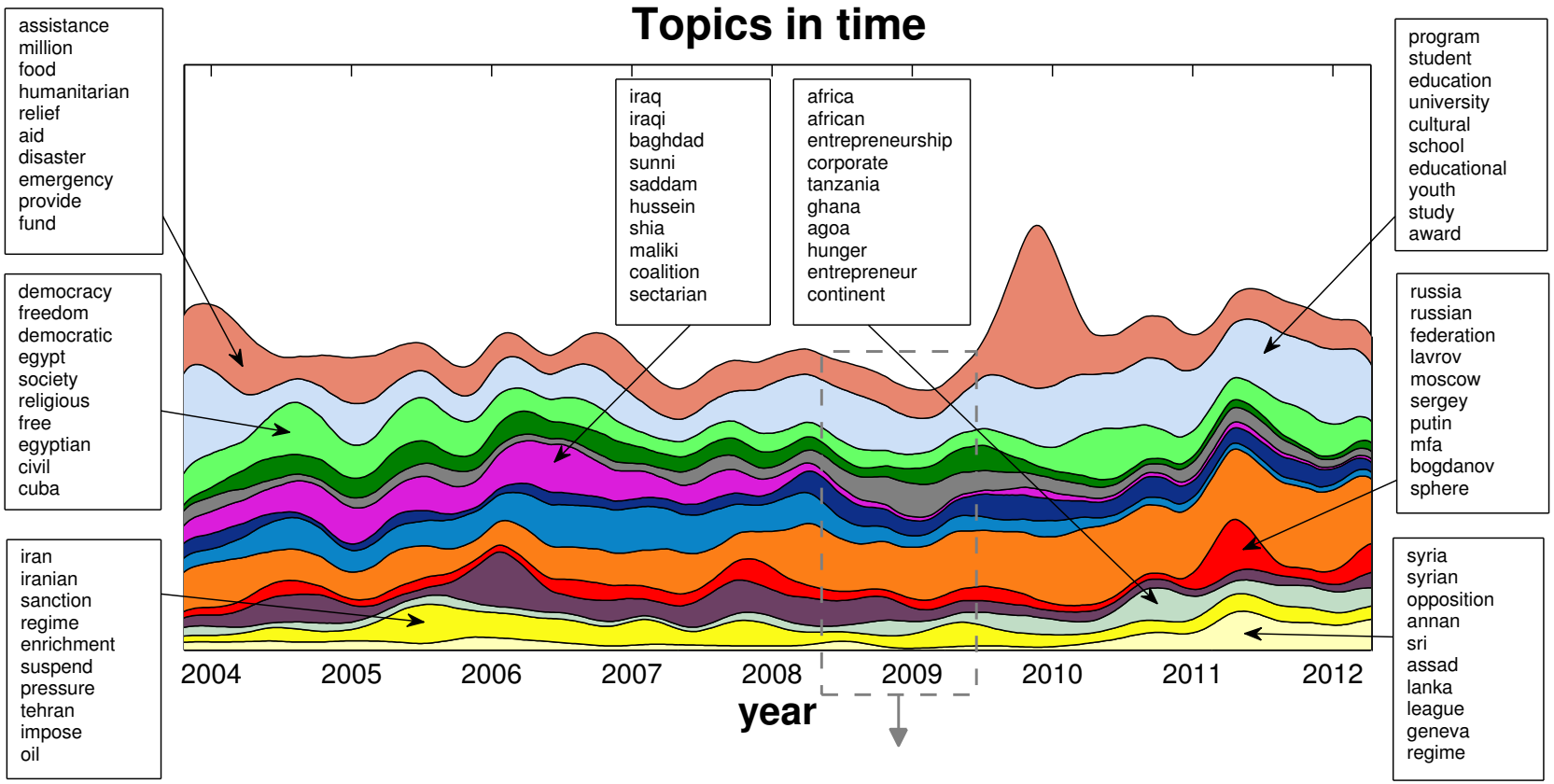
$$R_1(\Theta) = -\tau_1 \sum_{y \in Y} \mathrm{KL}\Big(\tfrac{1}{|T|} \| p(t|y)\Big) \to \max.$$

Сглаживание тем $p(y|t)$ в соседние моменты y, $y-1$:

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

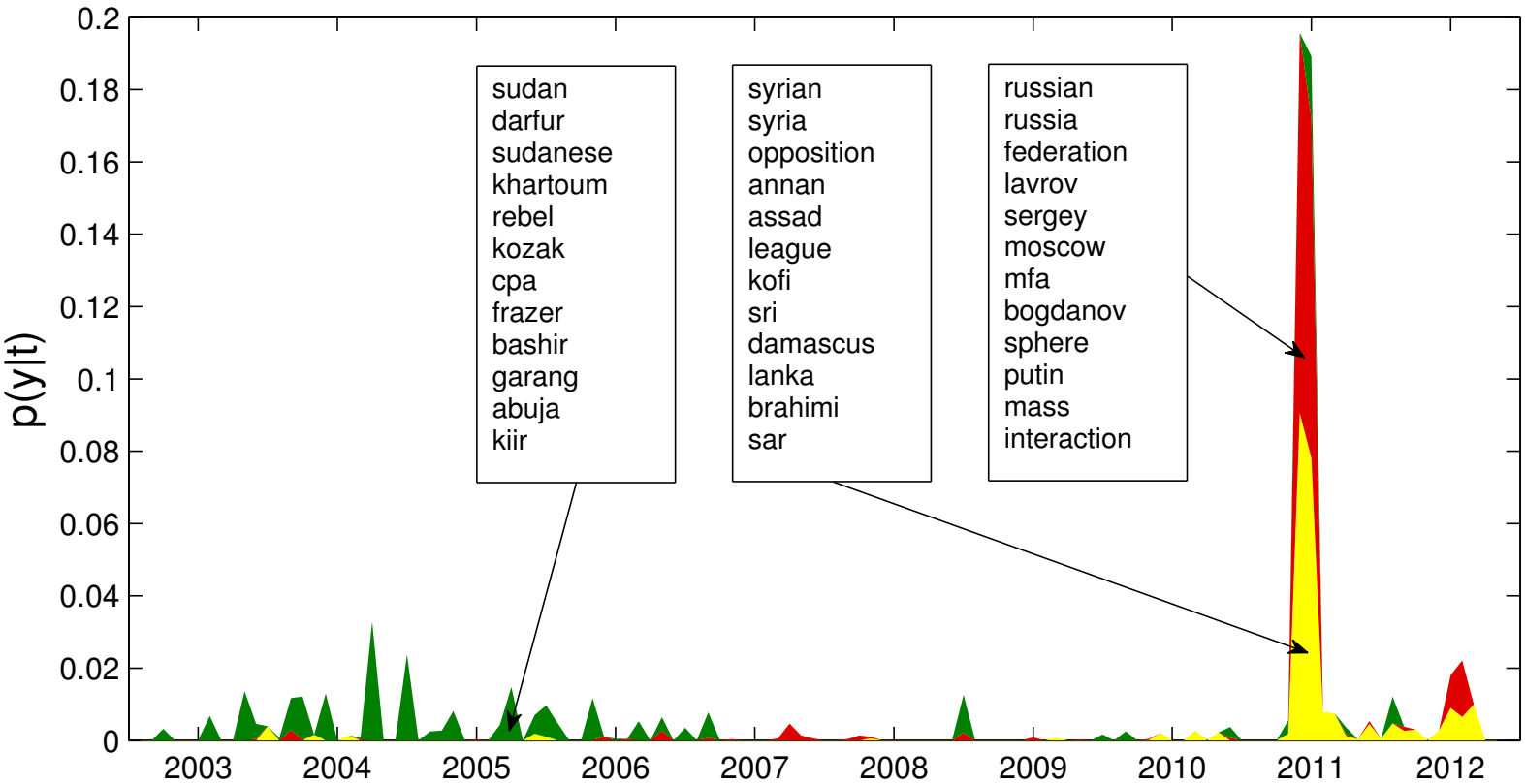
Эксперимент 5. Задача анализа потока пресс-релизов

Коллекция официальных пресс-релизов внешнеполитических ведомств ряда стран на английском языке.
Более 20 тыс. сообщений за 10 лет, 180Мб текста.



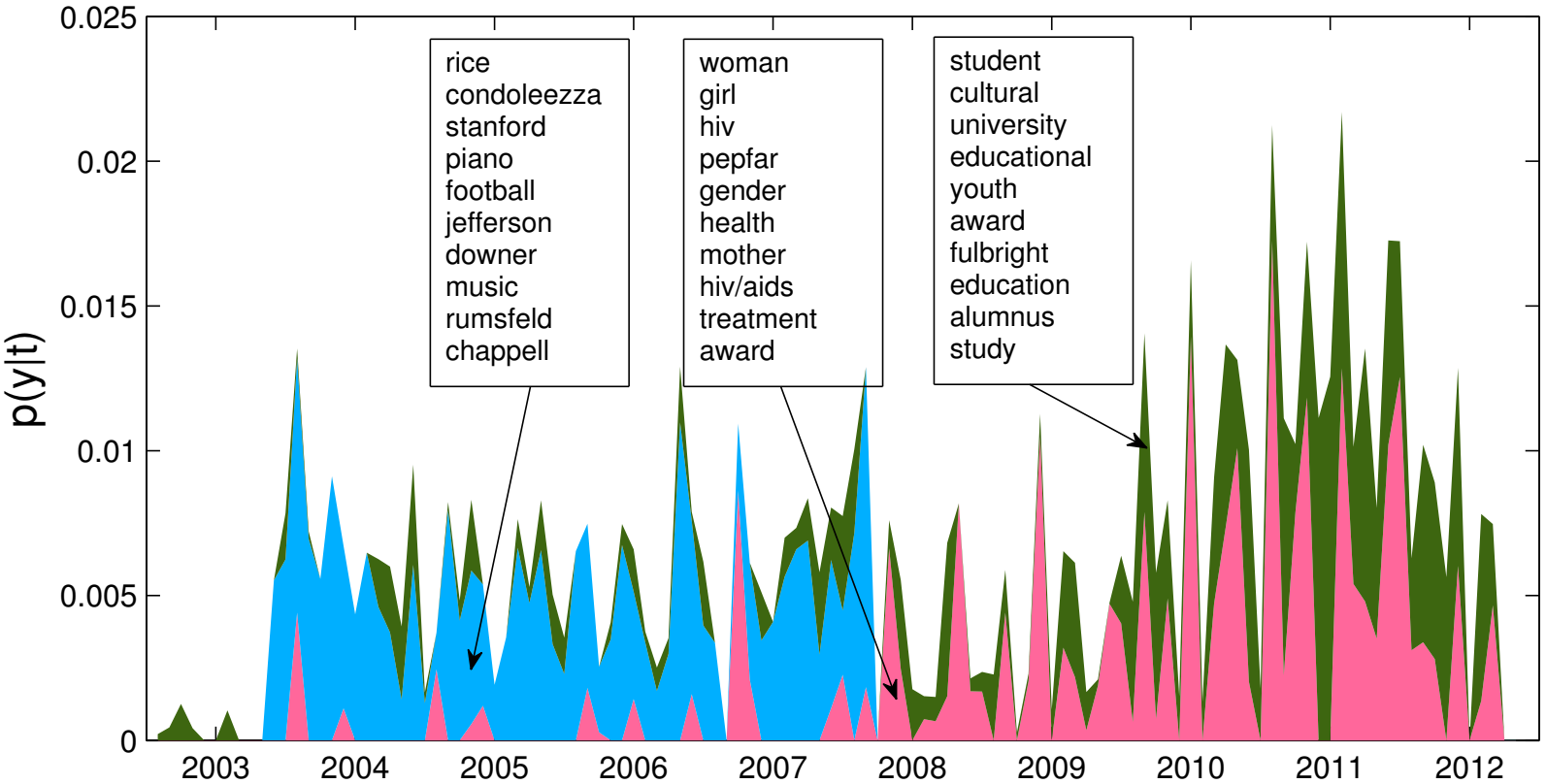
Эксперимент 5. Задача анализа потока пресс-релизов

Примеры событийных тем и момента их совместного всплеска



Эксперимент 5. Задача анализа потока пресс-релизов

Примеры перманентных тем



Эксперимент 6. Энтропийный регуляризатор для отбора тем

Чтобы сделать распределение $p(t)$ разреженным, максимизируем его KL-дивергенцию с равномерным распределением: $\mathrm{KL}\left(\frac{1}{|T|} \,\|\, p(t)\right) \to \max$:

$$R(\Theta) = -\tau n \sum_{t \in S} \frac{1}{|T|} \ln \underbrace{\sum_{d \in D} p(d)\theta_{td}}_{p(t)} \to \max.$$

Регуляризованный M-шаг разреживает строки Θ целиком:

$$\theta_{td} = \underset{t \in T}{\mathrm{norm}} \left(n_{td} \left(1 - \tau \frac{n}{n_t |T|} \right) \right).$$

Если $n_t < \tau \frac{n}{|T|}$, то все элементы t-й строки обращаются в нуль.

Эксперимент 6. Энтропийный регуляризатор для отбора тем

Коллекция статей NIPS (Neural Information Processing System)

- $|D| = 1566$ обучающих документов; $|D'| = 174$ тестовых
- $|W| \approx 1.3 \cdot 10^4$ — мощность словаря

Синтетическая коллекция:

- строим PLSA за 500 итераций, $|T_0| = 50$ тем на NIPS
- генерируем (n^0_{dw}) из полученных Φ и Θ:

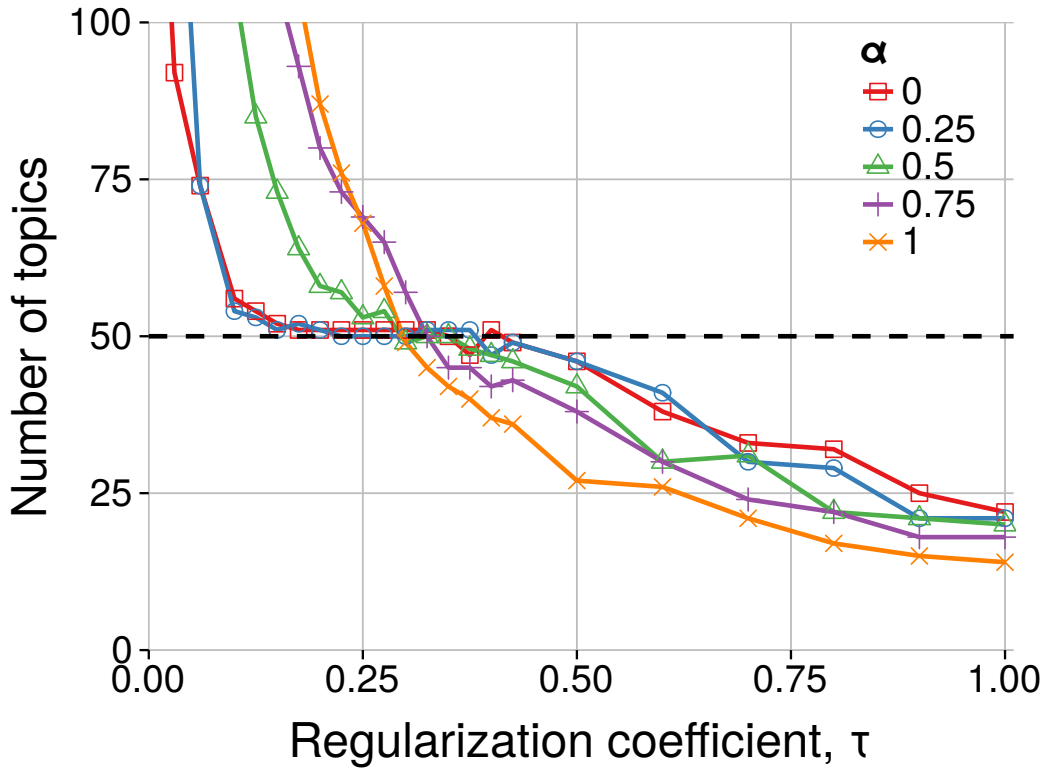
$$n^0_{dw} = n_d \sum_{t \in T} \phi_{wt} \theta_{td}$$

Параметрическое семейство полусинтетических данных:

- n^α_{dw} — смесь синтетических данных n^0_{dw} и реальных n_{dw}:

$$n^\alpha_{dw} = \alpha n_{dw} + (1 - \alpha) n^0_{dw}$$

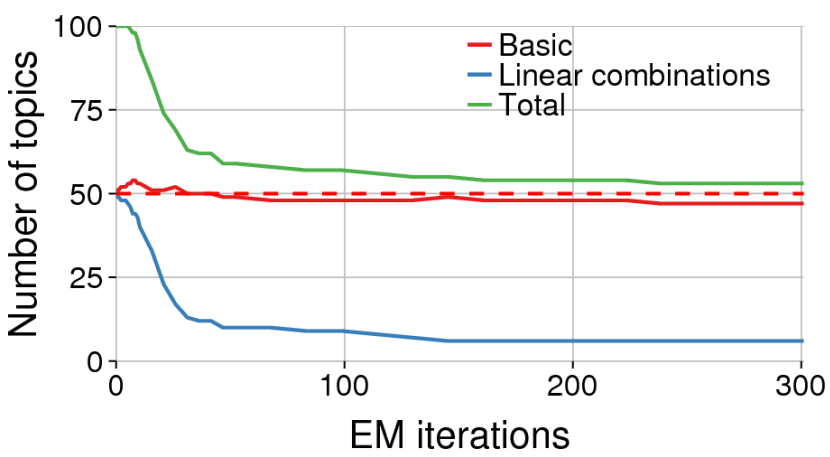
Попытка определения числа тем

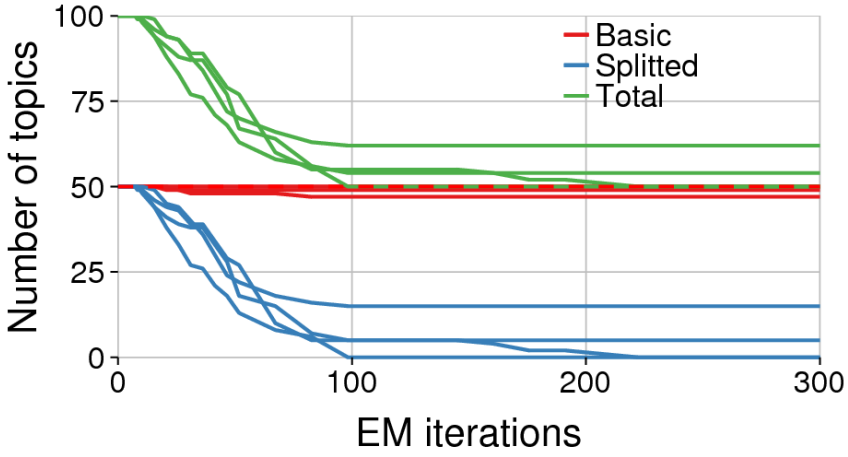


- На синтетических данных надёжно находим $|T| = 50$,
- в широком интервале значений коэффициента τ;
- однако на реальных данных нет столь чёткого интервала.

Удаление линейно зависимых и расщеплённых тем

- Добавили 50 линейных комбинаций тем в модельную Φ.
- Расщепили 50 тем, каждую на две подтемы в модельной Φ.



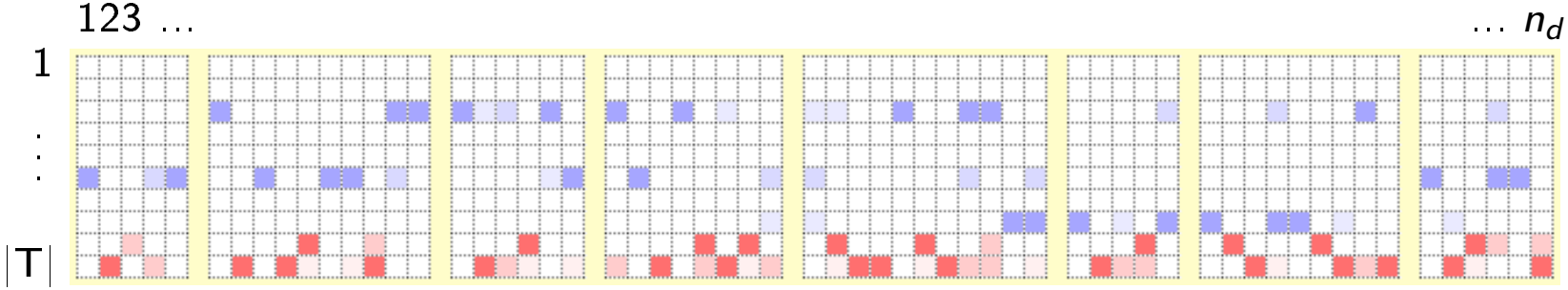


- Удаляются линейно зависимые и расщеплённые темы
- Остаются более различные темы исходной модели.

Тематическое моделирование последовательного текста

Документ $d = \{w_1, \ldots, w_{n_d}\}$, 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

Позиционный регуляризатор E-шага

Позиционный регуляризатор R_{di} зависит от позиции слова i в документе d и от параметров Φ, Θ через $p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$,

$$\mathscr{L}(\Phi,\Theta) + \sum_{d\in D}\sum_{i=1}^{n_d} R_{di}(p_{1dw_i},\ldots,p_{Tdw_i}) + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta},$$

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

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:

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

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:

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

Мотивации

Выборка может содержать не только пары (d, w), но также тройки, ..., 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

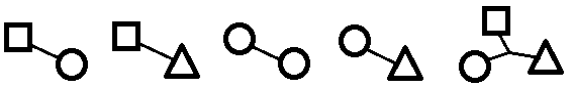
Хотим объяснить наблюдаемую выборку рёбер гиперграфа латентными тематическими профилями его вершин.

Тематическая модель гиперграфа: определения и обозначения

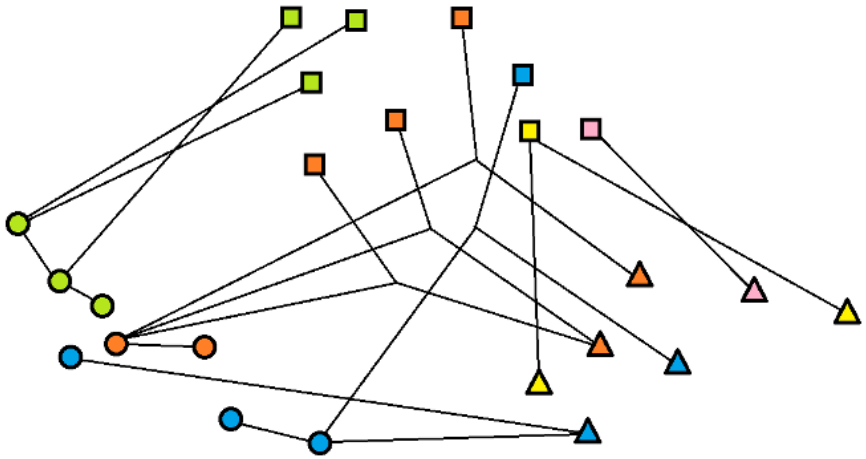
$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.
$V = V^1 \sqcup \cdots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

K — множество типов рёбер:



T — множество тем:



X^k — наблюдаемая выборка транзакций — рёбер типа k
ребро (d, x): вершина-контейнер $d \in V$ и вершины $x \subset V$,
n_{dx} — число вхождений ребра (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
$\phi_{kvt} = p_k(v|t)$ — для модальности v в теме t на рёбрах типа k

Задача максимизации log правдоподобия:

$$\sum_{k\in K} \tau_k \sum_{(d,x)\in X^k} n_{dx} \ln \sum_{t\in T} \theta_{td} \prod_{v\in x} \phi_{kvt} \to \max_{\Phi,\Theta},$$

$$\phi_{kvt} \geqslant 0, \quad \sum_{v\in V^m} \phi_{kvt} = 1, \ k \in K; \qquad \theta_{td} \geqslant 0, \quad \sum_{t\in T} \theta_{td} = 1,$$

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k\in K}\tau_k \sum_{(d,x)\in X^k} n_{dx} \ln \sum_{t\in T} \theta_{td} \prod_{v\in x} \phi_{kvt} \;+\; R(\Phi,\Theta) \to \max_{\Phi,\Theta}.$$

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdx} = p(t|d,x)$:

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

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

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

Пространственно-временная этно-тематическая модель

- Выделение этно-релевантных тем
  - регуляризатор частичного обучения по словарю этнонимов
  - альтернативный вариант: этнонимы как модальность
  - регуляризаторы для контрастирования малых тем
- Оценивание распределения тем по регионам
  - разреживание региональных тем
  - сглаживание общих тем
- Оценивание распределения тем по времени
  - разреживание событийных тем
  - сглаживание перманентных тем
- Повышение качества тематической модели
  - использование модальности авторов
  - лингвистическая регуляризация и выделение мультиграмм

Регуляризатор частичного обучения (обобщение LDA)

W_t^1 и W_t^0 — белый и чёрный список терминов темы t
T_d^1 и T_d^0 — белый и чёрный список тем документа d

Максимизируем сумму регуляризаторов:

$$R(\Phi,\Theta) = +\beta_1 \sum_{t\in T}\sum_{w\in W_t^1} \beta_{wt}^1 \ln\phi_{wt} + \alpha_1 \sum_{d\in D}\sum_{t\in T_d^1} \alpha_{td}^1 \ln\theta_{td} -$$
$$-\beta_0 \sum_{t\in T}\sum_{w\in W_t^0} \beta_{wt}^0 \ln\phi_{wt} - \alpha_0 \sum_{d\in D}\sum_{t\in T_d^0} \alpha_{td}^0 \ln\theta_{td} \to \max$$

Подставляем, получаем обобщение LDA:

$$\phi_{wt} = \underset{w}{\text{norm}}\Big(n_{wt} + \beta_1\beta_{wt}^1\big[w\in W_t^1\big] - \beta_0\beta_{wt}^0\big[w\in W_t^0\big]\Big)$$
$$\theta_{td} = \underset{t}{\text{norm}}\Big(n_{td} + \alpha_1\alpha_{td}^1\big[t\in T_d^1\big] - \alpha_0\alpha_{td}^0\big[t\in T_d^0\big]\Big)$$

Регуляризатор частичного обучения (второе обобщение)

Вместо $\ln(z)$ можно взять другую монотонную функцию, например, $\mu(z) = z$ и максимизировать сумму ковариаций:

$$R(\Phi, \Theta) = +\beta_1 \sum_{t \in T} \sum_{w \in W_t^1} \phi_{wt} + \alpha_1 \sum_{d \in D} \sum_{t \in T_d^1} \theta_{td} -$$
$$- \beta_0 \sum_{t \in T} \sum_{w \in W_t^0} \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in T_d^0} \theta_{td} \to \max$$

Подставляем, получаем ещё одно обобщение LDA:

$$\phi_{wt} = \underset{w}{\text{norm}} \Big(n_{wt} + \beta_1 \phi_{wt} \big[w \in W_t^1 \big] - \beta_0 \phi_{wt} \big[w \in W_t^0 \big] \Big)$$
$$\theta_{td} = \underset{t}{\text{norm}} \Big(n_{td} + \alpha_1 \theta_{td} \big[t \in T_d^1 \big] - \alpha_0 \theta_{td} \big[t \in T_d^0 \big] \Big)$$

Если θ_{td}^0 равномерно на T_d^1, то ковариация не накладывает ограничений на распределение θ_{td} между темами из T_d^1.

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!

References


Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.


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


Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models. Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.


Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. AIST'2014, Analysis of Images, Social networks and Texts. Springer, 2014. CCIS, Vol. 436. pp. 29–46.


Vorontsov K. V., Frei O. I., Apishev M. A., Romov P. A., Suvorova M. A., Yanina A. O. Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections. Topic Models: Post-Processing and Applications, CIKM 2015 Workshop, October 19, 2015, Melbourne, Australia.


Воронцов К. В., Фрей А. И., Апишев М. А., Ромов П. А., Суворова М. А., Янина А. О. BigARTM: библиотека с открытым кодом для тематического моделирования больших текстовых коллекций // DAMDID/RCDL'2015, Обнинск, 13–16 октября 2015.