Реализация мультимодальных регуляризованных тематических моделей в библиотеке с открытым кодом BigARTM

Мурат Апишев
great-mel@yandex.ru

МГУ им. М. В. Ломоносова, фак-т ВМК

16 апреля 2015

Тематическое моделирование

Тематическое моделирование — одно из приложений машинного обучения к анализу текстов.

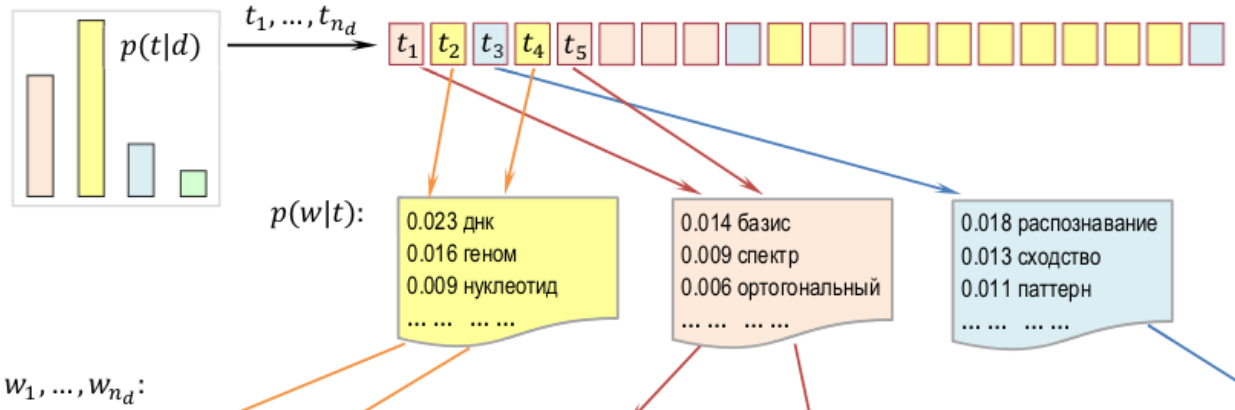
Дано: Корпус текстов D со словарём W в виде матрицы «слова – документы» F (Bag-of-words). На пересечении строки и столбца — вероятность встретить слово w в документе d.

Найти: Множество тем, описывающих D, т.е. матричное разложение $F = \Phi_{W \times T} \times \Theta_{T \times D}$.
$\phi_{wt} = p(w|t)$ — вероятность того, что слово w относится к теме t,
$\theta_{td} = p(t|d)$ — вероятность того, что документ d описывает тему t.

$$p(w|d) = \sum_t p(w|t)p(t|d)$$

Тематическое моделирование

$$p(w|d) = \sum_t p(w|t)p(t|d)$$



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).

Аддитивная регуляризация тематических моделей

APTM [1] — один из подходов к регуляризации процесса обучения тематических моделей.
Для нахождения матричного разложения будем максимизировать логарифм правдоподобия коллекции. Задача некорректно поставленная и должна быть регуляризована:

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

где n_{dw} — число раз, которое слово w встретилось в документе d.
Ограничения неотрицательности и нормировки

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

Регуляризаторы — это дополнительные ограничения на правдоподобие, позволяющие получить модель с нужными свойствами.

[1] Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.
Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. AIST. 2014

Основная теорема APTM

Теорема

Если функция $R(\Phi, \Theta)$ стохастических матриц Φ, Θ непрерывно дифференцируема и (Φ, Θ) является точкой локального максимума функции $L(\Phi, \Theta) + R(\Phi, \Theta)$, то для любой темы t и документа d выполняется система уравнений:

$$p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}; \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw};$$

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+; \quad \theta_{td} \propto \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+.$$

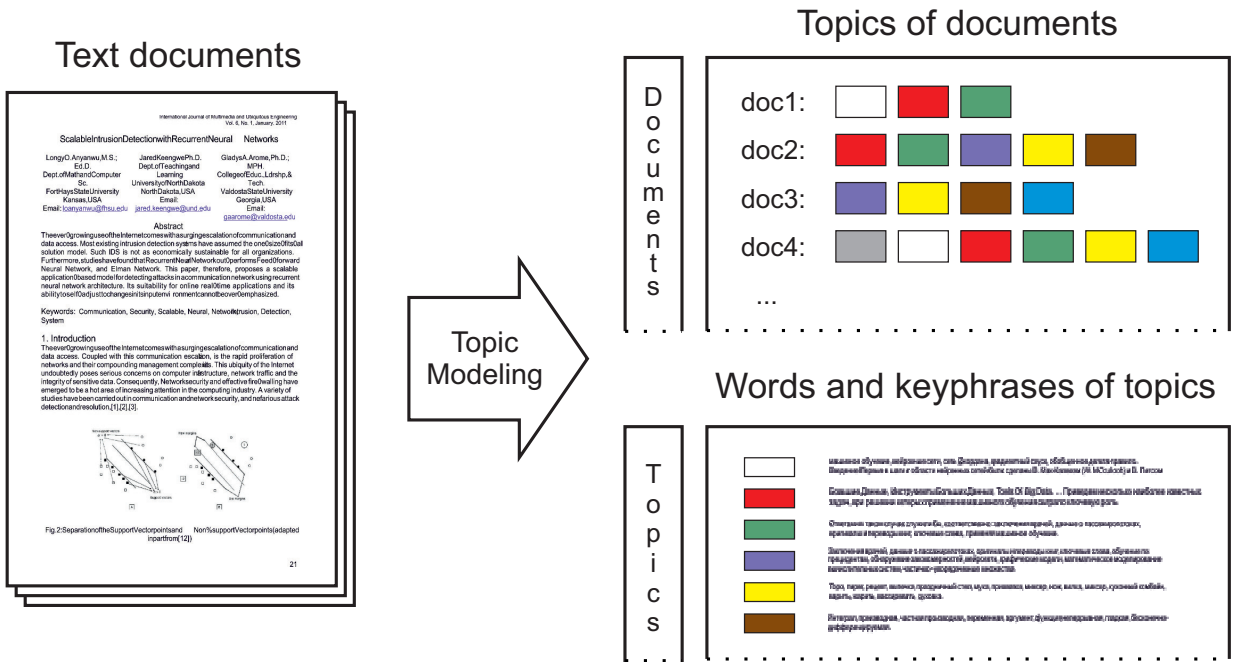
Данная задача может быть решена с помощью регуляризованного EM-алгоритма.

Мультимодальные тематические модели

Дана текстовая коллекция D, тематическое моделирование ищет:

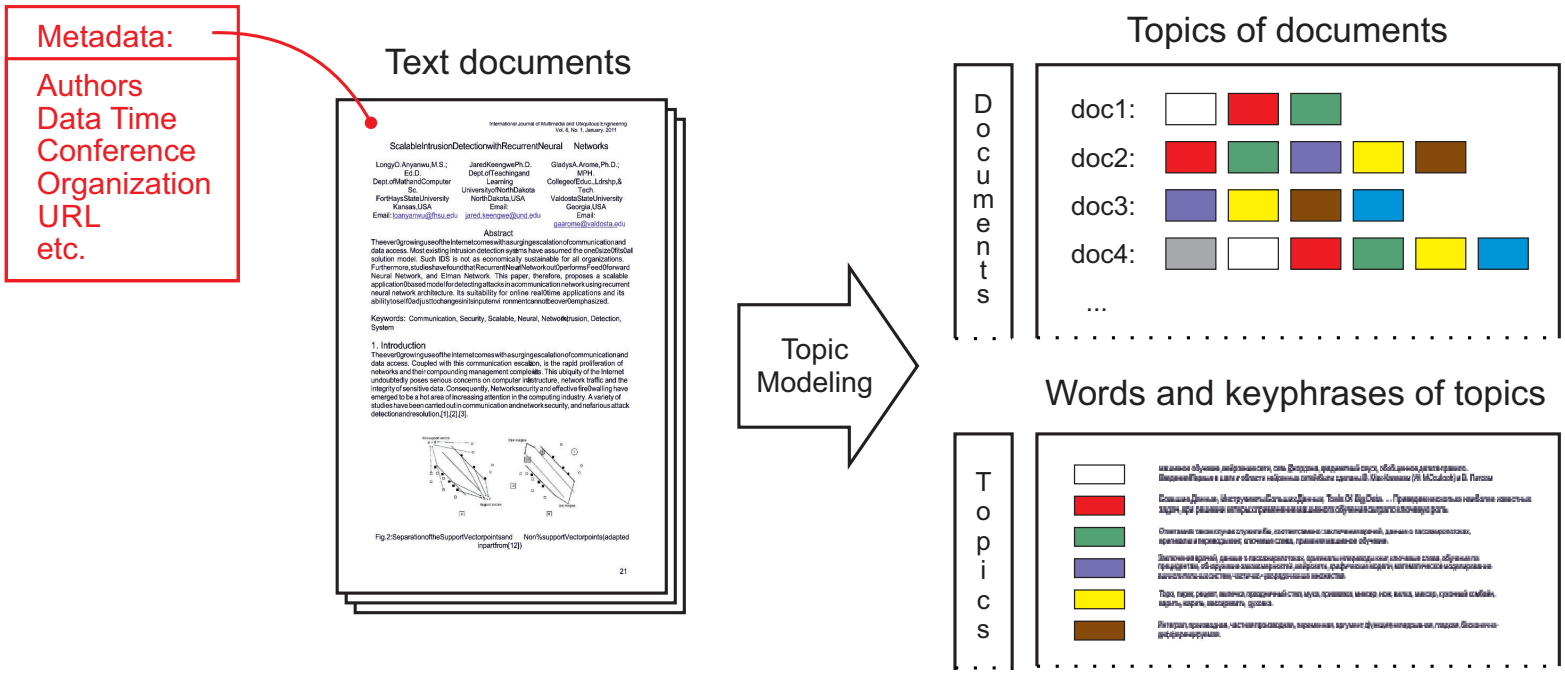
$p(t|d)$ — распределение по темам документа d,

$p(w|t)$ — распределение слов в теме t.



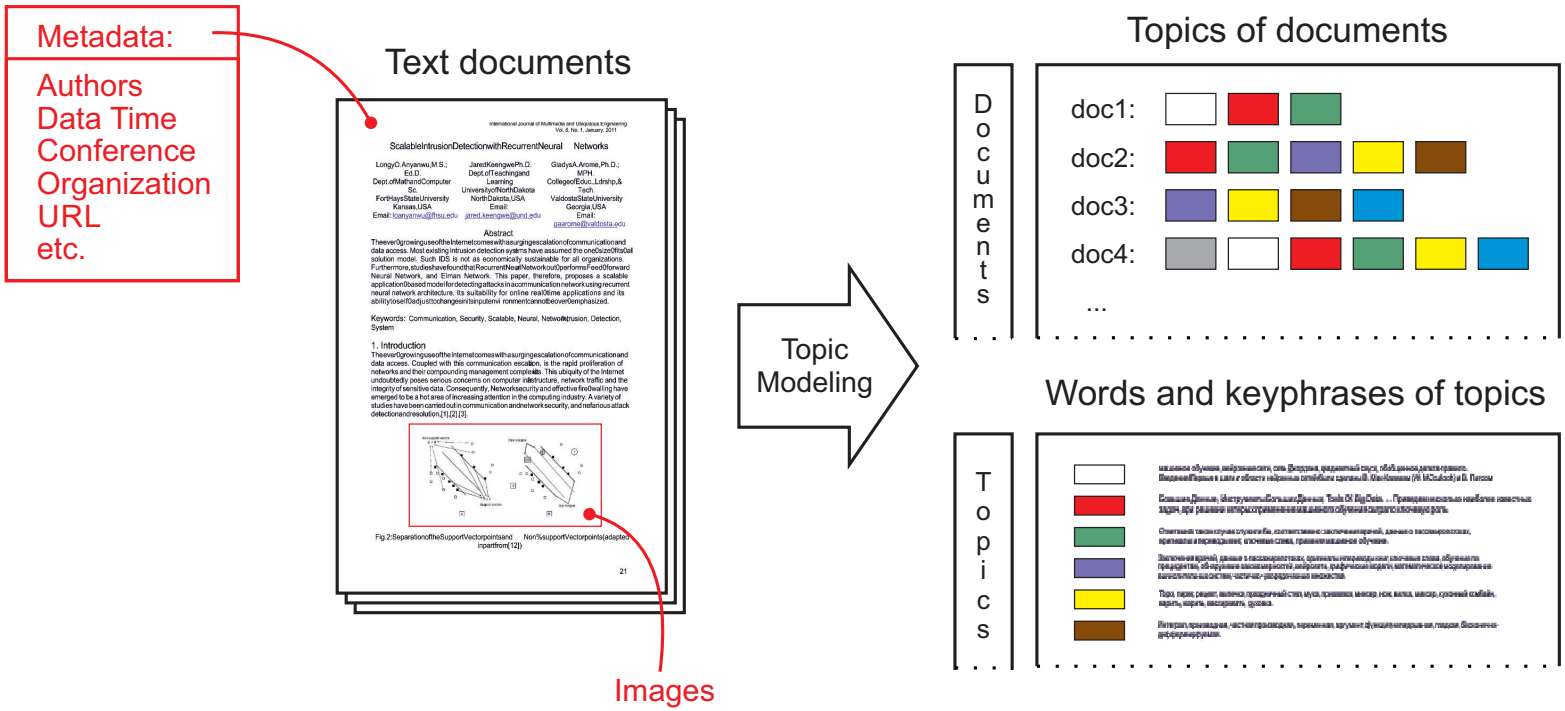
Мультимодальные тематические модели

M-APTM ищет распределения по темам $p(w|t)$, авторов $p(a|t)$, меток времени $p(y|t)$,



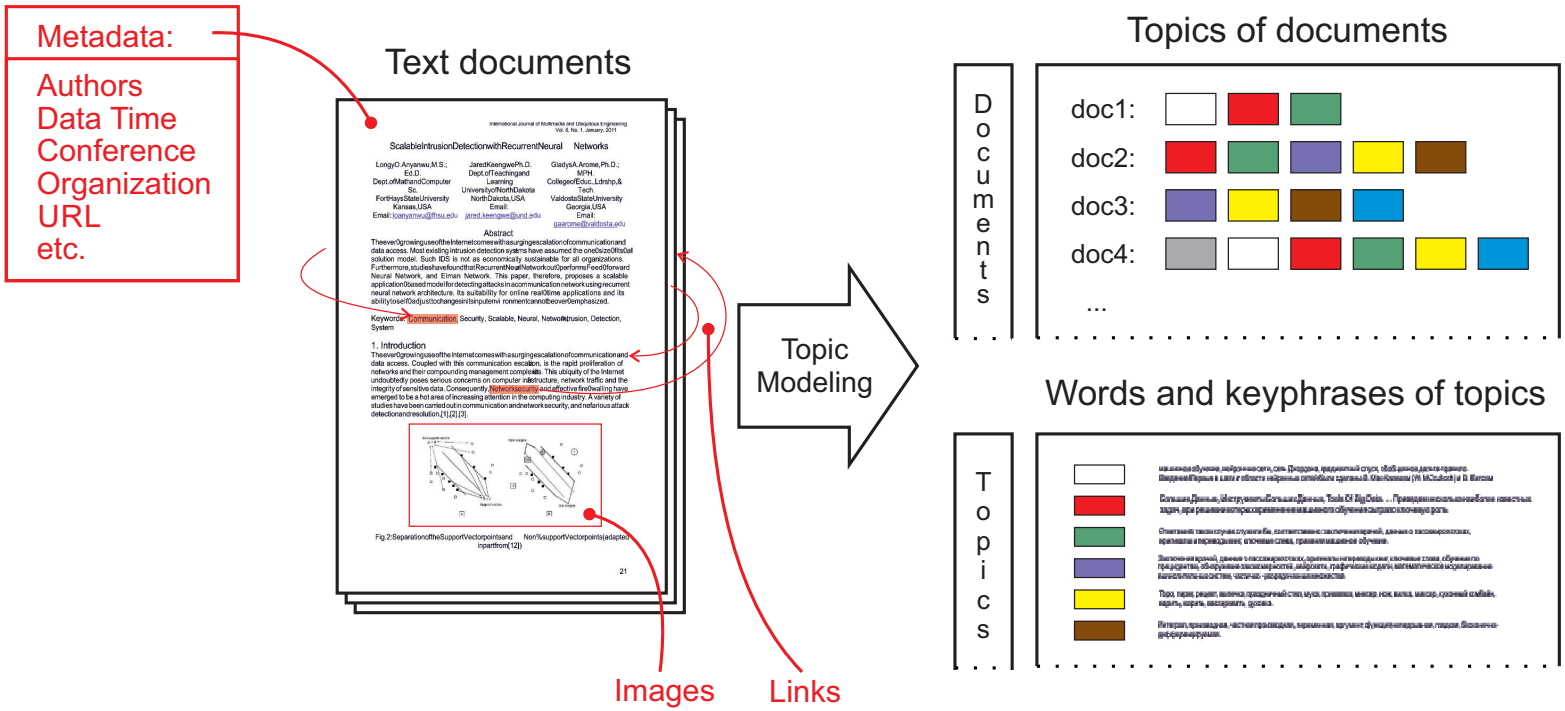
Мультимодальные тематические модели

M-APTM ищет распределения по темам $p(w|t)$, авторов $p(a|t)$, меток времени $p(y|t)$, объектов на картинках $p(o|t)$,



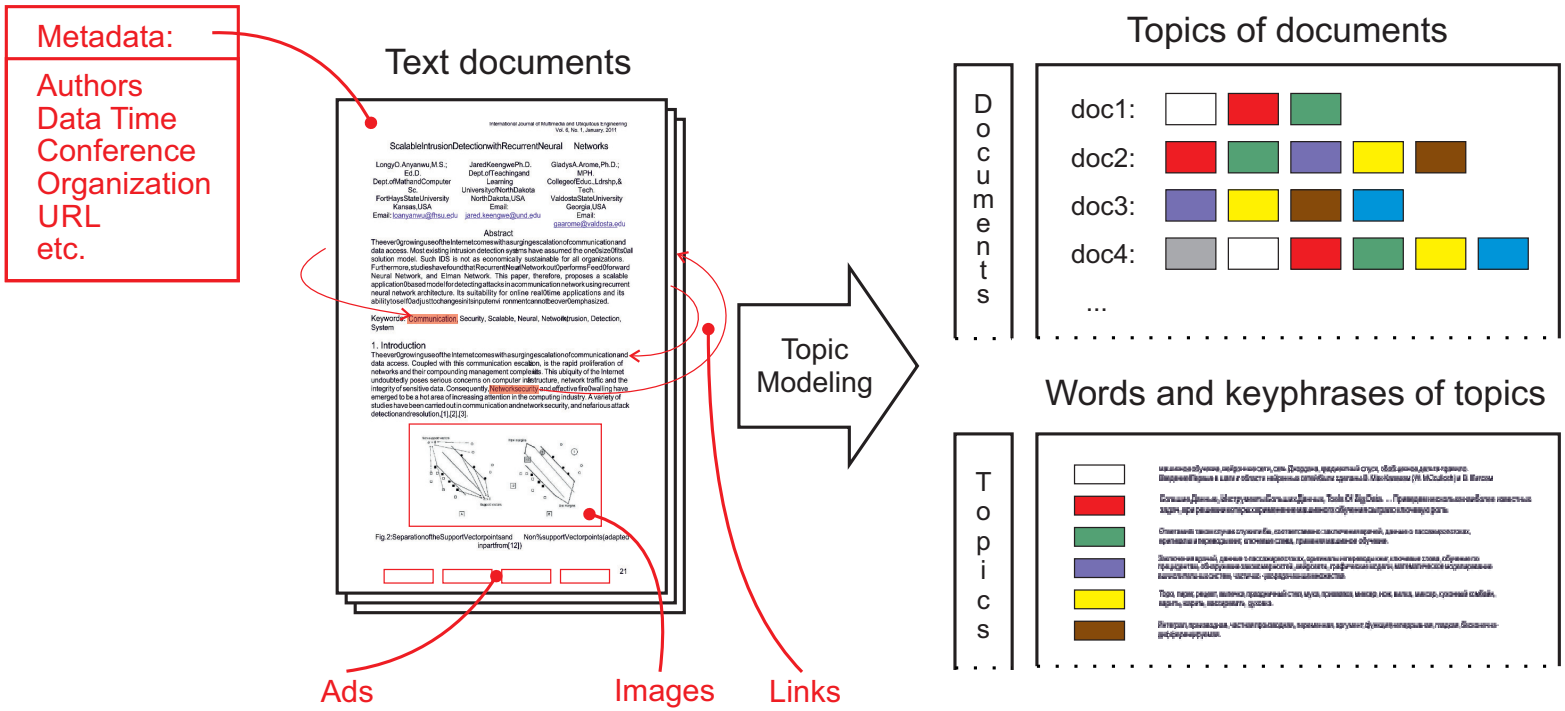
Мультимодальные тематические модели

M-APTM ищет распределения по темам $p(w|t)$, авторов $p(a|t)$, меток времени $p(y|t)$, объектов на картинках $p(o|t)$, связанных документов $p(d'|t)$,



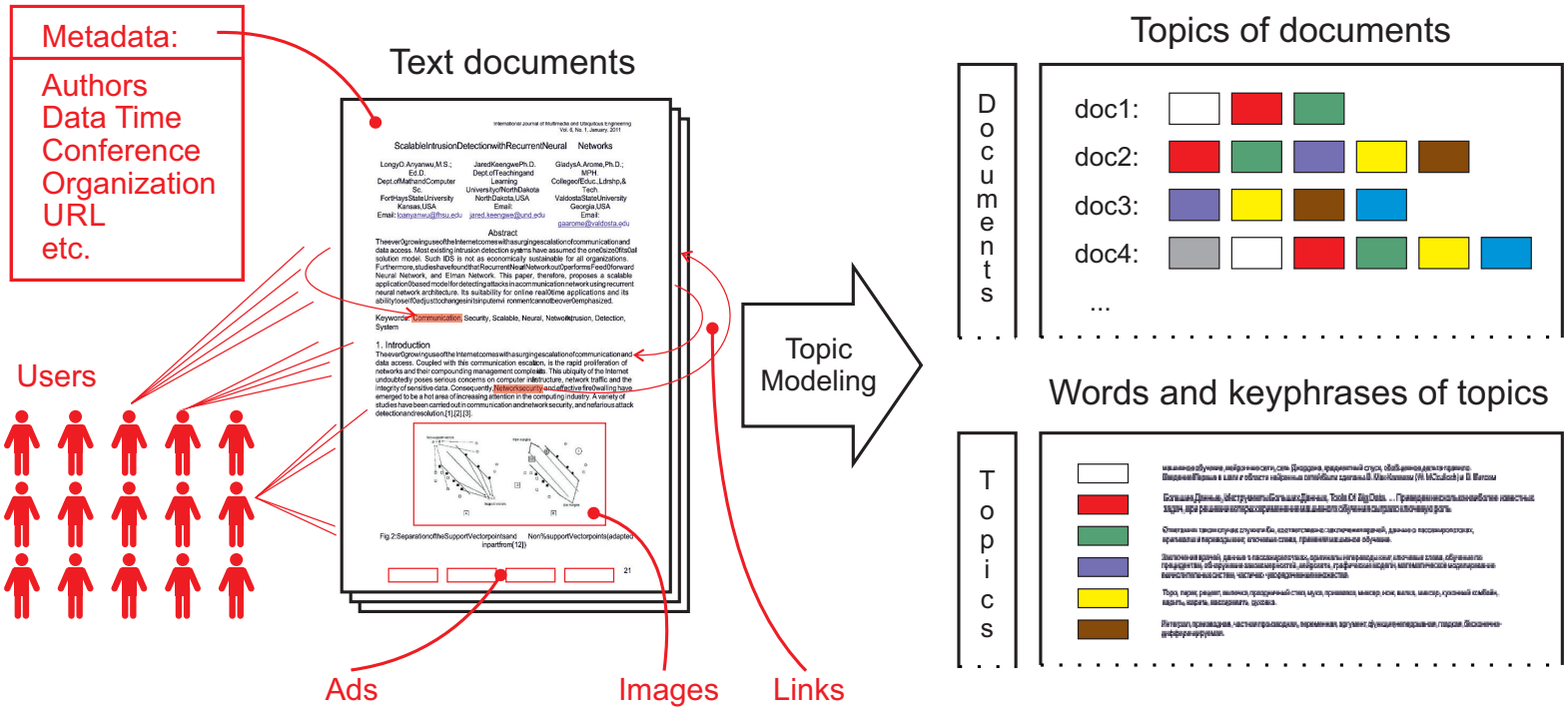
Мультимодальные тематические модели

M-APTM ищет распределения по темам $p(w|t)$, авторов $p(a|t)$, меток времени $p(y|t)$, объектов на картинках $p(o|t)$, связанных документов $p(d'|t)$, рекламных баннеров $p(b|t)$,



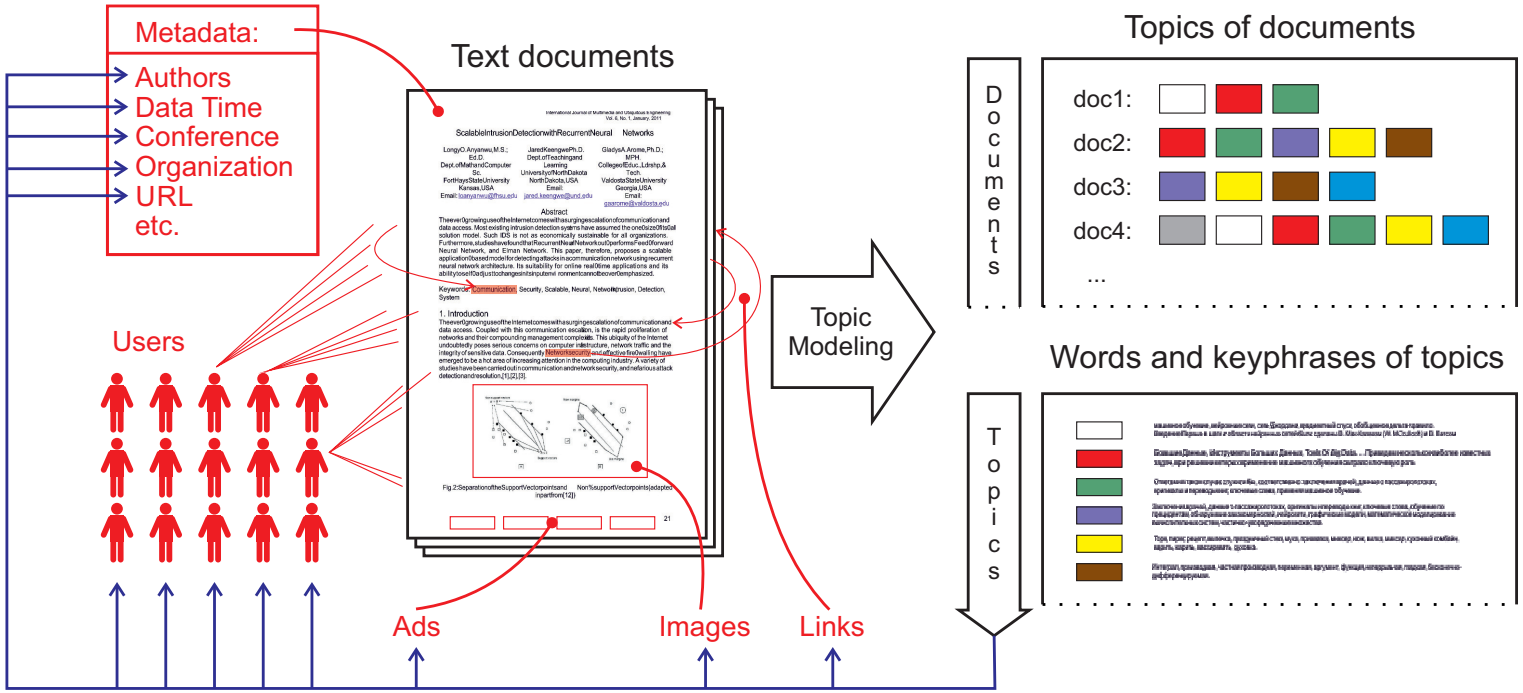
Мультимодальные тематические модели

M-APTM ищет распределения по темам $p(w|t)$, авторов $p(a|t)$, меток времени $p(y|t)$, объектов на картинках $p(o|t)$, связанных документов $p(d'|t)$, рекламных баннеров $p(b|t)$, пользователей $p(u|t)$,



Мультимодальные тематические модели

M-APTM ищет распределения по темам $p(w|t)$, авторов $p(a|t)$, меток времени $p(y|t)$, объектов на картинках $p(o|t)$, связанных документов $p(d'|t)$, рекламных баннеров $p(b|t)$, пользователей $p(u|t)$,
и объединяет все модальности в одну тематическую модель.



Мультимодальные тематические модели

Правдоподобие для мультимодальной модели примет вид

$$Q(\Phi,\Theta)=\sum_{j=1}^{m}\tau^j\sum_{d\in D}\sum_{x\in X_j}n_{dx}^j\ln\sum_{t\in T}\phi_{xt}^j\theta_{td}\to\max$$

τ^j — коэффициент, регулирующий значимость $j-$й модальности в модель.

Теорема

Если функция $R(\Phi,\Theta)$ стохастических матриц $\Phi^j, j=1,\cdots,m$, Θ непрерывно дифференцируема и (Φ,Θ) является точкой локального максимума функции $Q(\Phi,\Theta)+R(\Phi,\Theta)$, то для любой темы t и документа d выполняется система уравнений:

$$p_{tdw}^j=\frac{\phi_{xt}^j\theta_{td}}{\sum_{s\in T}\phi_{xs}^j\theta_{sd}};\quad n_{xt}^j=\sum_{d\in D}n_{dx}^j p_{tdx}^j;\quad n_{td}^j=\sum_{x\in d,x\in W^j}n_{dx}^j p_{tdx}^j;$$

$$\phi_{xt}^j\propto\left(n_{xt}^j+\phi_{xt}^j\frac{\partial R}{\partial\phi_{xt}^j}\right)_+;\quad \theta_{td}\propto\left(\sum_{j=1}^{m}\tau^j n_{td}^j+\theta_{td}\frac{\partial R}{\partial\theta_{td}}\right)_+.$$

BigARTM: особенности

BigARTM [2] — библиотека с открытым кодом для построения мультимодальных регуляризованных тематических моделей больших коллекций текстов.

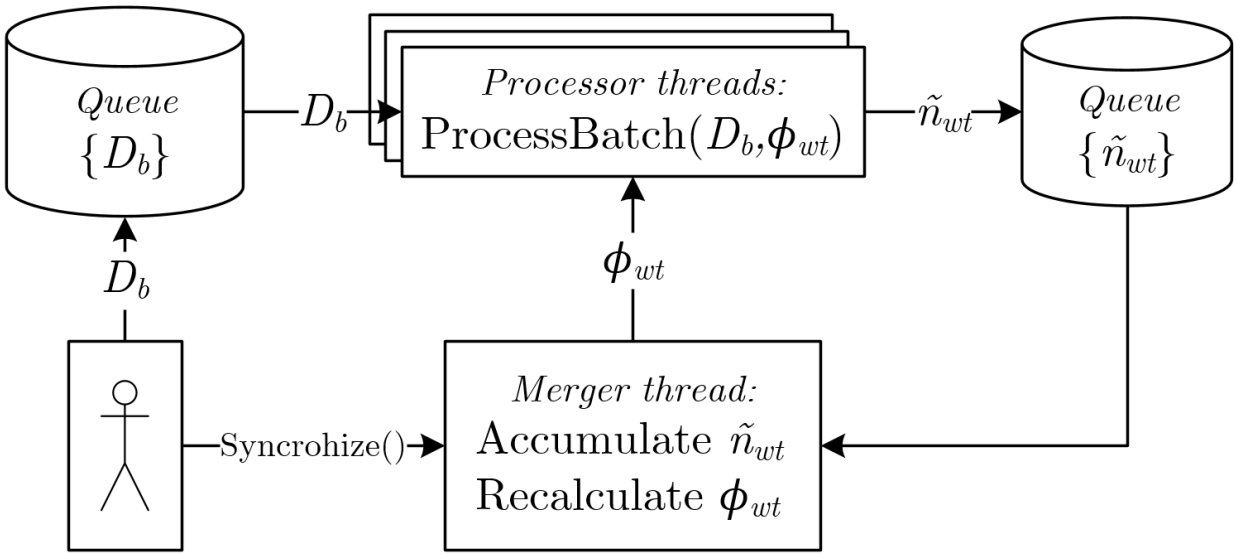
1. Эффективная параллельная и онлайновая обработка данных.
2. Кросплатформенность (Windows, Linux, Mac OS X).
3. Ядро на C++, API на C++ и Python.

[2] http://bigartm.org

BigARTM: архитектура

Параллелизм BigARTM основан на многопоточности.

- Множество *обработчиков* (Processor).
- Один *поток слияния* (Merger)
- Очередь заданий, откуда обработчики берут данные.
- Очередь слияния, в которую обработчики помещают результаты, и из которой их забирает поток слияния.



BigARTM vs. VW.LDA vs. Gensim

Библиотека	Число процессоров	Время обучения модели [3]
BigARTM	1	35 минут
LdaModel	1	369 минут
VW.LDA	1	73 минуты
BigARTM	4	9 минут
LdaMulticore	4	60 минут
BigARTM	8	4.5 минуты
LdaMulticore	8	57 минут

[3] Использовалась коллекция документов английской Википедии, $|D| \approx 3.7 \times 10^6$. Вычисления производились в облаке Amazon EC2.

Реализация механизма регуляризации

1. Механизм регуляризации реализован как для матрицы Θ, так и для Φ.
2. В первом случае регуляризирующие поправки вычисляются для каждого документа (т. е. для вектор-столбцов Θ) во время его обработки.
3. Регуляризация Φ производится каждый раз при её обновлении.
4. Каждый регуляризатор — плагин на C++, пользователь всегда может написать новый и добавить его в библиотеку по описанной схеме.

Список регуляризаторов

Реализованы следующие регуляризаторы:

- **Сглаживание и разреживание матрицы** Θ. Сглаживает/разреживает θ_d распределением α. Формула M-шага:

$$\theta_{td} \propto (n_{td} + \alpha_0 \alpha_t)_+$$

- **Сглаживание и разреживание матрицы** Φ. Сглаживает/разреживает ϕ_t распределением β. Формула M-шага:

$$\phi_{wt} \propto (n_{wt} + \beta_0 \beta_w)_+$$

- **Декоррелирование тем в матрице** Φ. Разреживает темы, делая их как можно более различными. Формула M-шага:

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

- **Label Regularization** Φ. Приближает оценку распределения слов, полученную моделью, к частотным, полученным по коллекции. Пусть $\hat{p}_w$ — частота слова w в выборке. Формула M-шага:

$$\phi_{wt} \propto \left(n_{wt} + \tau \hat{p}_w \frac{\phi_{wt} n_t}{\sum_{s \in T} \phi_{ws} n_w} \right)_+ .$$

Реализация мультимодальных моделей

1. Модифицирован способ хранения матрицы Φ.
2. Реализовано хранение метки модальности для каждого слова из словаря.
3. Реализована раздельная нормировка подматриц Φ^j по модальностям.
4. Реализован API для регуляризаторов отдельных модальностей.
5. Сделанные модификации не ухудшили производительность.

Описание эксперимента с регуляризацией

- Производилось сравнение моделей LDA и ARTM.
- Эксперименты ставились на коллекции Wikipedia, $\approx 3.7 \times 10^6$ документов, объём словаря $\approx 2 \times 10^5$ слов.
- Онлайн-алгоритм (один проход по коллекции)

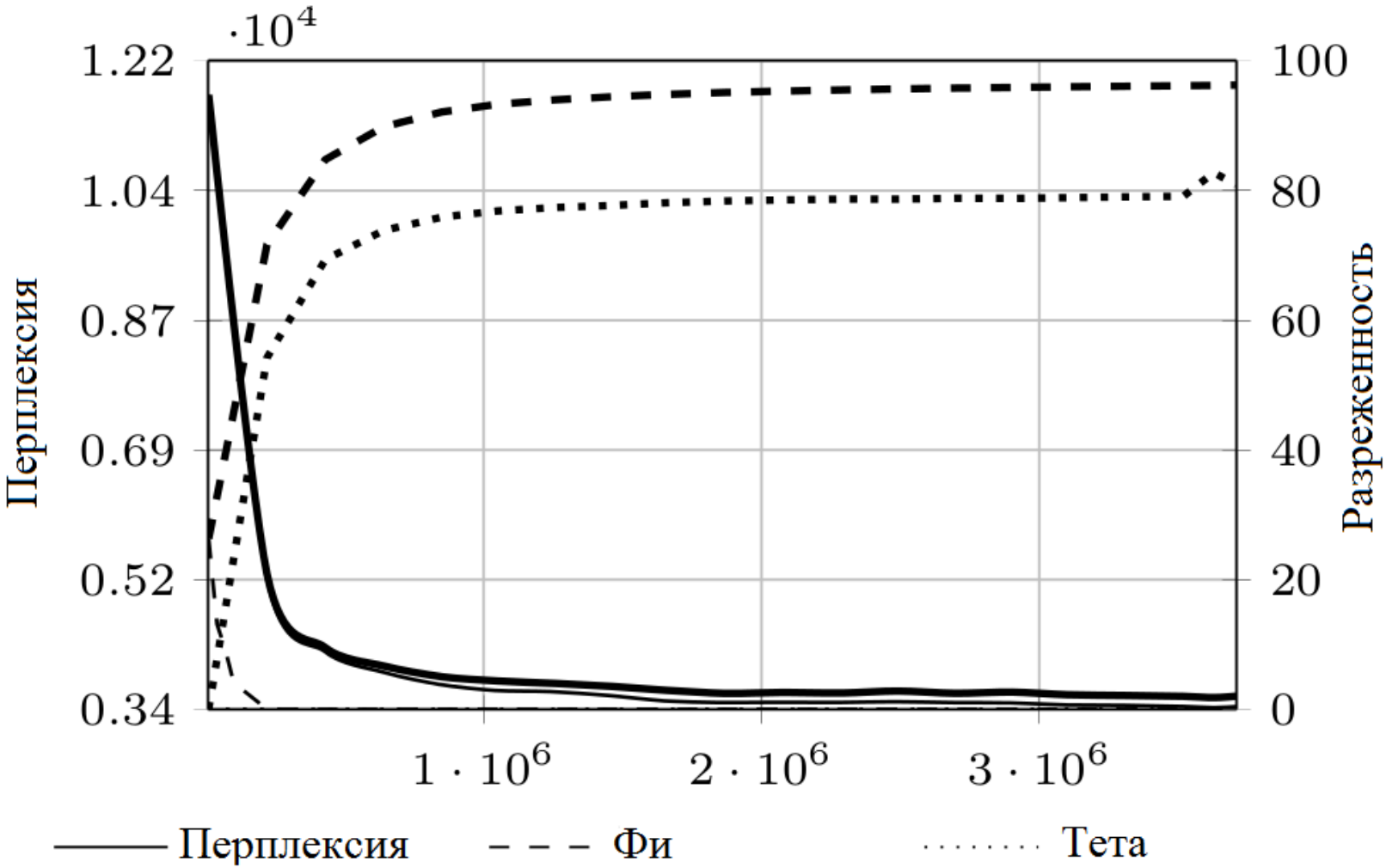
Метрики качества

- Перплексия

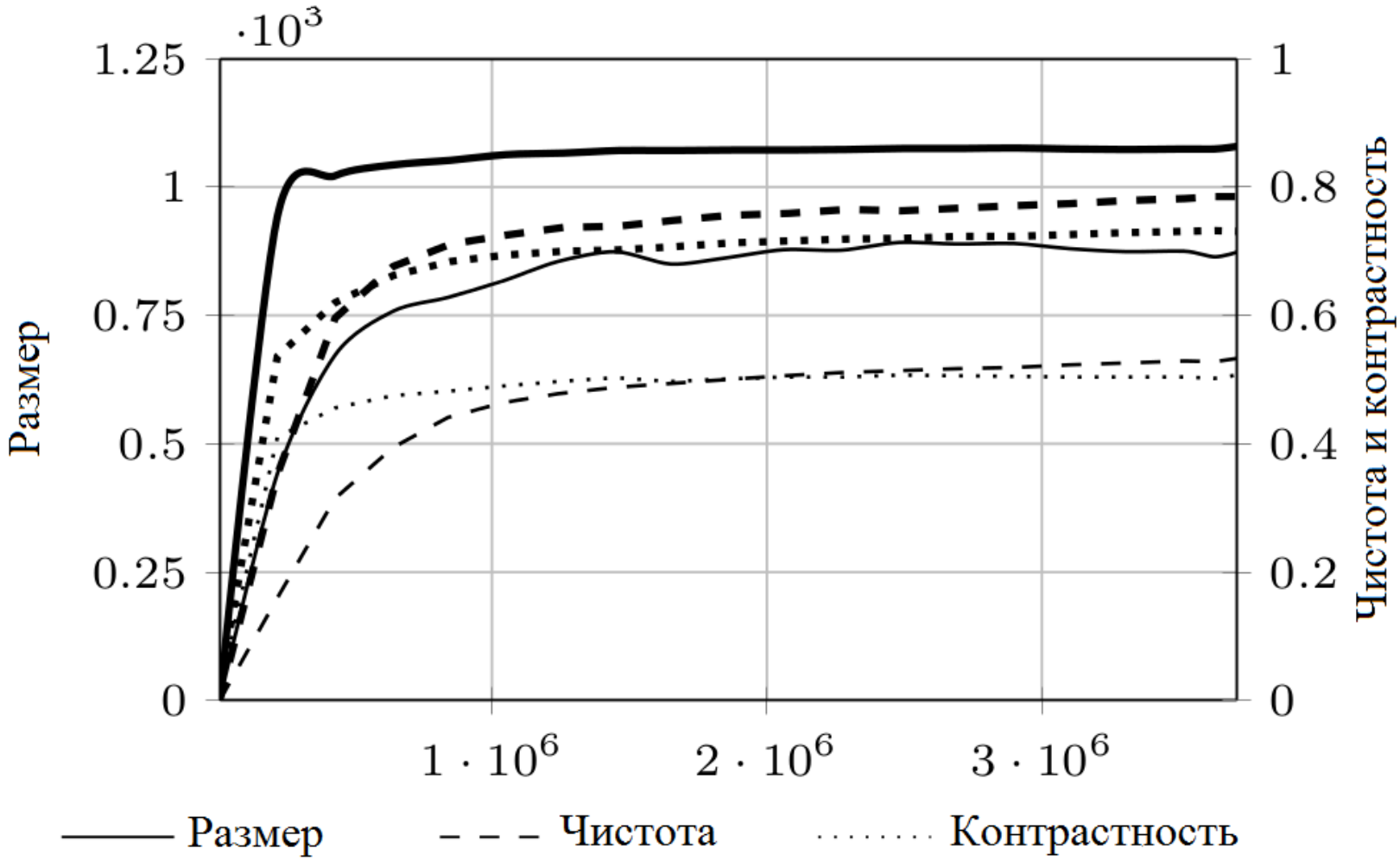
$$\mathcal{P}(D) = \exp\left(-\frac{1}{n}\mathcal{L}(\Phi,\Theta)\right) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw}\ln p(w|d)\right)$$

- Разреженности матриц Φ и Θ
- Характеристики *ядер тем*: если для слова вероятность $p(t|w)$ больше некоторого порога в теме t — относим его к ядру W_t этой темы. Характеристики:
  1. Размер: $|W_t|$
  2. Чистота: $\sum_{t\in W_t} p(w|t)$
  3. Контрастность: $|W_t|^{-1}\sum_{t\in W_t} p(t|w)$

Результаты: перплексия и разреженности



Результаты: характеристики ядер тем



Описание эксперимента с классификацией

- При наличии в документах модальности меток классов можно обучить тематическую модель классификации.
- В общем случае тематическим моделям пока не удаётся обойти традиционные алгоритмы классификации, такие как SVM.
- Но в случае сложных задач с большим числом несбалансированных, пересекающихся и взаимнозависимых классов ситуация меняется [4].
- В данном эксперименте используется коллекция EUR-lex. Параметры (после предобработки): $\approx$ 19000 документов, $\approx$ 21000 слов в словаре, $\approx$ 3200 классов.
- В статье [Rubin 2012] производилось сравнение алгоритма Dependency LDA и SVM. Сравниваемся с обоими по метрикам из статьи.

[4] «T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers. Statistical topic models for multi-label document classification, 2012».

Метрики и результаты

Используемые метрики качества классификации:

- Площадь под кривой precision-recall ($\mathrm{AUC}_{\mathrm{PR}}$).
- Площадь по ROC-кривой ($\mathrm{AUC}_{\mathrm{ROC}}$).
- Доля документов, у которых самая вероятная метка неверная (One Error).
- Доля документов, не классифицированных полностью правильно (Is Error).

	$\lvert T\rvert_{optimal}$	$\mathrm{AUC}_{\mathrm{PR}} \Uparrow$	$\mathrm{AUC}_{\mathrm{ROC}} \Uparrow$	One Error $\Downarrow$	Is Error $\Downarrow$
ARTM	10000	**0.513**	0.980	**29.1**	**95.5**
DLDA	200	0.492	**0.982**	32.0	97.2
$\mathrm{SVM}_{\mathrm{best}}$	$\approx$ 21000	0.435	0.975	31.6	98.1

Выводы

- APTM даёт возможность легко строить тематические модели, удовлетворяющие большому количеству ограничений.
- M-APTM позволяет строить тематические модели, учитывающие всю необходимую метаинформацию.
- BigARTM — параллельная реализация онлайнового EM-алгоритма для построения мультимодальных тематических моделей с возможностью комбинирования регуляризаторов.
- В экспериментах на больших коллекциях BigARTM показывает высокую производительность и качество.