# Энтропийный регуляризатор отбора тем в вероятностных тематических моделях

Плавин Александр

Московский физико-технический институт, Москва

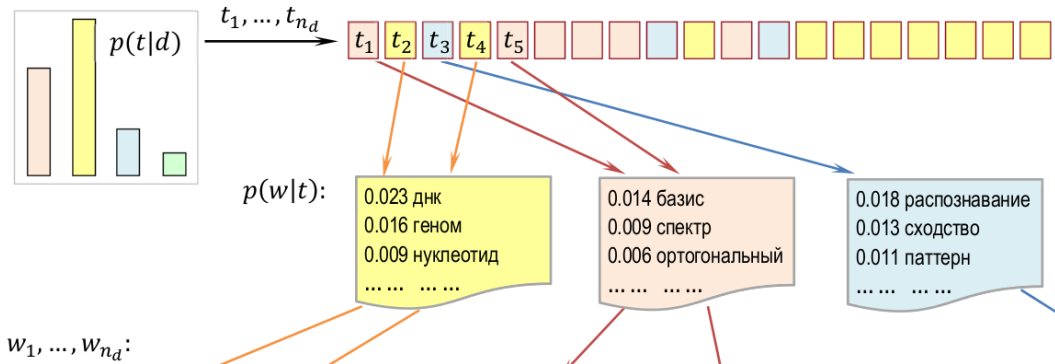

Светлогорск • 23 сентября 2015 года

# Задача выявления тем в коллекции документов

**Дано:**

$n_{dw}$ — число вхождений слова $w \in W$ в документ $d \in D$.

Каждое вхождение порождается некоторой неизвестной темой:




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).

# Задача выявления тем в коллекции документов

**Найти:**

- $T$ — множество тем,

распределения:

- $\Theta \equiv \{\theta_{td}\} \equiv \{p(t|d)\}$ — тем в документах,
- $\Phi \equiv \{\phi_{wt}\} \equiv \{p(w|t)\}$ — слов в темах,

такие, что:

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

Максимизация правдоподобия в модели PLSA:

$$\mathscr{L}(\Phi, \Theta) = \prod_{d \in D, w \in d} p(w|d)^{n_{dw}} = \prod_{d,w} \left( \sum_{t \in T} \theta_{td}\phi_{wt} \right)^{n_{dw}} \to \max_{\Phi, \Theta}$$

# Проблема определения числа тем

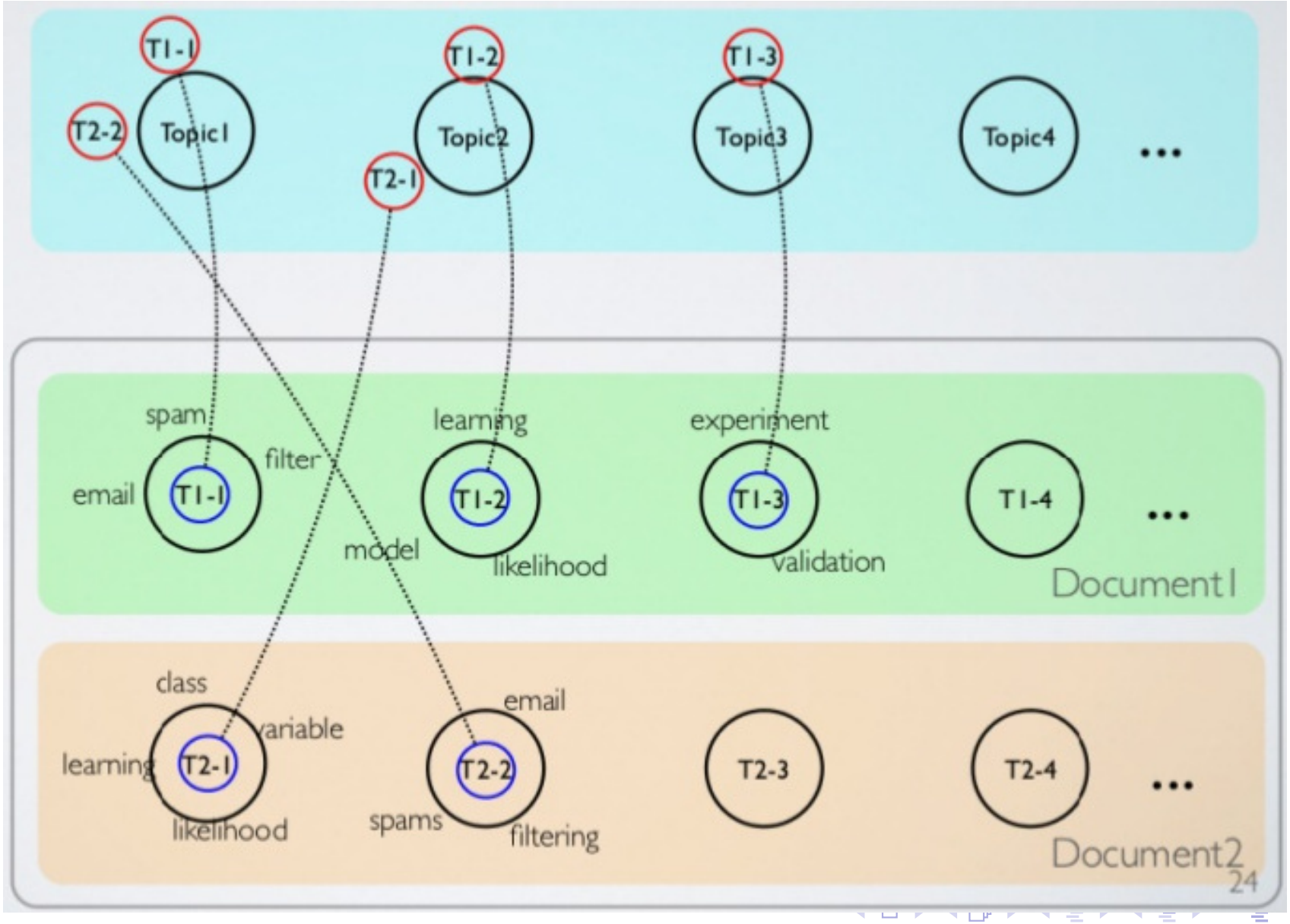
Число тем — задаваемый извне параметр.

Важен для интерпретируемости:

- Задано мало тем $\Rightarrow$ различные темы сливаются вместе.
- Задано много тем $\Rightarrow$ появляются дубликаты, комбинации уже имеющихся.

HDP — иерархические процессы Дирихле — популярный байесовский подход к определению числа тем.

# HDP: франшиза китайских ресторанов



# Базовый метод: ARTM

Подход ARTM *(аддитивная регуляризация тематических моделей)* — максимизация *регуляризованного* логарфима правдоподобия:

$$\ln \mathscr{L}(\Phi, \Theta) + \sum_i \tau_i R_i(\Phi, \Theta) \to \max_{\Phi,\Theta}.$$

Здесь:

- $R_i(\Phi, \Theta)$ — регуляризаторы, задающие дополнительные требования к модели,
- $\tau_i$ — коэффициенты регуляризации, устанавливающие баланс между этими требованиями.

# Обучение модели: EM-алгоритм

- **E-шаг** — формула Байеса:

$$p(t|d,w) = norm_t\left(p(w|t)p(t|d)\right) = norm_t\left(\phi_{wt}\theta_{td}\right)$$

- **M-шаг** — принцип максимума правдоподобия:

$$\phi_{wt} = norm_w\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right), \quad \text{где } n_{wt} = \sum_{d\in D} n_{dw}p(t|d,w)$$

$$\theta_{td} = norm_t\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \quad \text{где } n_{td} = \sum_{w\in W} n_{dw}p(t|d,w);$$

$$\text{здесь } norm_i(a_{ij}) = \frac{(a_{ij})_+}{\sum_i (a_{ij})_+}, \qquad a_+ = \max(a,0).$$

# Предлагаемый метод: регуляризатор в ARTM

Будем максимизировать расстояние (KL-дивергенцию) между равномерным распределением $p_U(t) = \frac{1}{|T|}$ и модельным $p(t)$:

$$R(\Phi, \Theta) = KL\left(p_U \| p\right) = KL\left(\frac{1}{|T|} \middle\| \frac{n_t}{n}\right).$$

Формулы M-шага:

$$\phi_{wt} = norm_w(n_{wt}), \qquad \theta_{td} = norm_t\left(n_{dt}\left(1 - \tau \frac{n}{|T|}\frac{1}{n_t}\right)\right),$$

где $n_t = \sum_d n_{td}$ — число слов, отнесённых моделью к теме $t$.

# Эксперимент: набор данных

**Исходная коллекция (обозначим $n^1_{dw}$):**

Статьи с конференции NIPS:

$|D| = 1740, |W| \approx 1.3 \cdot 10^4, n \approx 2.3 \cdot 10^6$.

**Синтетические данные (обозначим $n^0_{dw}$):**

На основе $\Phi$, $\Theta$ модели PLSA с 50 темами на данных NIPS:

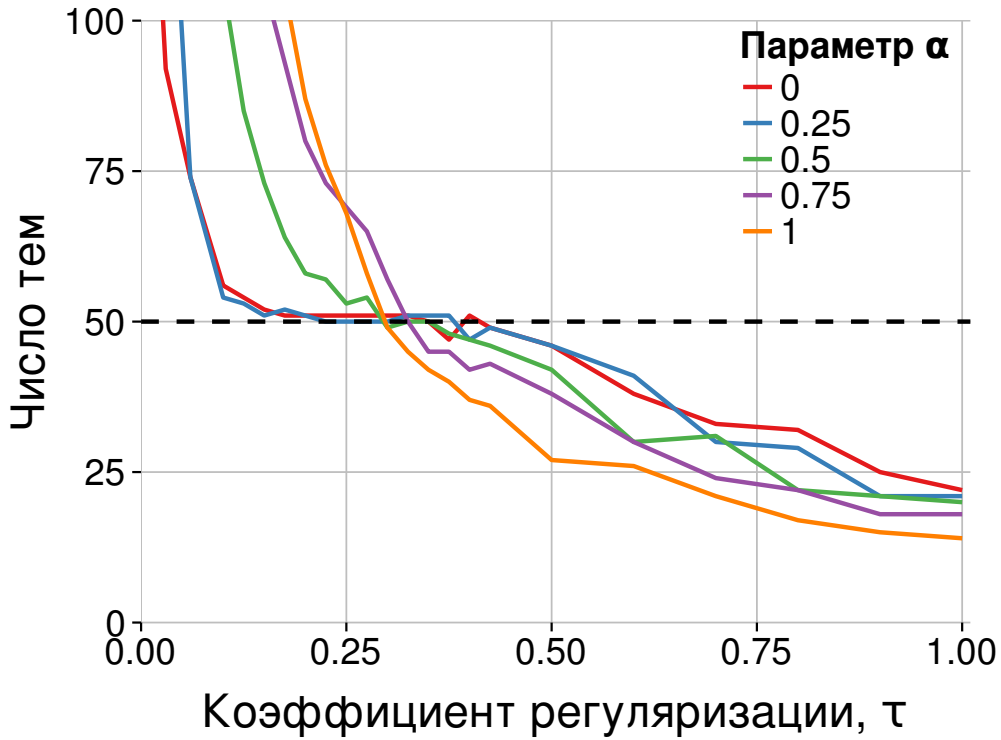
$$n^0_{dw} = n_d \cdot p(w|d) \equiv n_d \cdot (\Phi\Theta)_{wd}.$$

**Параметрическое семейство смешанных данных:**

Для $\alpha \in [0, 1]$ определим $n^\alpha_{dw} = \alpha n^1_{dw} + (1 - \alpha) n^0_{dw}$ — смешанные данные.

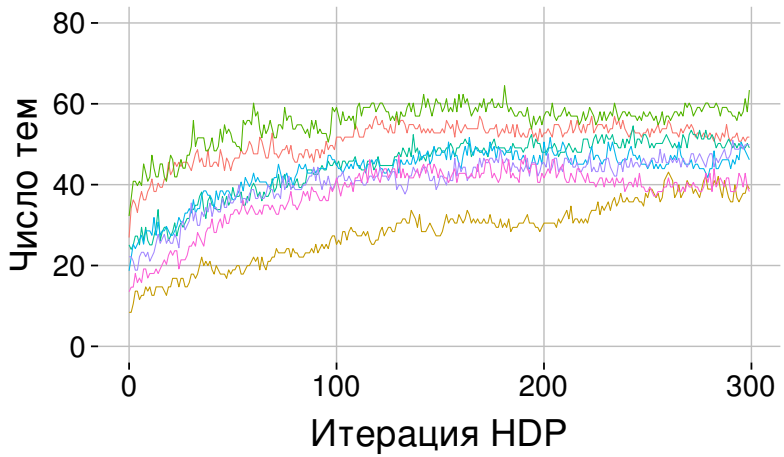
# Эксперимент: определение истинного числа тем

Получаемое число тем при различных значениях параметра $\alpha$ и коэффициента регуляризации $\tau$:

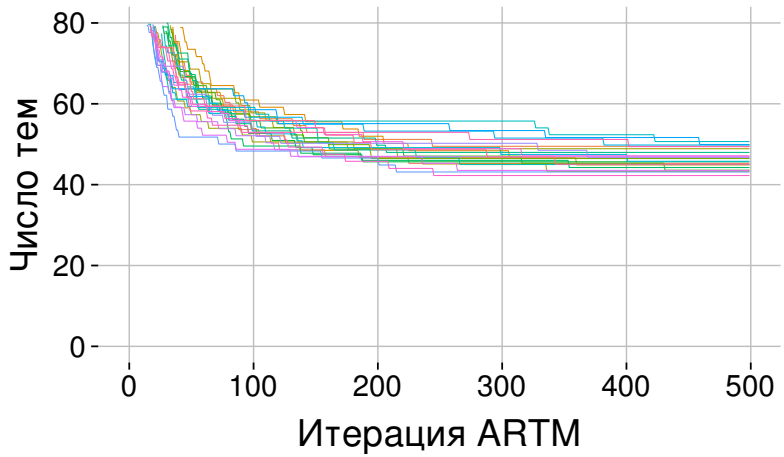


# Эксперимент: устойчивость получаемых значений

Для фиксированных значений параметров $\eta$ и $\tau$:




(a) HDP

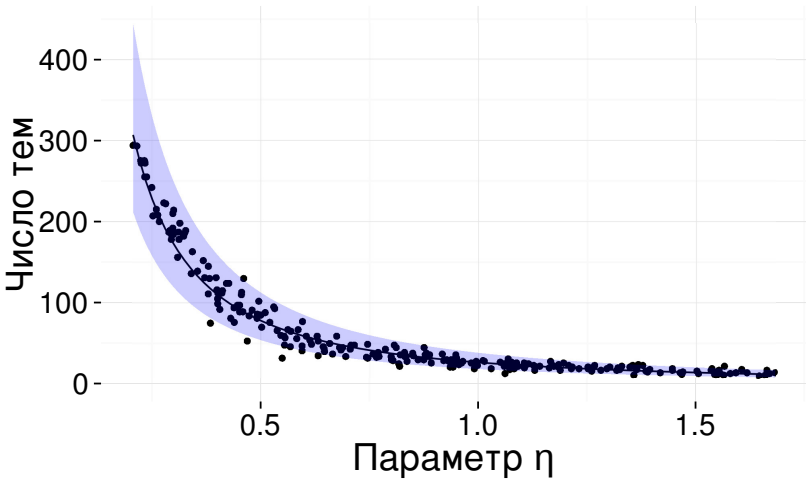



(b) ARTM

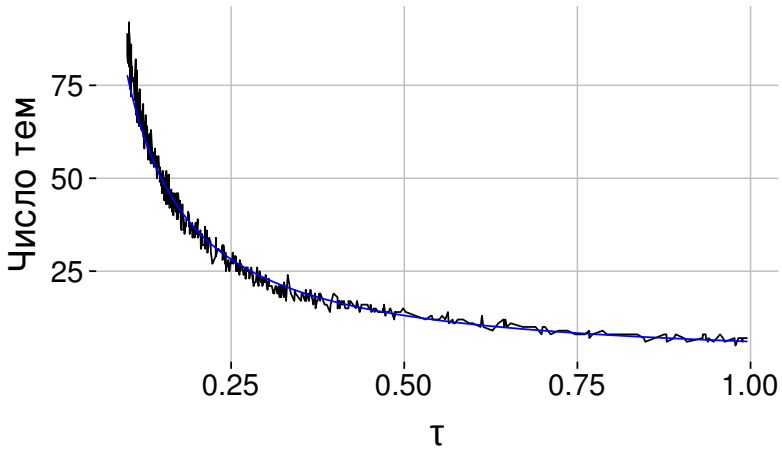
ARTM с предлагаемым регуляризатором даёт меньший разброс числа тем.

# Эксперимент: устойчивость получаемых значений

Для различных значений параметров $\eta$ и $\tau$:




(c) HDP




(d) ARTM

Определяемое HDP число тем также зависит от параметра алгоритма.

# Происходит ли удаление лишних тем?

Виды избыточных тем в модели:

- Выпуклые комбинации нескольких других: $\phi_{wt'} = \sum_t \alpha_{tt'}\phi_{wt}$, некоторым образом распределённые по документам.
- Расщеплённые темы: $\phi_{wt} = \sum_{t'} \alpha_{tt'}\phi_{wt'}$, одинаковым образом используемые в документах.
- Дубликаты: повторяющиеся идентичные темы.

# Эксперимент: происходит ли удаление лишних тем?

**Гипотеза**

Предлагаемый регуляризатор в ARTM удаляет все описанные виды лишних тем.

Добавим в синтетическую коллекцию $n_{dw}^0$ искусственные темы:

- Выпуклые комбинации исходных: $\phi_{wt'} = \sum_t \alpha_{tt'}\phi_{wt}$, строки $\theta_{t'}$ — случайные.
- Расщеплённые исходные: $\phi_{wt} = \sum_{t'} \alpha_{tt'}\phi_{wt'}$, строки $\theta_{t'}$ — совпадающие с $\theta_t$.

# Происходит ли удаление комбинаций тем?

Пусть:

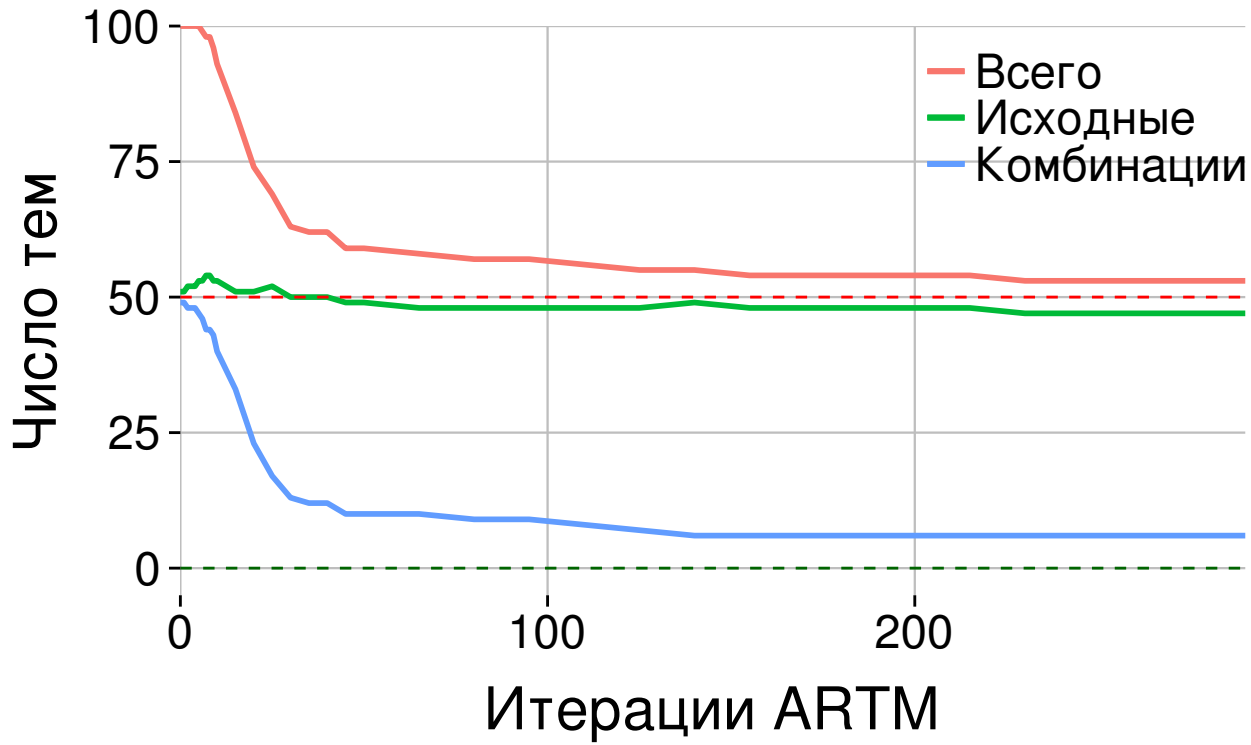
- Коллекция в точности соответствует порождающей модели: $n_{dw} = \frac{n}{|D|} \sum_{t \in T} \phi_{wt} \theta_{td}$.
- Для каждой пары тем есть документ, содержащий только эти темы.
- Количество документов, содержащих некоторые $K$ тем, находится между $c^K |D|$ и $C^K |D|$.
- В восстановленных темах есть только исходные темы или их выпуклые комбинации: $\hat{\Phi} = A\Phi$, причём $|\hat{T}| < 2|T|$ и все ненулевые $A_{ij} > \alpha > 0$.

Тогда, если предлагаемый регуляризатор оставит $|T|$ тем, то в них будет не более $|T| \frac{C^2}{\alpha c - C^2}$ комбинаций.

Например, при $c = \frac{1}{1000}$, $C = \frac{2}{1000}$, $\alpha = 0.1$: $\frac{C^2}{\alpha c - C^2} \approx 0.04$.

# Эксперимент: удаление комбинаций тем

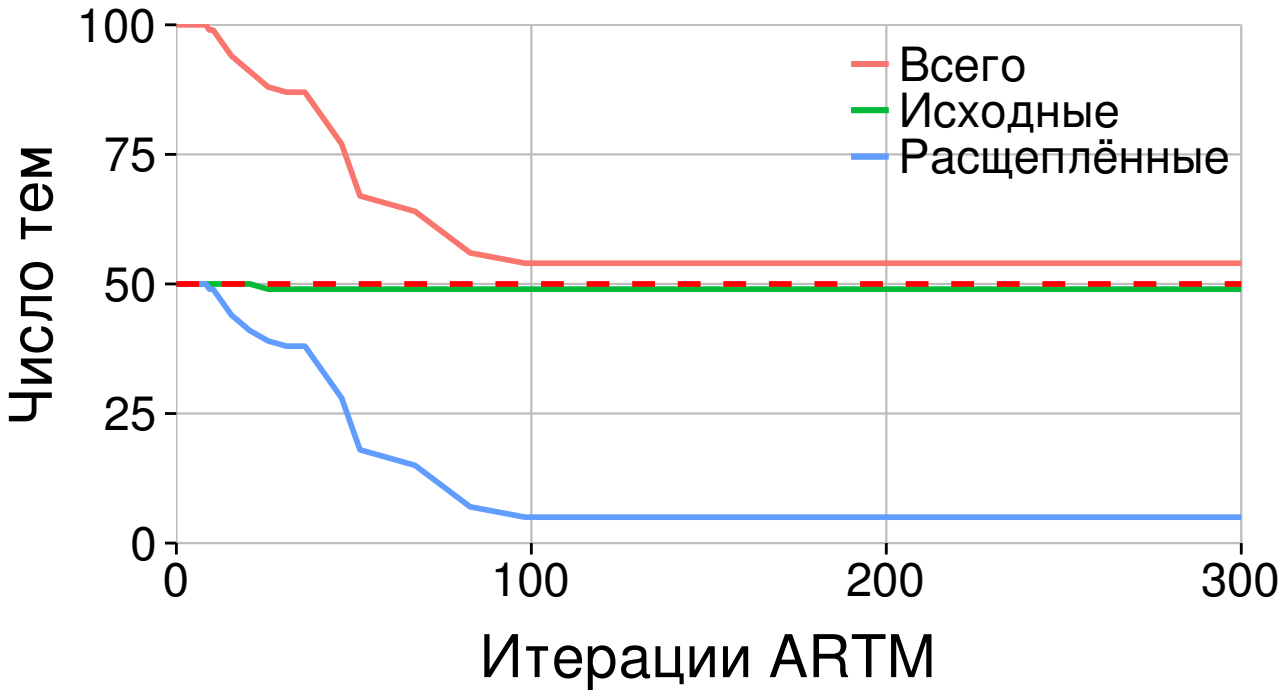
Набор данных: синтетическая коллекция $n_{dw}^0$ + добавленные 50 выпуклых комбинаций, в каждой по 5 исходных тем.




Удаляются преимущественно темы-комбинации.

# Эксперимент: удаление расщеплённых тем

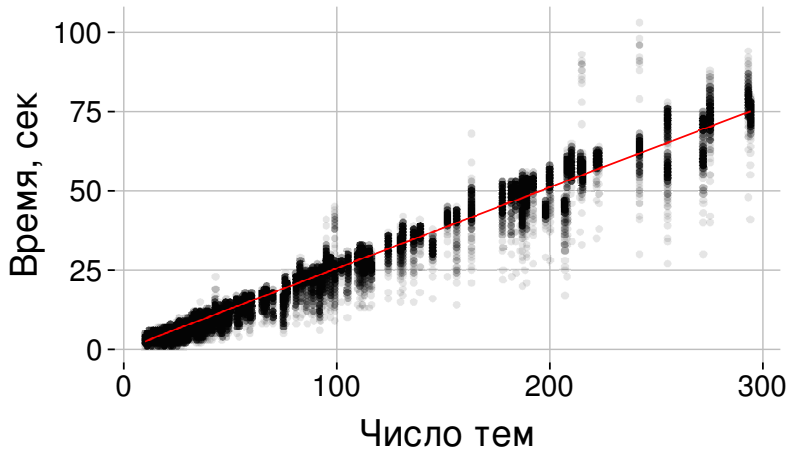
Набор данных: синтетическая коллекция $n^0_{dw}$ + добавленные 50 расщеплённых тем — расщепление 10 исходных на 5 долей.



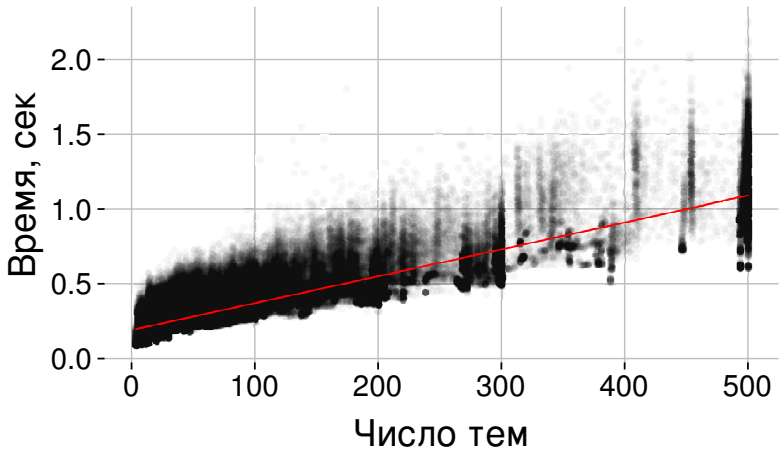

Удаляются преимущественно расщеплённые темы.

# Эксперимент: время работы

Набор данных: статьи с конференции NIPS с $|D| = 1740, |W| \approx 1.3 \cdot 10^4, n \approx 2.3 \cdot 10^6$.




(e) HDP




(f) ARTM

Рис.: Время работы одной итерации

Например, при 200 темах и 500 итерациях прирост скорости около 100 раз: 7 часов для HDP против 4.5 минут для ARTM.

# Результаты

Предложен метод последовательного отбора тем для модели ARTM,

- который определяет число тем устойчивее и быстрее, чем стандартный метод HDP,
- удаляет в первую очередь комбинации тем и расщеплённые темы,
- позволяет находить истинное число тем, если оно существует.

## Направления дальнейших исследований

- Стратегия автоматического выбора значения коэффициента регуляризации $\tau$.
- Обобщение метода: добавление новых тем в модель.
- Поочерёдное применение шагов отбора тем и добавления новых (Add-Del).

# Публикации

- *Плавин А.В.* Оптимизация числа тем в вероятностных тематических моделях с помощью регуляризатора строкового разреживания // Конференция МФТИ, 2014.
- *Плавин А.В.* Отбор тем в вероятностных тематических моделях // Ломоносов-2015, МГУ.
- *Vorontsov K. V., Potapenko A. A., Plavin A. V.* Additive Regularization of Topic Models for Topic Selection and Sparse Factorization // The Third International Symposium On Learning And Data Sciences (SLDS 2015). April 20-22, 2015. Royal Holloway, University of London, UK.