Отбор тем в вероятностных тематических моделях

Плавин Александр

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф-м.н. К. В. Воронцов

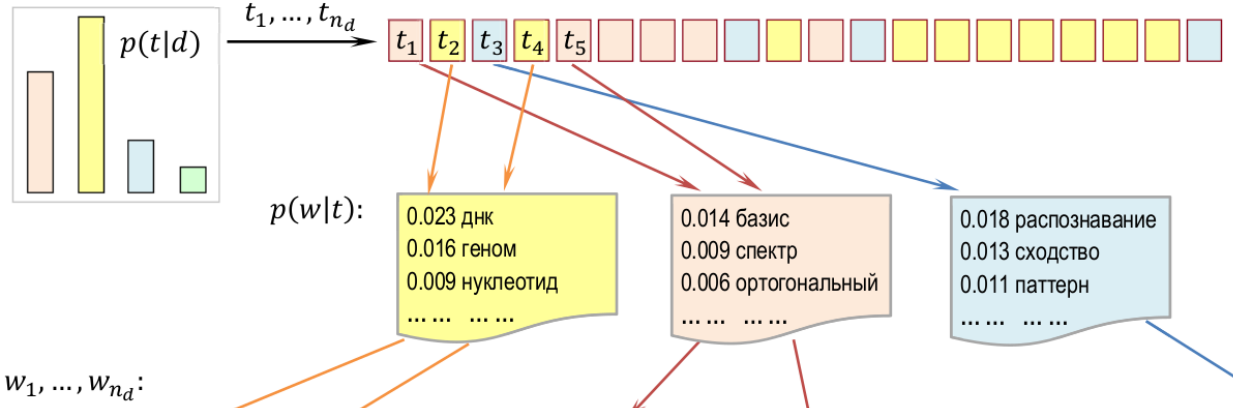
23 июня 2015 года

План

Задача выявления тем в коллекции документов

Дано:

Коллекция текстовых документов: n_{dw} — число вхождений слова $w \in W$ в документ $d \in D$. Каждое вхождение каждого слова 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:

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

Решение: PLSA

Максимизация правдоподобия:

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

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 подход к определению числа тем.

Однако,

- введение дополнительных требований к модели затруднено,
- число тем определяется им неустойчиво.

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,
- τ_i — коэффициенты регуляризации, устанавливающие баланс между этими требованиями.

Обучение модели: EM-алгоритм

- **E-шаг** — формула Байеса:

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

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

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

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

Предлагаемый метод: регуляризатор в ARTM

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

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

Формулы M-шага:

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

Набор данных

Исходная коллекция:

Статьи с конференции NIPS (обозначим n^1_{dw}):

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

Синтетические данные:

На основе сгенерированной простой модели коллекции n^1_{dw} с 50 темами:

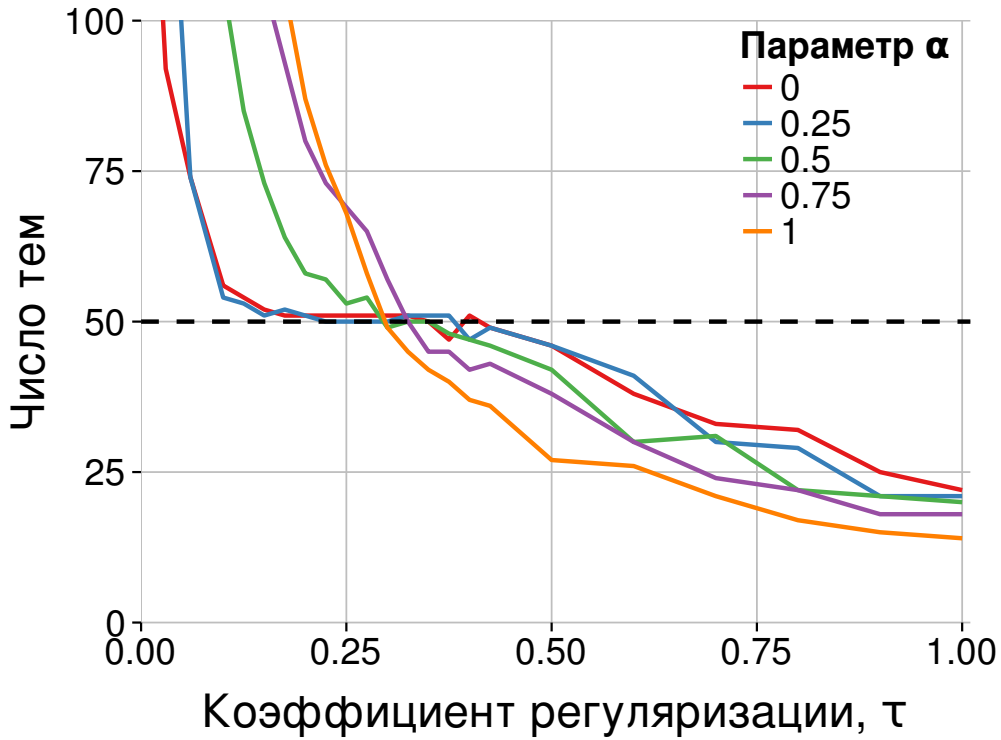
$$n^0_{dw} = n_d \cdot p(w|d) \equiv n_d \cdot (\Phi\Theta)_{wd}.$$

Параметрическое семейство смешанных данных:

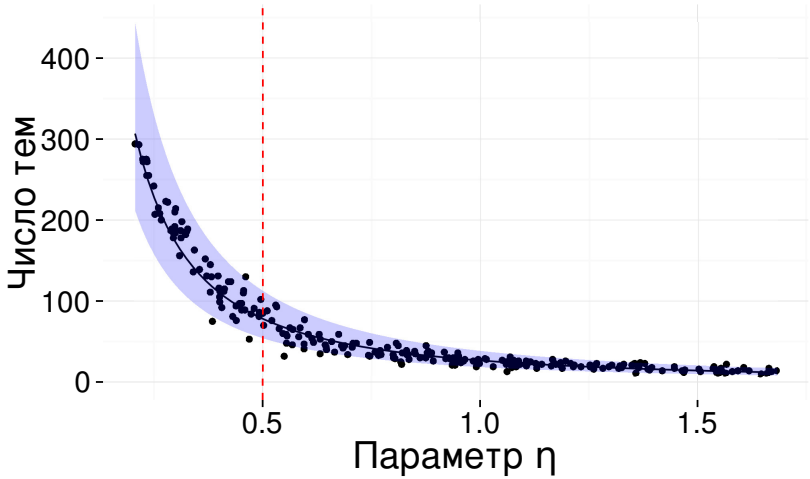
Для $\alpha \in [0, 1]$ определим $n^\alpha_{dw} = \alpha n^1_{dw} + (1 - \alpha) n^0_{dw}$ — смешанные данные.

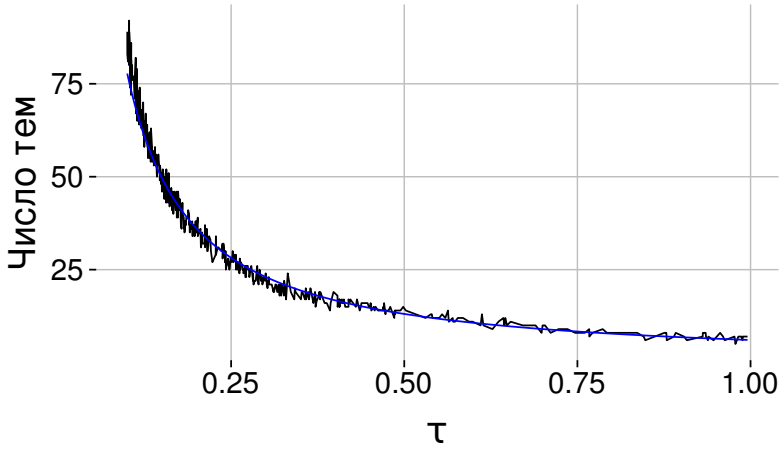
Определение истинного числа тем

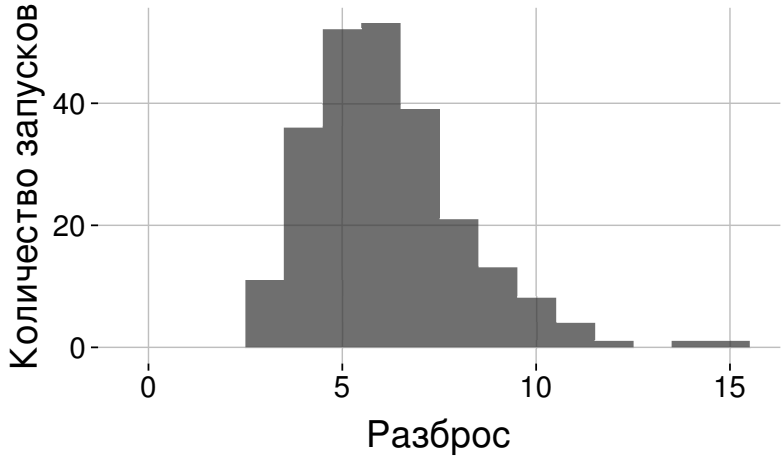
Получаемое число тем при различных значениях параметра α и коэффициента регуляризации τ:

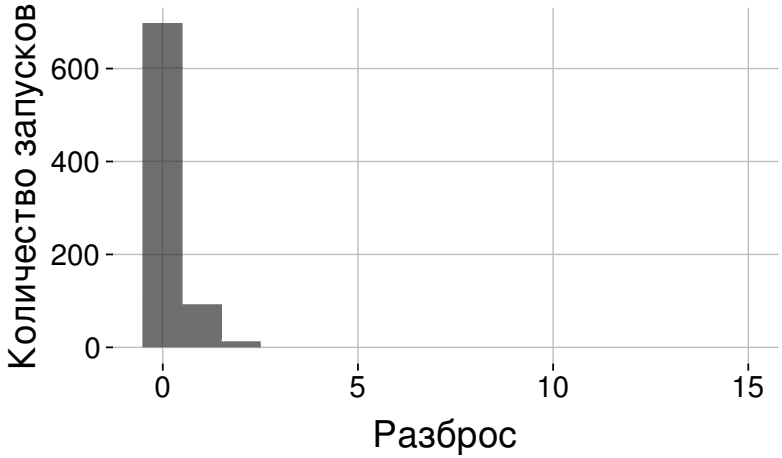


Устойчивость получаемых значений









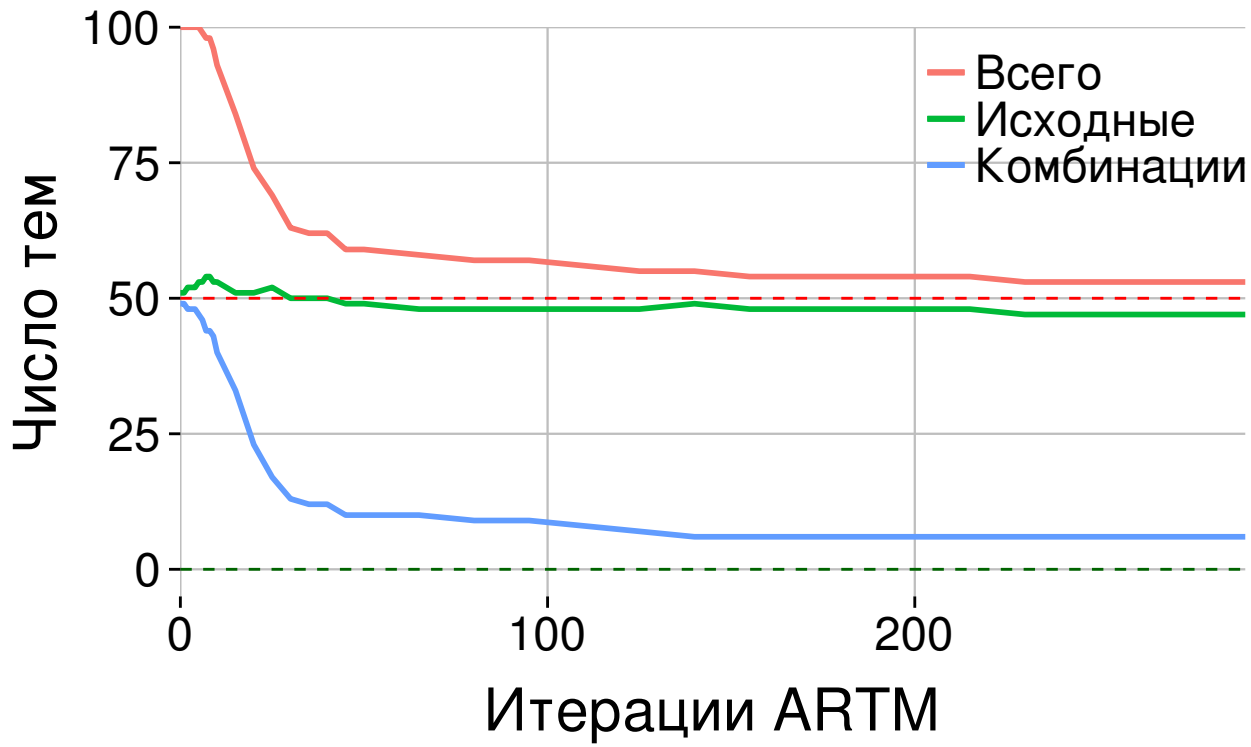
(a) HDP (b) ARTM

Удаление линейно зависимых тем

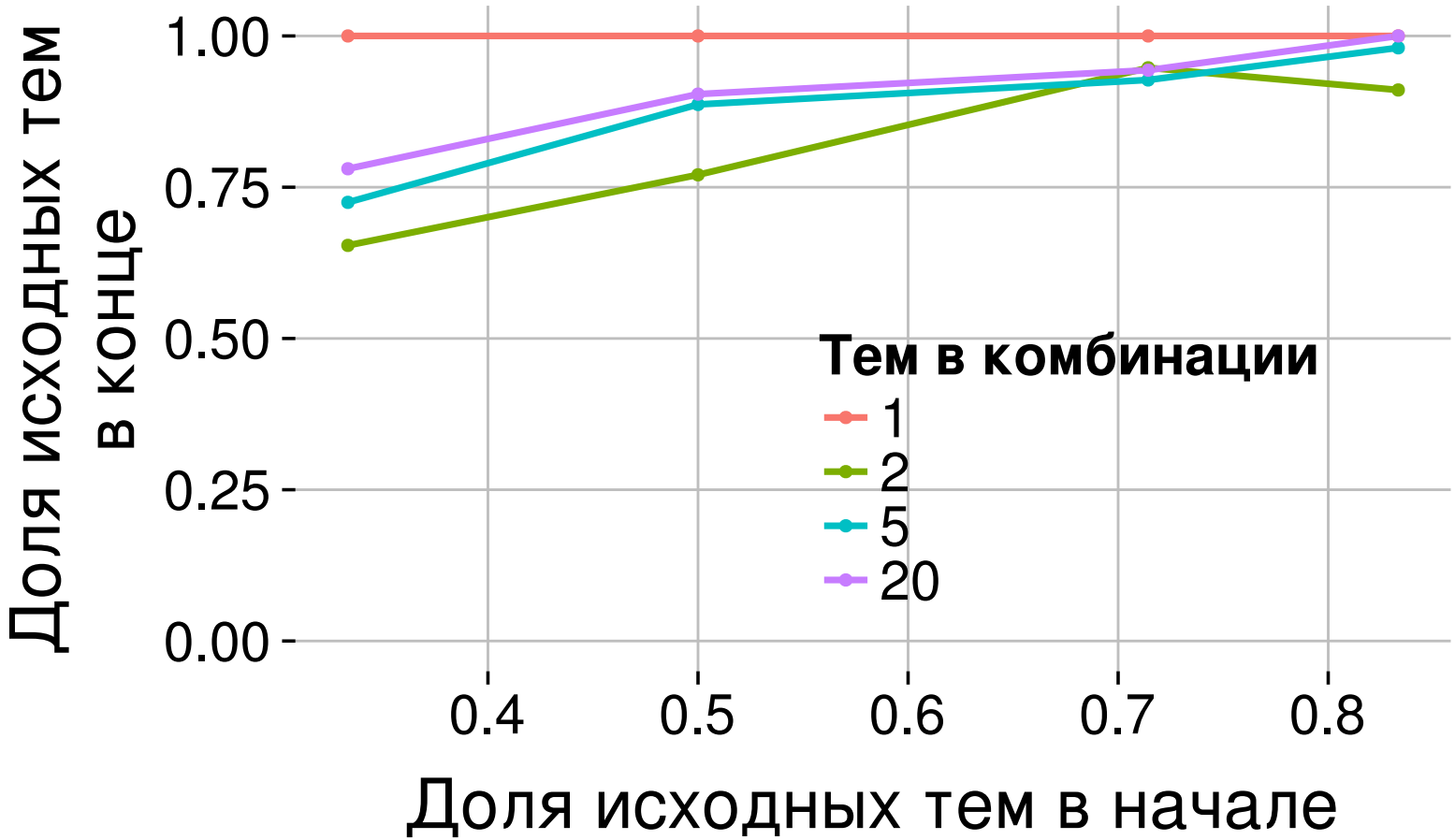
Данные:

Синтетическая коллекция n_{dw}^0 + добавленные линейные комбинации тем.

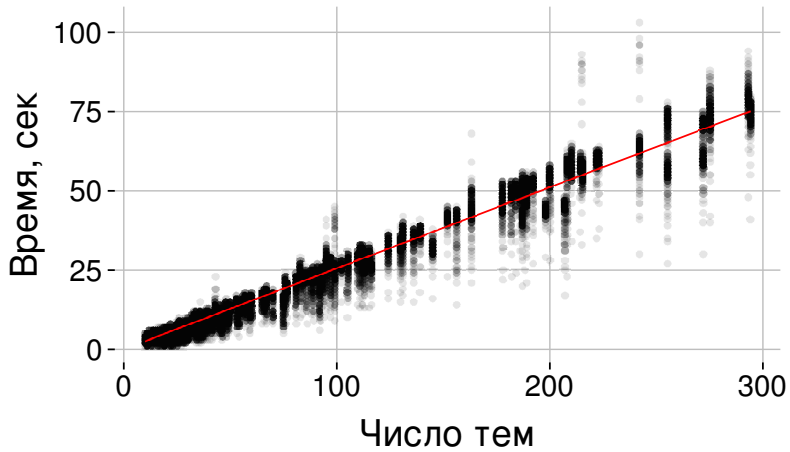
Выбран оптимальный коэффициент регуляризации τ.



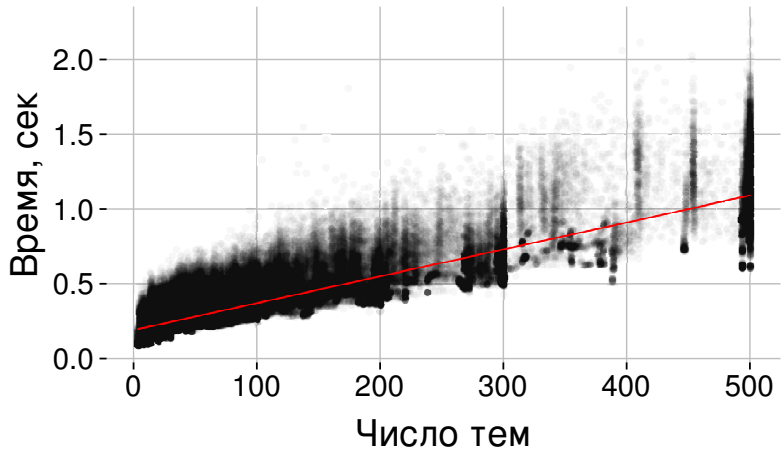
Удаление линейно зависимых тем



Время работы



(c) HDP



(d) ARTM

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

Результаты, выносимые на защиту

- Предложен регуляризатор последовательного отбора тем для модели ARTM.
- Показано, что он определяет число тем намного устойчивее и значительно быстрее, по сравнению со стандартным методом HDP.
- Показано, что он удаляет в первую очередь комбинации тем и расщеплённые темы.
- Показано, что он позволяет определять истинное число тем, если оно существует.

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.