Московский государственный университет имени М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

# Адаптивная регуляризация вероятностных тематических моделей

**Выполнил:**
студент 417 группы
Дойков Никита Владимирович

**Научный руководитель:**
д.ф-м.н. Воронцов Константин Вячеславович

Москва, 2015

# Содержание

**Аннотация**

Данная работа посвящена вероятностному тематическому моделированию — статистическому методу анализа текстовых документов.

В рамках подхода аддитивной регуляризации ARTM [27] предлагается непараметрическая темпоральная тематическая модель, учитывающая метки времени документов. Использование дополнительной априорной информации позволяет улучшать качество модели и визуализировать динамику тем во времени.

Описывается способ адаптивной нормировки коэффициентов регуляризации, упрощающий процесс настройки модели.

Предложенная модель тестируется на коллекции пресс-релизов министерств иностранных дел ряда стран за промежуток времени длиной более десяти лет.

# 1 Введение

Данная работа посвящена вероятностному тематическому моделированию — статистическому методу анализа коллекций текстовых документов. Для набора документов вводятся скрытые переменные — *темы*, которые описывают процесс порождения данных.

Многие тематические модели предполагают, что порядок документов в коллекции не важен, хотя в реальных задачах документы зачастую упорядочены. Ставится задача построения *темпоральной* тематической модели, где для каждого документа известна метка времени. Требуется использовать в модели эту дополнительную информацию в целях улучшения качества и визуализации динамики тем во времени.

Работа организована следующим образом.

Во втором разделе приводится описание базовой тематической модели PLSA [13]. Эксперименты на синтетических данных выявляют её недостатки. Описывается подход аддитивной регуляризации тематических моделей ARTM [27] с примерами регуляризаторов.

Третий раздел содержит перечисление основных метрик качества, используемых при построении тематических моделей.

Четвертый раздел посвящен темпоральным тематическим моделям. Предлагается непараметрическая модель, которая тестируется на реальной коллекции документов — пресс-релизах министерств иностранных дел ряда стран за промежуток времени длиной более десяти лет. Показан процесс настройки модели и визуализация тем.

В пятом разделе приводится способ адаптивной нормировки коэффициентов регуляризации. Предложенная репараметризация упрощает подбор коэффициентов на практике и позволяет сбалансировать воздействие регуляризаторов на отдельные темы и документы.

# 2 Тематические модели PLSA и ARTM

## 2.1 Постановка задачи

Тематическая модель коллекции документов PLSA (Probabilistic Latent Semantic Analysis) была предложена Томасом Хофманом в 1999 году [13].

Пусть имеется конечное множество слов $W$, называемое *словарем* и конечное множество документов $D$ — *коллекция*.

Для каждого документа $d \in D$ известно, какие слова он содержит и частоты вхождений всех слов. Таким образом, документ представляется в виде *«мешка слов»*.

Обозначим через $n_{dw}$ число вхождений слова $w$ в документ $d$, $n_d \equiv \sum\limits_w n_{dw}$ — длина документа $d$.

Предположим, что имеется также некоторое конечное множество скрытых переменных $T$ — *темы*.

На множестве $D \times W \times T$ введем вероятностное пространство с вероятностной мерой $p(d, w, t)$.

Примем *гипотезу условной независимости* — вероятность появления слова $w$, относящегося к теме $t$ в документе $d$ описывается общим для всей коллекции распределением $p(w|t)$ и не зависит от документа $d$:

$$p(w|d,t) \;=\; p(w|t). \tag{1}$$

Используя формулу полной вероятности и гипотезу условной независимости, получим:

$$p(w|d) = \sum_{t \in T} p(t|d)p(w|t). \tag{2}$$

Предполагается, что для каждого слова в документе автор сначала выбирает тему $t$ из распределения $p(t|d)$, а затем нужное слово из распределения $p(w|t)$.

Условные вероятности $\phi_{wt} \equiv p(w|t),\ \theta_{td} \equiv p(t|d)$ являются неизвестными параметрами модели. В новых обозначениях формула (2) записывается в следующем виде:

$$p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}. \tag{3}$$

*Задача тематического моделирования* состоит в нахождении стохастических матриц $\Phi = (\phi_{wt})_{W \times T}$ и $\Theta = (\theta_{td})_{T \times D}$, столбцы которых — распределения слов по темам и тем по документам.

Известными данными в этой задаче является матрица $F = (\hat{p}(w|d))_{W \times D}$ частотных оценок вероятностей $p(w|d)$:

$$\hat{p}(w|d) = \frac{n_{dw}}{n_d}.$$

Для решения задачи тематического моделирования в такой постановке применяется метод максимального правдоподобия:

$$\log \mathscr{L}(\Phi, \Theta) = \log \prod_i p(w_i, d_i|\Phi, \Theta) \longrightarrow \max_{\Phi,\Theta}. \tag{4}$$

Взяв логарифм от произведения, сгруппировав одинаковые слагаемые и отбросив константу, получаем следующую задачу оптимизации:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \log p(w|d) = \sum_{d \in D} \sum_{w \in W} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{td} \longrightarrow \max_{\Phi,\Theta}. \tag{5}$$

Стандартным методом решения задачи (5) является EM-алгоритм [19]. Это итерационный процесс, который последовательно выполняет две операции:

1. *E-шаг (expectation).* По текущему приближению параметров $(\Phi, \Theta)$ оценивается апостериорное распределение скрытых переменных:

$$p(t|d, w) = \frac{p(t|d)p(w|t)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}.$$

2. *M-шаг (maximization).* Зафиксировав распределение $p(t|d, w)$ скрытых переменных, получаем аналитическое решение задачи (5) максимизации правдоподобия:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \quad \theta_{td} = \frac{n_{td}}{n_d}, \tag{6}$$

где переменные $n_{wt}$, $n_t$ и $n_{td}$ выражаются через распределения $p(t|d, w)$ и имеют смысл счетчиков, аналогичных $n_{dw}$:

$$n_{wt} \equiv \sum_{d \in D} n_{dw} p(t|d, w), \quad n_{td} \equiv \sum_{w \in W} n_{dw} p(t|d, w),$$
$$n_t \equiv \sum_{w \in W} n_{wt} = \sum_{d \in D} n_{td}.$$

## 2.2 Неоднозначность решения PLSA

Задача оптимизации (5) эквивалентна задаче разложения заданной стохастической матрицы $F$ в произведение $\Phi\Theta$ двух стохастических матриц меньших размеров, минимизирующих дивергенцию Кульбака-Лейблера:

$$KL(F\|\Phi\Theta) \longrightarrow \min_{\Phi,\Theta}. \tag{7}$$

Идея разложения матрицы в произведение двух других матриц широко используется в задачах анализа данных. Известным методом является SVD-разложение, которое, однако, не подходит для тематического моделирования, т.к. минимизирует норму Фробениуса, а полученные матрицы в общем случае не являются стохастическими (стохастическая матрица — матрица, столбцы которой представляют собой дискретные распределения вероятностей).

Существуют методы, получающие *неотрицательное* матричное разложение, например, алгоритм HALS (Hierarchical Alternating Least Squares) [8], минимизирующий норму Фробениуса, или градиентный спуск с мультипликативными обновлениями [16], работающий и с дивергенцией Кульбака-Лейблера.

В случае точного представления $F = \Phi\Theta$, условие стохастичности исходной матрицы $F$ позволяет легко отнормировать результат неотрицательного матричного разложения. Но на практике мы получаем лишь приближенное решение $\Phi\Theta \approx F$, после нормировки которого результат может ухудшиться значительно.

Главной же проблемой здесь является то, что задача (7) невыпукла, допускает неединственное решение, поэтому EM-алгоритм и другие итерационные методы в лучшем случае могут гарантировать лишь нахождение *локального* минимума, а их результат будет сильно зависеть от начального приближения.

Проведем эксперименты на модельных данных, показывающие зависимость точности восстановления матриц $\Phi$ и $\Theta$ от выбора начального приближения и от доли нулей в исходной матрице $\Phi$.

На **рис. 1** показан пример сходимости EM-алгоритма на модельных данных при 30 различных случайных начальных приближениях.

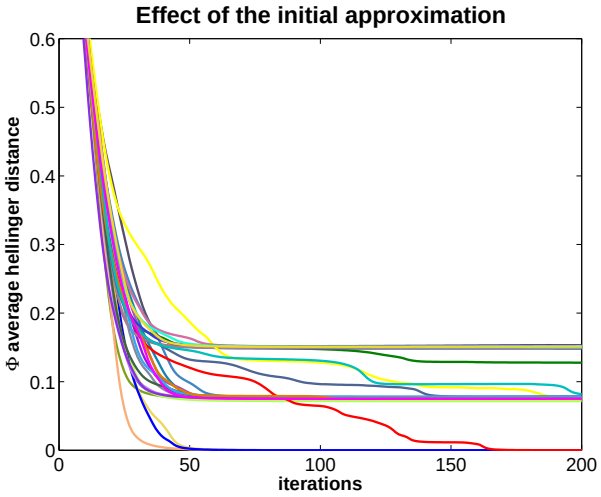



**Рис. 1:** Сходимость EM-алгоритма для разных начальных приближений.

Столбцы матрицы $\Phi$ генерировались из симметричного распределения Дирихле: $\phi_t \sim \mathrm{Dir}(\beta)$. Для генерации каждого столбца матрицы $\Theta$ генерировалось равномерно случайное число $T_d$ от 2 до 10 — количество тем в документе $d$, затем равномерно выбиралось подможество из $T_d$ тем, вероятности которых устанавливались одинаковыми.

Параметры модели: $|W| = 1000,\ |D| = 300,\ |T| = 20,\ \beta = 0.01$.

EM-алгоритм восстанавливал матрицы $\Phi$ и $\Theta$ по их произведению, после чего с помощью Венгерского алгоритма [15] производилась оптимальная перестановка тем, минимизирующая среднее *расстояние Хеллингера* между исходными $p(w|t)$ и найденными распределениями $\hat{p}(w|t)$:

$$\mathrm{Hellinger}\big(p(w|t), \hat{p}(w|t)\big) \;=\; \left( \frac{1}{2} \sum_{w \in W} \Big( \sqrt{p(w|t)} \;-\; \sqrt{\hat{p}(w|t)} \Big)^2 \right)^{\frac{1}{2}}.$$

Видно, что идеально восстановить матрицу $\Phi$ получилось лишь в небольшом числе случаев. Также остается открытым вопрос момента остановки данного итерационного алгоритма: число итераций, необходимых для достижения сходимости, может быть непредсказуемо велико.

На **рис. 2** приведена зависимость точности восстановления $\Phi$, $\Theta$ и их произведения $F$ от доли нулевых элементов в исходной матрице $\Phi$ для EM-алгоритма и метода неотрицательного матричного разложения HALS с нормировкой полученных матриц после каждой итерации.

Параметр $\beta$ симметричного распределения Дирихле перебирался по сетке так, чтобы получить матрицы $\Phi$ разной степени разреженности: от 0 до 99 процентов нулей.

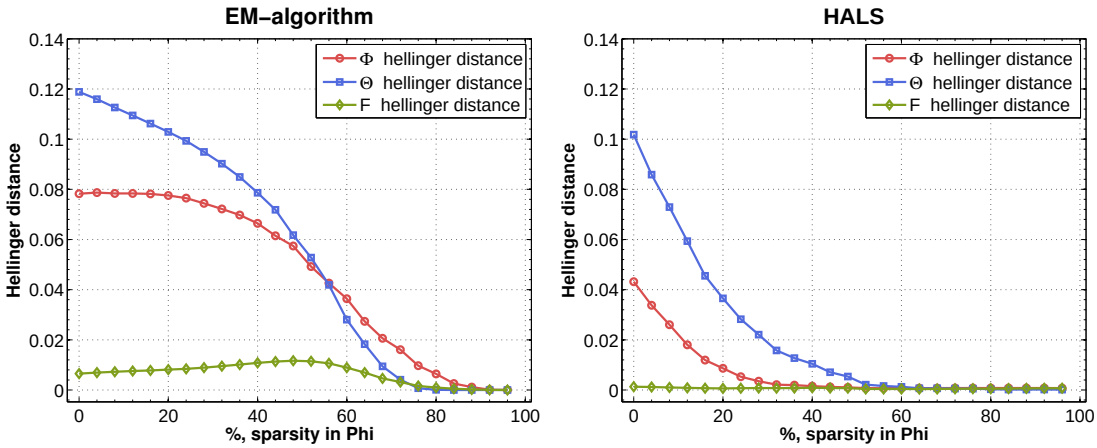



**Рис. 2:** Влияние разреженности Φ на качество работы итерационных методов.

Матрица $F$ восстанавливается неплохо практически при любом уровне разреженности. Но для полного восстановления Φ и Θ алгоритму HALS требуется более 60% нулей, EM-алгоритму — более 85%.

Стоит отметить, что реальные данные зачастую удовлетворяют требованию разреженности. Удается строить тематические модели, в которых число нулевых элементов в матрице Φ может достигать 98%, число нулевых элементов в матрице Θ — более 90% [27].

Для устранения проблемы неоднозначности решения задачи построения тематической модели воспользуемся подходом аддитивной регуляризации ARTM (Additive Regularization of Topic Models) [26].

## 2.3 Регуляризация задачи тематического моделирования

В подходе ARTM авторы предлагают в оптимизационной задаче (4) добавить к логарифму правдоподобия еще $r$ функционалов: $R_i(\Phi, \Theta),\ i = 1, \ldots, r$ называемых *регуляризаторами*, каждый со своим неотрицательным весом $\tau_i$:

$$\begin{cases} R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta), \qquad \log \mathscr{L}(\Phi,\Theta) + R(\Phi,\Theta) \longrightarrow \max\limits_{\Phi,\Theta}, \\ \phi_{wt} \geqslant 0, \quad \theta_{td} \geqslant 0, \quad \sum\limits_w \phi_{wt} = 1, \quad \sum\limits_t \theta_{td} = 1. \end{cases} \tag{8}$$

Приведем определения и теорему из [27], которые позволяют решать новую оптимизационную задачу методом, похожим по своей структуре на EM-алгоритм для PLSA.

**Определение 1.** Тема $t \in T$ называется *перерегуляризованной*, если

$$n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} \leqslant 0 \quad \text{для всех слов } w \in W.$$

**Определение 2.** Документ $d \in D$ называется *перерегуляризованным*, если

$$n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \leqslant 0 \quad \text{для всех тем } t \in T.$$

**Теорема 1.** *Если функция $R(\Phi, \Theta)$ непрерывно дифференцируема и $(\Phi, \Theta)$ — локальный максимум задачи (8) с ограничениями на стохастичность матриц $\Phi$ и $\Theta$, тогда для любой не перерегуляризованной темы $t \in T$ и любого не перерегуляризованного документа $d \in D$ справедлива следующая система уравнений:*

$$\begin{aligned} n_{dwt} &= n_{dw}\frac{\phi_{wt}\theta_{td}}{\sum_{s\in T}\phi_{ws}\theta_{sd}}; \\ \phi_{wt} &\propto \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+; \qquad & n_{wt} &= \sum_{d\in D} n_{dwt}; \\ \theta_{td} &\propto \left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+; & n_{td} &= \sum_{w\in W} n_{dwt}; \end{aligned} \tag{9}$$

*где $(x)_+ = \max\{x, 0\}$ — операция положительной срезки.*

Формулы (9) дают необходимое условие экстремума и при $R \equiv 0$ совпадают с формулами *M-шага* (6) классического EM-алгоритма для модели PLSA.

Для решения регуляризованной задачи оптимизации можно использовать тот же метод, но с модифицированным *M-шагом*. Построенный таким образом EM-алгоритм можно рассматривать как частный случай метода простой итерации.

## 2.4 Используемые регуляризаторы

Будем использовать два конкретных регуляризатора: *сглаживание* и *разреживание*, которые в наборе с другими подробно описаны в [27].

**Сглаживание фоновых тем.** Зафиксируем в нашей модели некоторое число тем $|T|$ и выделим подмножество $T_0 \subset T$ для слов, составляющих общую лексику. Текст на естественном языке обычно на $80\% - 90\%$ состоит из таких слов.

Основной интерес для нас представляют специализированные, тематические слова, поэтому будем считать, что общая лексика относится к фоновой теме, которая должна стремиться к естественному распределению слов в языке $\beta_w$.

Распределение $\beta_w$ можно взять равномерным, оценить по внешней коллекции, например, по Википедии, или по исходной с помощью частот вхождения слов по всем документам:

$$\hat{\beta}_w = \frac{n_w}{n} = \frac{\sum_{d \in D} n_{dw}}{\sum_{d \in D} n_d}.$$

Потребуем, чтобы распределение $\phi_{wt}$ было близко к распределению $\beta_w$ для фоновых тем $t \in T_0$ по дивергенции Кульбака-Лейблера:

$$KL(\beta_w \| \phi_{wt}) \longrightarrow \min \quad \Longleftrightarrow \quad R(\Phi) = \sum_{w \in W} \beta_w \log \phi_{wt} \longrightarrow \max.$$

Продифференцируем $R(\Phi)$ и, согласно общей формуле М-шага (9), получим:

$$\phi_{wt} \ \propto \ n_{wt} \ + \ \beta_0 \beta_w, \quad \text{для } t \in T_0, \tag{10}$$

где $\beta_0 \geqslant 0$ — коэффициент регуляризации.

Аналогично, для каждого документа будем приближать фоновые темы по дивергенции Кульбака-Лейблера к равномерному распределению:

$$\theta_{td} \ \propto \ n_{td} \ + \ \alpha_0 \alpha_t [t \in T_0], \quad d \in D. \tag{11}$$

Описанная регуляризация является аналогом известной вероятностной тематической модели LDA (Latent Dirichlet Allocation) [5], в которой предполагается, что столбцы матриц $\Phi$ и $\Theta$ генерируются из распределений Дирихле.

**Разреживание предметных тем.** Каждая предметная тема должна состоять из множества специализированных слов, отличающих данную тему от остальных тем и от распределения слов в языке. Также каждый документ должен содержать лишь малое число тем, поэтому будем отдалять наши распределения $\phi_{wt}$ и $\theta_{td}$ от априорных распределений $\beta_w$ и $\alpha_t$ по дивергенции Кульбака-Лейблера.

В итоге формулы М-шага для предметных тем $t \in T \setminus T_0$ получаются очень похожими на формулы сглаживания (10) и (11), но только с другим знаком:

$$\begin{aligned} \phi_{wt} \ &\propto \ (n_{wt} \ - \ \beta_1 \beta_w)_+; \\ \theta_{td} \ &\propto \ (n_{dt} \ + \ \alpha_0 \alpha_t [t \in T_0] \ - \ \alpha_1 \alpha_t [t \notin T_0])_+. \end{aligned} \tag{12}$$

Здесь $\beta_1 \geqslant 0$ и $\alpha_1 \geqslant 0$ — еще два коэффициента регуляризации, являющиеся параметрами модели и требующие подбора.

# 3 Оценки качества тематических моделей

От распределений $\phi_{wt}$ и $\theta_{td}$, полученных в ходе построения тематической модели, требуется обладание многими полезными свойствами: разреженностью — большим числом нулей, отсутствием фоновых слов в предметных темах, различностью предметных тем друг от друга, плавностью изменения тем во времени, а главное — интерпретируемостью.

## 3.1 Когерентность

Наиболее популярная оценка интерпретируемости тематических моделей называется *согласованностью* или *когерентностью* [21, 18] и может вычисляться следующим образом:

- Находим частоты встречаемости слов по отдельности и в парах:

  $N_{w_i}$ — число документов, содержащих слово $w_i$;

  $N_{w_i w_j}$ — число документов, содержащих пару слов $(w_i, w_j)$ вместе.

  При этом подсчет может вестись как по исходной коллекции, так и по внешним источникам, например, Википедии. Число вхождений также можно считать разными способами: как по целым документам, так и в окне размера $10 - 20$ слов.

- PMI (Pointwise Mutual Information) пары слов $(w_i, w_j)$:

  $$\text{PMI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \approx \log \frac{N_{w_i w_j}|D| + \varepsilon}{N_{w_i} N_{w_j}},$$

  где $\varepsilon$ — добавка, чтобы логарифм всегда был определен, предложенная в [25]. В экспериментах бралось значение $\varepsilon = 10^{-6}$.

  Если слова независимы: $p(w_i, w_j) = p(w_i)p(w_j)$, PMI равен нулю:

  $$\text{PMI}(w_i, w_j) = \log \frac{p(w_i)p(w_j)}{p(w_i)p(w_j)} = \log 1 = 0.$$

  Чем более два слова согласованы друг с другом, тем чаще они встречаются вместе и тем больше их PMI.

- *Когерентность* темы $t$ определяется как среднее арифметическое попарных PMI среди $k = 10$ наиболее вероятных слов темы:

  $$\text{Coherence}(t) = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \text{PMI}(w_{(i)}, w_{(j)}).$$

  Здесь: $w_{(i)}$ — слово с $i$-м по убыванию значением вероятности.

Интерпретируемые темы обладают большой когерентностью. В экспериментах [22] было показано, что эта мера хорошо коррелирует с разметкой тем на интерпретируемые и не интерпретируемые, сделанной экспертами-асессорами.

## 3.2 Дополнительные оценки качества

Помимо когерентности, будем следить за набором дополнительных характеристик, позволяющих наблюдать за процессом сходимости и определять, обладают ли искомые распределения $\phi_{wt}$ и $\theta_{td}$ теми или иными свойствами.

1. **Перплексия** — величина, выражающаяся через правдоподобие выборки и позволяющая отслеживать сходимость метода оптимизации:

$$\text{Perplexity}(\Phi, \Theta) \;=\; \exp\Big(-\frac{1}{n}\sum_{d\in D}\sum_{w\in W} n_{dw}\log p(w|d)\Big),$$

$$p(w|d) \;=\; \sum_{t\in T}\phi_{wt}\theta_{td}, \quad n \;\equiv\; \sum_{d\in D}\sum_{w\in W} n_{dw}.$$

   Численное значение перплексии не имеет интерпретации и позволяет лишь сравнивать алгоритмы между собой. Значения чем меньше, тем лучше.

   В экспериментах значение перплексии делилось на 1000 для попадания её в интервал $[0, 1]$ и отображения вместе с остальными метриками.

2. **Разреженность матриц $\Phi$ и $\Theta$** — доля нулевых элементов.

   В предметных темах разреженность достигает $90 - 95\%$, поэтому, для хорошей тематической модели разреженность необходима.

Лексическим *ядром темы* будем называть множество слов, отличающих данную тему от остальных. Одним из способов формализации понятия ядра темы является [27]:

$$W_t \;=\; \{w \in W \mid p(t|w) > \delta\}.$$

Параметр $\delta = 0.25$. Подбирается с тем расчетом, чтобы *размер ядра* $|W_t|$ был от 20 до 200 слов.

На основе ядра темы строятся следующие две оценки:

4. **Чистота** — суммарная вероятность слов ядра:

$$\text{Purity}(t) \;=\; \sum_{w\in W_t} p(w|t) \;=\; \sum_{w\in W_t}\phi_{wt}$$

   Показывает насколько хорошо тема описывается своим ядром. Чем выше, тем лучше.

5. **Контрастность** — средняя вероятность встретить слова ядра в конкретной теме:

$$\text{Contrast}(t) \;=\; \frac{1}{|W_t|}\sum_{w\in W_t} p(t|w)$$

При большой контрастности тема однозначно угадывается по своему ядру, при малой — тема размывается, становится нечеткой.

Дополнительно введем две оценки, показывающие, насколько хорошо фоновые слова отделяются от тематических, и насколько различными оказываются тематики:

6. **Зашумленность.**

Пусть $\beta_w$ — естественное распределение слов в языке.

$$\text{Noisiness}(t) \;=\; \sum_{w\in W} \phi_{wt}\beta_w \;\Big/\; \sum_{w\in W} \beta_w^2$$

— взвешенная средняя вероятность встретить в произвольном тексте слова из темы. Принимает значения из $[0, 1]$. Большие значения зашумленности говорят о наличии в теме фоновых слов.

7. **Rescaled Dot Product.**

Для двух фиксированных тем $t_1, t_2 \in T$ обозначим

$$P(w) = p(w|t_1), \qquad Q(w) = p(w|t_2)$$

— векторы распределений слов по темам.

Rescaled Dot Product (отнормированное скалярное произведение) двух тематик $t_1$ и $t_2$ определяется как:

$$\text{Rescaled Dot Product}(t_1, t_2) \;=\; \frac{P\cdot Q \;-\; d_{\text{Min}}}{d_{\text{Max}} \;-\; d_{\text{Min}}},$$

$$d_{\text{Min}} = \overrightarrow{P}\cdot\overleftarrow{Q}, \qquad d_{\text{Max}} = \overrightarrow{P}\cdot\overrightarrow{Q},$$

где $\overrightarrow{P}$ — вектор, значения которого отсортированы по убыванию, $\overleftarrow{P}$ — по возрастанию; $P\cdot Q$ — скалярное произведение между векторами.

В работе [7] было показано, что данная мера схожести между темами наиболее хорошо коррелирует с экспертными оценками схожести.

Значения, близкие к 1, указывают на сильную корреляцию между темами.

Для исследования моделей со временем, где каждому документу $d$ привязана метка времени $y_d \in Y$ из множества временных отчетов, добавим еще три характеристики:

8. **Разреженность распределений $p(t|y)$** — доля нулевых элементов среди всех распределений тем во времени. Позволяет оценивать воздействие на модель регуляризатора *разреживания* $p(t|y)$.

9. **Колебание темы во времени:**

$$\text{Variation}(t) \;=\; \sum_{y \in Y} \big|\, \sqrt{p(y|t)} \;-\; \sqrt{p(y-1|t)}\, \big|;$$

   Меньшие значения соответствуют более плавному изменению темы во времени.

10. **Событийность темы.**

    Большое число тем для потоков текстовых документов можно разбить на два класса: постоянные темы и темы-события.

    Постоянные темы присутствуют на протежении всего промежутка времени, распределение $p(y|t)$ для такой темы близко к равномерному.

    Тема-событие характеризуется внезапным появлением и постепенным затуханием во времени, её распределение $p(y|t)$ обладает большим числом нулей.

    Определим величину *событийности* темы как долю чисел в распределении $p(y|t)$, меньших заданного порога $\varepsilon$.

    Будем говорить, что тема — событие, если её событийность больше некоторой величины $\delta$.

    Константа $\varepsilon$ определяет какие значения в $p(y|t)$ мы полагаем равными нулю и служит для устранения шума.

На выходе тематической модели мы получаем матрицы $\Phi$ и $\Theta$. Чтобы оценить качество модели в целом, требуется усреднение.

В экспериментах бралась *медиана* по множеству тем для каждой характеристики, т.к. она обладает большей устойчивостью к разбросу значений, чем, например, среднее арифметическое.

# 4 Тематическая модель со временем

К документам может быть привязана дополнительная метаинформация, описывающая источники, авторов или дату публикации. В тематическом моделировании такая информация интересна как сама по себе, для интерпретации и визуализации полученных тем, так и для улучшения качества модели, путем учета внешних связей между документами.

Рассмотрим способ моделирования документов, привязанных ко времени.

Будем считать, что для каждого документа $d$ известен единственный момент времени $y_d = y(d) \in Y$ из некоторого линейно-упорядоченного множества временных меток.

## 4.1 Обзор известных моделей

Существует множество работ, посвященных анализу тематических моделей коллекций документов со временем.

Некоторые авторы строят модели, никак не учитывающие время, а затем проводят *post hoc* анализ динамик изменения тем [11, 12].

Ещё одним универсальным приёмом является использование онлайн алгоритмов построения тематических моделей [2, 10]. Статические модели PLSA или LDA последовательно обрабатывают приходящие пачки документов: запускается обычный алгоритм построения модели, для которого в качестве начального приближения используется результат построения в предыдущий момент времени.

Способы явного включения времени в вероятностную модель чаще всего основаны на байесовском подходе: желаемые особенности добавляются с помощью указания априорных распределений на параметры.

Можно выделить два наиболее популярных направления: использование непрерывного априорного распределения $p(y|t)$ времени для каждой темы и модели с дискретным временем, основанные на *Марковском свойстве*.

Относящаяся к первому классу модель TOT (Topics Over Time, [29]) расширяет модель LDA, задавая априорное бета-распределение времени для каждой темы:

$$p(y|t) = \text{Beta}(y|a_t, b_t),$$

— время считается непрерывным из отрезка $[0, 1]$. Предполагается, что для каждого документа метка времени $y_d$ генерируется из смеси:

$$p(y|d) = \sum_{t \in T} p(y|t)p(t|d).$$

Характерный вид бета-распределения не позволяет моделировать сложные периодические явления, а из-за ограниченности времени ($y \in [0, 1]$) обработка новых документов не допускается.

Несмотря на эти недостатки, модель TOT получила большое развитие в литературе, имеются её обобщения для учёта корреляций между темами [17], авторов документов [20] и для непараметрического случая, когда число тем определяется по коллекции автоматически, с помощью *иерархических процессов Дирихле* [9].

В моделях с *Марковским свойством* время дискретизируется, документы с одинаковой меткой времени собираются в пачки, для каждой из которых строится своя тематическая модель. Модели для соседних моментов времени притягиваются друг к другу с помощью некоторого распределения.

Так, в [4, 28] построенные в соседние моменты времени тематические модели LDA связываются с помощью логит-нормального распределения, а в [30] — с помощью распределения Дирихле. Существует большое число обобщений данного подхода на непараметрический случай [23, 32, 1, 3].

Особенностью такого подхода является динамичность тем с течением времени: темы могут значительно расширяться и даже переходить в другие предметные области. Например, представим подмножества документов с тремя четкими парами тем: «птицы и полёты», «полёты и физика», «физика и квантовая механика». Тематическая модель этой коллекции с динамическими темами может привести исследователя к нежелательному выводу о существовании в ней связи между птицами и квантовой механикой. Модель со статическими, хорошо различающимися между собой темами, позволяет избегать таких недостатков.

## 4.2 Описание модели

Расширим наше вероятностное пространство на множество $D \times W \times T \times Y$.

Примем *гипотезу условной независимости* для распределения тем во времени — вероятность метки $y$ в фиксированном документе $d$ не зависит от выбранной темы $t$:

$$p(y|t, d) \;=\; p(y|d). \tag{13}$$

В нашей модели будем считать, что распределение временных меток по документам является вырожденным:

$$p(y|d) \;=\; [y = y_d].$$

Найдем распределение временных меток для каждой темы:

$$\begin{aligned} p(y|t) &= \sum_{d \in D} p(y|t,d)p(d|t) = \sum_{d \in D} [y = y_d] \frac{p(t|d)p(d)}{p(t)} = \\ &= \sum_{d \in D_y} \frac{\theta_{td} p(d)}{p(t)} = \frac{1}{p(t)} \sum_{d \in D_y} \theta_{td} p(d), \end{aligned} \tag{14}$$

где $D_y = \{d \in D \mid y_d = y\}$ — множество документов, привязанных к моменту времени $y$.

По формуле Байеса выводим распределение тем для фиксированного момента времени:

$$p(t|y) = \frac{p(y|t)p(t)}{p(y)} = \frac{1}{p(y)} \sum_{d \in D_y} \theta_{td} p(d). \tag{15}$$

С помощью счетчиков: $n_d$, $n_t$ и $n_y \equiv \sum\limits_{d \in D_y} n_d$ — можем получить следующие частотные оценки для вероятностей:

$$\begin{aligned} \hat{p}(y|t) &= \frac{1}{n_t} \sum_{d \in D_y} \theta_{td} n_d, \\ \hat{p}(t|y) &= \frac{1}{n_y} \sum_{d \in D_y} \theta_{td} n_d. \end{aligned} \tag{16}$$

## 4.3 Регуляризаторы времени

Распределения $p(y|t)$ и $p(t|y)$ можно использовать как для визуализации тематической модели, так и для улучшения её качества.

Определим два регуляризатора, использующих метки времени:

**Разреживание $p(t|y)$** — основано на предположении, что в каждый момент времени число актуальных тем невелико.

Действуя аналогично разреживанию предметных тем (12), будем отдалять распределение $p(t|y)$ от равномерного по дивергенции Кульбака-Лейблера:

$$R_1(\Theta) = -\sum_{y \in Y} \sum_{t \in T} \frac{1}{|T|} \log p(t|y) \longrightarrow \max_{\Theta},$$

$$\frac{\partial R_1}{\partial \theta_{td}} = -\frac{1}{|T|} \frac{p(d)/p(y)}{p(t|y)}, \qquad \text{где } y = y(d).$$

Производная регуляризатора не изменится, если вместо $p(t|y)$ разреживать распределение $p(y|t)$. Отсюда получаем другую интерпретацию этого регуляризатора: разреживаем моменты времени, в которые тема присутствует.

Распределение $p(y|t)$ можно рассматривать как нормированный временной ряд. Тогда каждый регуляризатор стремится придать этому ряду те или иные свойства: разреженность, сглаженность, периодичность и т.п.

**Плавность изменения $p(y|t)$** — с течением времени тема должна изменяться плавно, с редкими скачками:

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

Преимуществом этого функционала перед более гладкими аналогами, например, суммой квадратов отклонений, является отбор тем во времени, подобно отбору признаков в $L_1$-регуляризации линейной регрессии.

Похожая регуляризация успешно применяется при решении задач разладки временного ряда [14]. Модуль позволяет определять моменты времени, когда временной ряд переключается с одного режима на другой, а также поощряет зануление участков ряда с незначительными отклонениями от нуля.

Недостаток такого подхода: недифференцируемость модуля в нуле.

Для вывода формулы М-шага рассмотрим регуляризованную задачу максимизации правдоподобия в общем случае, вместе с *негладким* регуляризатором:

$$\begin{cases} \sum\limits_{d\in D}\sum\limits_{w\in W} n_{dw}\log\sum\limits_{t\in T}\phi_{wt}\theta_{td} \;+\; R(\Phi,\Theta) \;-\; \sum\limits_{i=1}^{r}\tau_i\big|R_i(\Phi,\Theta)\big| \;\longrightarrow\; \max\limits_{\Phi,\Theta}, \\ \phi_{wt}\geqslant 0, \quad \theta_{td}\geqslant 0, \quad \sum\limits_{w}\phi_{wt}=1, \quad \sum\limits_{t}\theta_{td}=1, \end{cases} \tag{17}$$

где первое слагаемое есть правдоподобие модели, $R(\Phi,\Theta)$ — гладкий регуляризатор, а $R_i(\Phi,\Theta)$ суть линейные по $\Phi$ и $\Theta$ функции (линейность используется для выполнения условий *регулярности* [31]), $\tau_i \geqslant 0$ — коэффициенты регуляризации.

Справедливо следующее утверждение:

**Теорема 2.** *Если $(\Phi,\Theta)$ — локальный максимум задачи (17) с указанными ограничениями на $R(\Phi,\Theta)$ и $R_i(\Phi,\Theta)$, тогда для любой не перерегуляризованной темы $t\in T$ и любого не перерегуляризованного документа $d\in D$ справедлива следующая система уравнений:*

$$\phi_{wt} \;\propto\; \left(n_{wt} \;+\; \phi_{wt}\Big(\frac{\partial R}{\partial\phi_{wt}} \;-\; \sum_{i=1}^{r}\tau_i\frac{\partial R_i}{\partial\phi_{wt}}\operatorname{sign}(R_i)\Big)\right)_+; \tag{18}$$

$$\theta_{td} \;\propto\; \left(n_{td} \;+\; \theta_{td}\Big(\frac{\partial R}{\partial\theta_{td}} \;-\; \sum_{i=1}^{r}\tau_i\frac{\partial R_i}{\partial\theta_{td}}\operatorname{sign}(R_i)\Big)\right)_+;$$

*где*

$$\operatorname{sign}(x) \;=\; \begin{cases} 1, & x > 0, \\ -1, & x < 0, \\ [-1,\ 1], & x = 0. \end{cases}$$

*Примечание.* В случае, когда аргумент функции sign равен нулю, выражения (18) означают принадлежность величин $\phi_{wt}$ и $\theta_{td}$ множествам, указанным в правых частях.

*Доказательство.* Введём для каждой величины под модулем $G_i = R_i(\Phi, \Theta)$ две дополнительные переменные $G_i^+$ и $G_i^-$ — положительную и отрицательную срезку:

$$G_i^+ \;\equiv\; \max\{G_i, 0\}, \qquad G_i^- \;\equiv\; \max\{-G_i, 0\};$$

$$G_i \;=\; G_i^+ \;-\; G_i^-, \qquad |G_i| \;=\; G_i^+ \;+\; G_i^-.$$

Добавив в общей сложности $2\,r$ новых переменных $\left\{G_i^+, G_i^-\right\}_{i=1}^r \equiv G$, получим новую задачу оптимизации с ограничениями, эквивалентную исходной:

$$\sum_{d\in D}\sum_{w\in W} n_{dw} \log \sum_{t\in T} \phi_{wt}\theta_{td} \;+\; R(\Phi,\Theta) \;-\; \sum_{i=1}^{r} \tau_i \left(G_i^+ + G_i^-\right) \;\longrightarrow\; \max_{\Phi,\Theta,G},$$

Ограничения типа равенство:

$$G_i^+ - G_i^- \;=\; R_i(\Phi,\Theta), \quad \sum_w \phi_{wt} = 1, \quad \sum_t \theta_{td} = 1.$$

Ограничения типа неравенство:

$$\theta_{td} \geqslant 0, \quad \phi_{wt} \geqslant 0, \quad G_i^+ \geqslant 0, \quad G_i^- \geqslant 0.$$

Запишем Лагранжиан:

$$\begin{aligned} L \;=\; & \sum_d \sum_w n_{dw} \log \sum_t \phi_{wt}\theta_{td} \;+\; R(\Phi,\Theta) \;-\; \sum_i \tau_i \left(G_i^+ + G_i^-\right) \;+ \\ & + \sum_t \lambda_t (1 - \sum_w \phi_{wt}) \;+\; \sum_w \sum_t \lambda_{wt}\phi_{wt} \;+ \\ & + \sum_d \mu_d (1 - \sum_t \theta_{td}) \;+\; \sum_t \sum_d \mu_{td}\theta_{td} \;+ \\ & + \sum_i \eta_i (G_i^+ - G_i^- - R_i(\Phi,\Theta)) \;+\; \sum_i \xi_i G_i^+ \;+\; \sum_i \zeta_i G_i^-, \end{aligned}$$

где $\{\lambda_t,\ \lambda_{wt},\ \mu_d,\ \mu_{td},\ \eta_i,\ \xi_i,\ \zeta_i\}$ — набор *двойственных* переменных.

По теореме Каруша-Куна-Таккера [6], если $(\Phi, \Theta, G)$ — точка локального максимума задачи с ограничениями, то существует набор двойственных переменных такой, что:

$$\nabla_{(\Phi,\Theta,G)} L \;=\; 0,$$
$$\mu_{td} \geqslant 0, \quad \lambda_{wt} \geqslant 0, \quad \xi_i \geqslant 0, \quad \zeta_i \geqslant 0,$$
$$\mu_{td}\theta_{td} \;=\; 0, \quad \lambda_{wt}\phi_{wt} \;=\; 0, \quad \xi_i G_i^+ \;=\; 0, \quad \zeta_i G_i^- \;=\; 0.$$

Докажем отсюда утверждение теоремы для $\Theta$, для $\Phi$ доказательство полностью аналогичное.

Посчитаем производные:

$$\frac{\partial L}{\partial \theta_{td}} \;=\; \sum_w n_{dw} \frac{\phi_{wt}}{p(w|d)} \;+\; \frac{\partial R}{\partial \theta_{td}} \;-\; \mu_d \;+\; \mu_{td} \;-\; \sum_i \eta_i \frac{\partial R_i}{\partial \theta_{td}} \;=\; 0.$$

Домножим на $\theta_{td}$:

$$n_{td} \;+\; \theta_{td} \left( \frac{\partial R}{\partial \theta_{td}} \;-\; \sum_i \eta_i \frac{\partial R_i}{\partial \theta_{td}} \right) \;=\; \theta_{td}\mu_d.$$

Просуммируем по $t$ — получим выражение для $\mu_d$:

$$\mu_d \;=\; n_d \;+\; \sum_t \theta_{td} \left( \frac{\partial R}{\partial \theta_{td}} \;-\; \sum_i \eta_i \frac{\partial R_i}{\partial \theta_{td}} \right).$$

Если документ $d$ не *перерегуляризован* получим:

$$\theta_{td} \;\propto\; \left( n_{td} \;+\; \theta_{td} \Big( \frac{\partial R}{\partial \theta_{td}} \;-\; \sum_i \eta_i \frac{\partial R_i}{\partial \theta_{td}} \Big) \right)_+ .$$

Осталось найти выражение для $\eta_i$.

Производные по введённым переменным:

$$\begin{aligned} \frac{\partial L}{\partial G_i^+} \;&=\; -\tau_i \;+\; \eta_i \;+\; \xi_i \;=\; 0, \\ \frac{\partial L}{\partial G_i^i} \;&=\; -\tau_i \;-\; \eta_i \;+\; \zeta_i \;=\; 0; \end{aligned} \tag{19}$$

$\Rightarrow$

$$\xi_i \;=\; \tau_i \;-\; \eta_i \;\geqslant\; 0,$$

$$\zeta_i \;=\; \tau_i \;+\; \eta_i \;\geqslant\; 0;$$

— условия неотрицательности сопряженных переменных, соответствующих ограничениям неравенствам.

Отсюда:

$$|\eta_i| \leqslant \tau_i.$$

Домножим условия (19) на $G_i^+$ и $G_i^-$ соответственно и сложим:

$$-\tau_i(G_i^+ + G_i^-) + \eta_i(G_i^+ - G_i^-) = 0.$$

Рассмотрев два случая ($G_i^+ + G_i^- \neq 0$) и ($G_i^+ + G_i^- = 0$), окончательно получаем:

$$\eta_i = \tau_i \operatorname{sign}(R_i).$$

□

Условию линейности по $\Phi$ и $\Theta$ удовлетворяют выражения вида $p(y|t) - p(y-1|t)$, $p(y+1|t) - 2p(y|t) + p(y-1|t)$ и другие аналоги разностных производных для $p(y|t)$.

Применяя **теорему 2** к регуляризатору *плавного изменения* $p(y|t)$ получим следующую формулу M-шага модифицированного EM-алгоритма:

$$\theta_{td} \propto \left( n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} + \theta_{td}\frac{n_d}{n_t}\tau_2\Big(-\operatorname{sign}\big(p(y|t) - p(y-1|t)\big) + \operatorname{sign}\big(p(y+1|t) - p(y|t)\big)\Big)\right)_+,$$

В экспериментах значение функции sign в нуле бралось либо равным 0, либо генерировалось случайно из равномерного распределения, заметной разницы между этими двумя подходами выявлено не было.

## 4.4 Эксперименты

Эксперименты проводились на коллекции пресс-релизов министерств иностранных дел ряда стран за промежуток времени длиной более десяти лет.

Всего около 20 000 документов.

Каждое сообщение состояло из заголовка, основного текста и дополнительной информации: даты создания документа, источника, и страны происхождения. Все сообщения на английском языке.

Предварительная обработка включала извлечение данных, избавление от мусора (случайно попавших элементов верстки) и лемматизацию текста с помощью лексической базы WordNet[1] и библиотеки NLTK[2].

На выходе лемматизатора получались слова, приведенные к нормальной форме, после чего составлялся словарь и матрица частот.

Слова, встречающиеся суммарно менее 10 раз, отбрасывались как слишком редкие, шумовые, не несущие частотной тематической информации.

**Настройка тематической модели**

Для настройки параметров тематической модели были отобраны 2000 документов. Объем словаря $|W|$ составил 5070 слов. Число тем было фиксированным $|T| = 100$.

В экспериментах проводилось по 100 итераций EM-алгоритма из случайных начальных приближений. Сильной зависимости от начального приближения (как в модельных данных на **рис. 1**) выявлено не было, поэтому дальнейшие запуски выполнялись для одного-двух приближений.

Использовалась следующая тактика регуляризации: выделялись три фоновых темы с целью притягивания стоп-слов, которые подвергались сглаживанию с первой итерации. Для остальных тем применялись разреживающие регуляризаторы, которые включались с $15 - 20$ итерации и постепенно увеличивали свое действие.

Подбор коэффициентов регуляризации выполнялся поочередно: фиксировались все коэффициенты, кроме одного, значения которого варьировались на отрезке $[0, \tau_{\max}]$. Выбиралось значение с большей когерентностью, при этом велся учет остальных характеристик и ручное оценивание тем. Допускались значения с не самой большой (но локально максимальной) когерентностью, если другие параметры модели были значительно лучше.

После выбора оптимального параметра, он становился фиксированным, а поиск продолжался для остальных коэффициентов. Данный процесс является близким ана-

[1] http://wordnet.princeton.edu

[2] http://www.nltk.org

логом покоординатного спуска, но с постоянным контролем модели и выбором решений, лучших по совокупности критериев.

На **рис. 3** приведено сравнение изначальной PLSA и наилучшей регуляризованной модели ARTM.

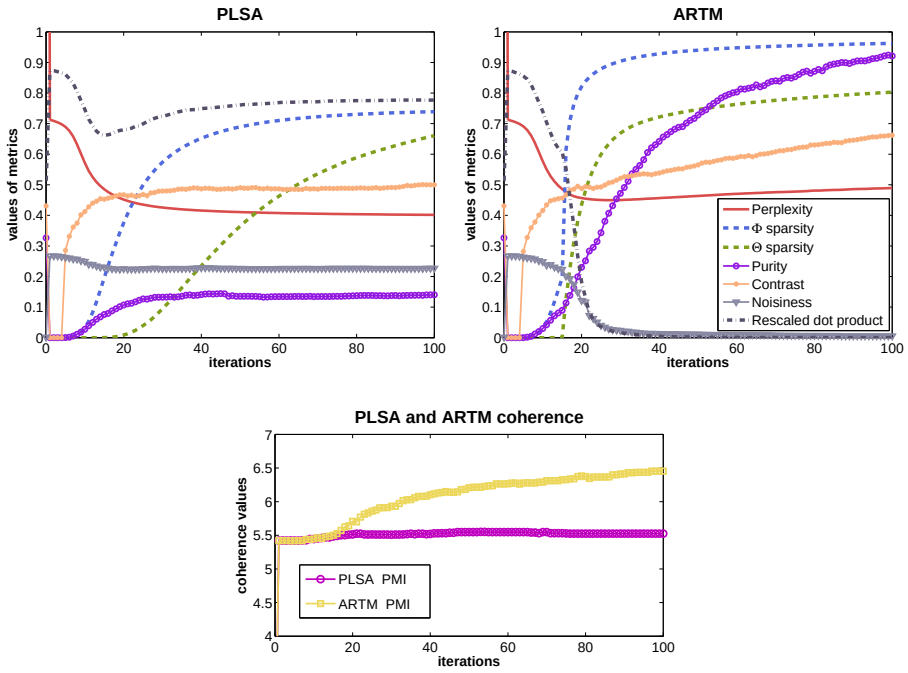



**Рис. 3: Сравнение моделей PLSA и регуляризованной ARTM.**
На верхних графиках показан процесс сходимости от числа итераций, на нижнем: когерентности обоих моделей. Запуск производился из одного и того же начального приближения.

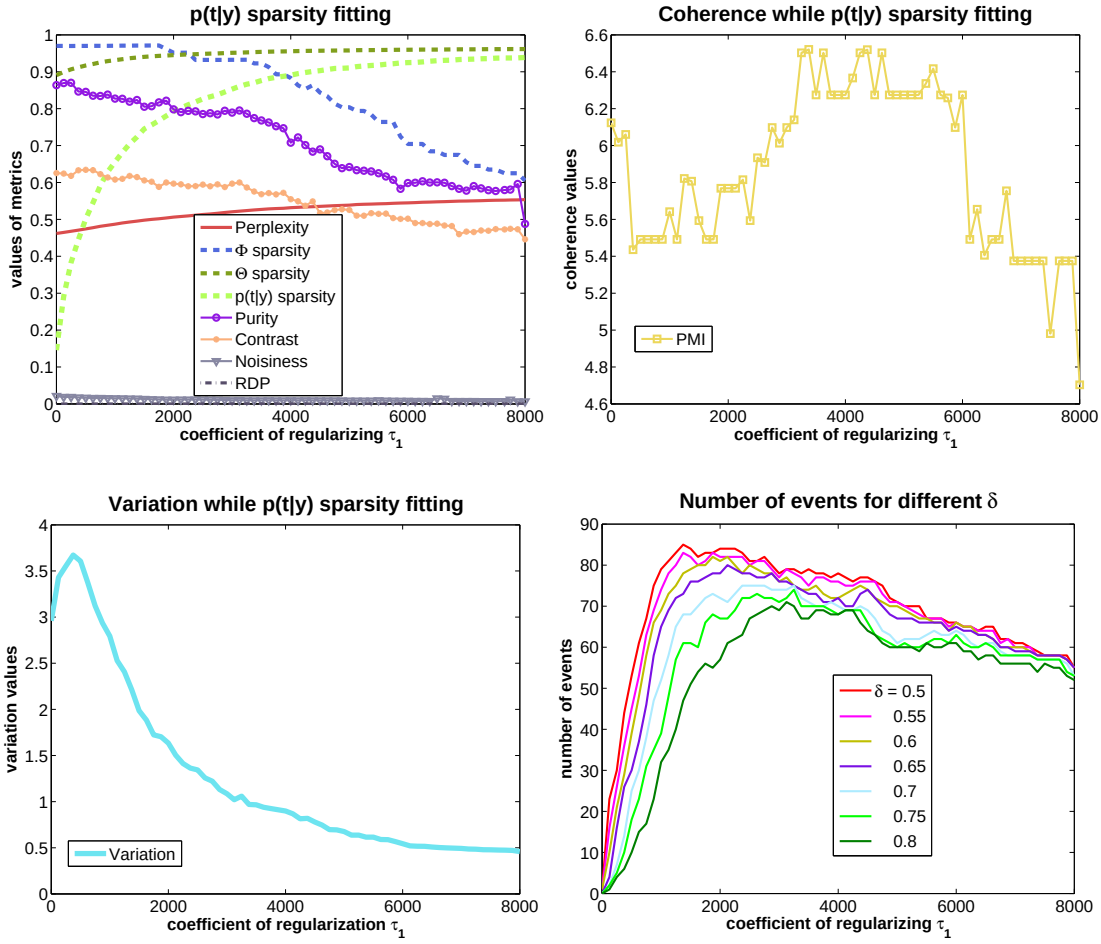
### Добавление времени

После настройки базовой тематической модели, в неё было добавлено вычисление распределений $p(t|y)$ и $p(y|t)$, визуализация тематик во времени и регуляризаторы разреживания и плавного изменения $p(y|t)$. На **рис. 4** и **5** представлены характеристики модели в зависимости от коэффициентов регуляризации.

Регуляризатор разреживания $p(t|y)$ увеличивает долю нулевых элементов в распределениях $p(t|y)$, при правильном подборе коэффициента регуляризации способствует улучшению интерпретируемости и уменьшению колебания тем во времени.

При большем разреживании число тем-событий начинает сходиться к определенному значению и становится всё менее чувствительным к порогу $\delta$, разделяющем по метрике событийности постоянные темы от кратковременных.

**Рис. 4: Настройка параметра регуляризации $\tau_1$ — разреживание распределения $p(t|y)$.**
Оптимальными значениями для $\tau_1$ видятся $4000 - 6000$: доля нулей в распределениях $p(t|y)$ достигает 90%, интерпретируемость улучшается.
Дальнейшее увеличение $\tau_1$ ведет к значительному падению когерентности.




Регуляризатор плавного изменения $p(y|t)$ способствует уменьшению колебания тем во времени, стремясь сблизить соседние значения вероятностей $p(y|t)$ и $p(y-1|t)$.

Имеет смысл применять его *совместно* с разреживанием $p(t|y)$, иначе модель ведет себя нестабильно, процесс сходимости начинает сильно зависеть от начального приближения, а темы перестают иметь визуальную интерпретацию.

Также увеличение плавности $p(y|t)$ ведёт к образованию всё большего числа постоянных тем, в связи с чем могут исчезать интересуемые темы-события. В таких случаях предлагается увеличивать параметр $|T|$ — число тем в модели.

**Рис. 5: Настройка параметра регуляризации $\tau_2$ плавного изменения $p(y|t)$.**

Разреживание $p(t|y)$ фиксировано: $\tau_1 = 4000$.

Представляются оптимальными значения $\tau_2 \leqslant 0.012$, при которых происходит уменьшение колебания тем во времени и улучшение визуального представления тем, а когерентность и остальные характеристики остаются на приемлемом уровне

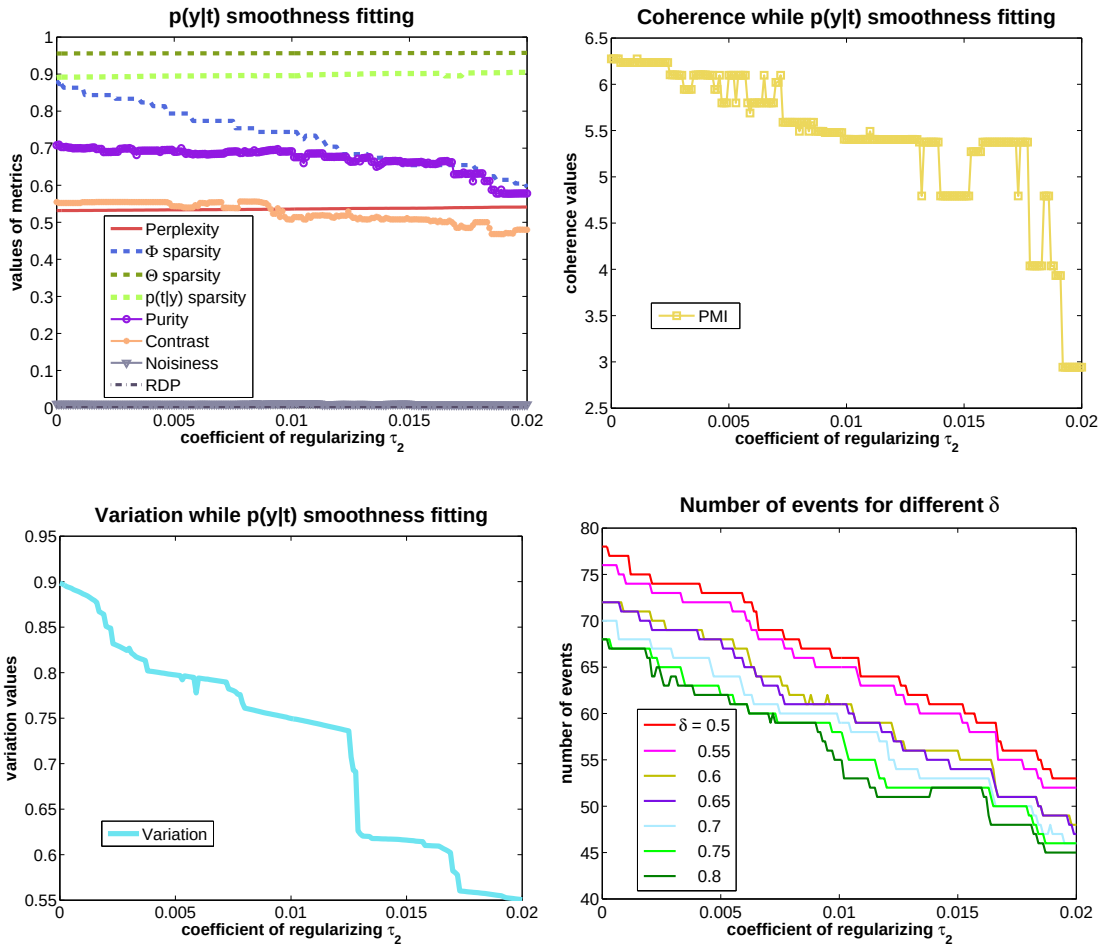
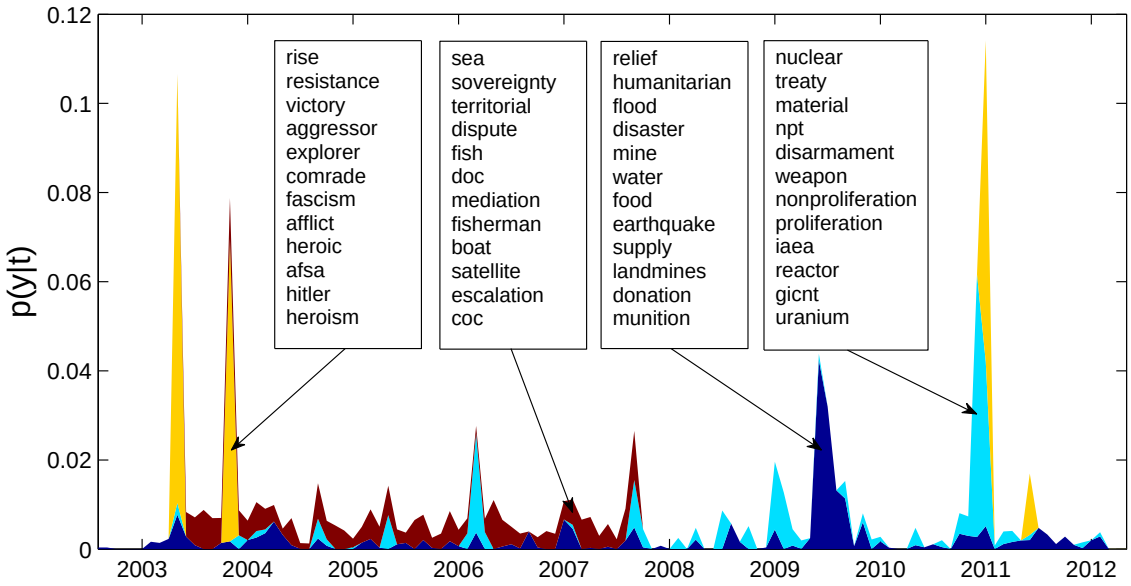
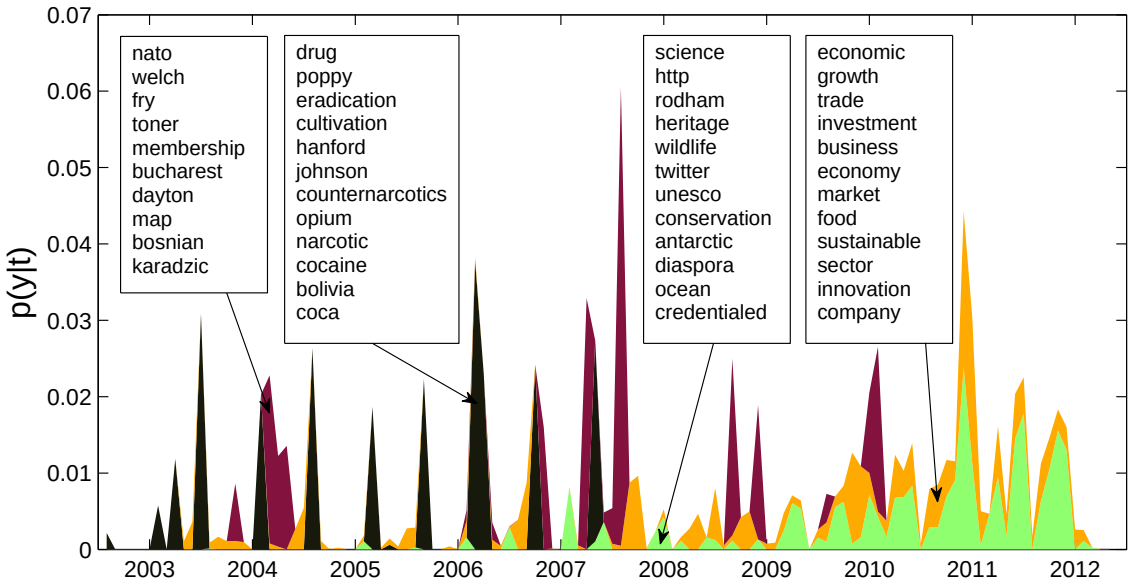
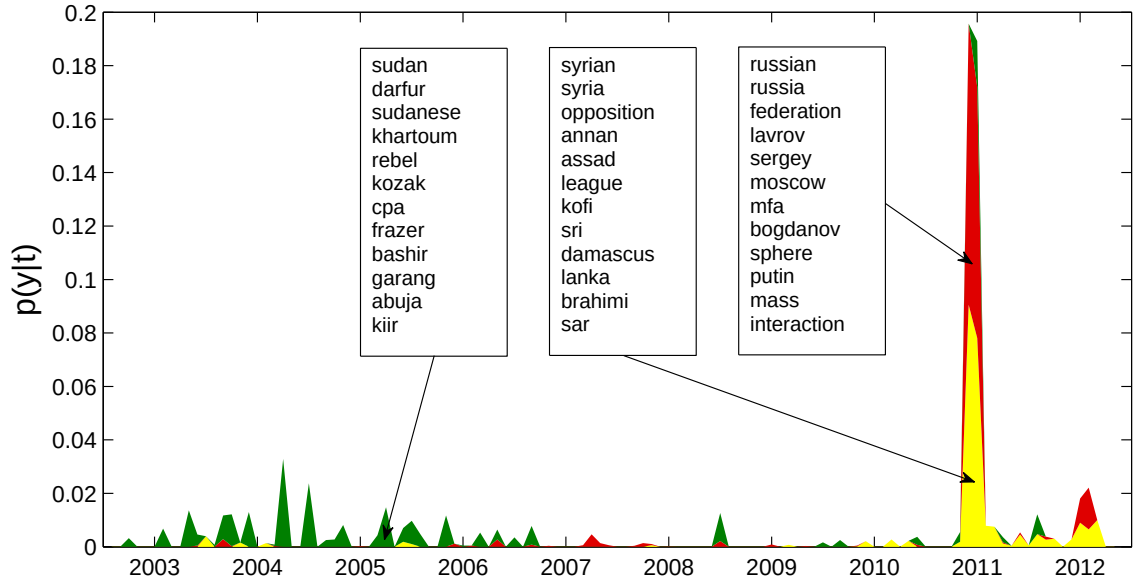
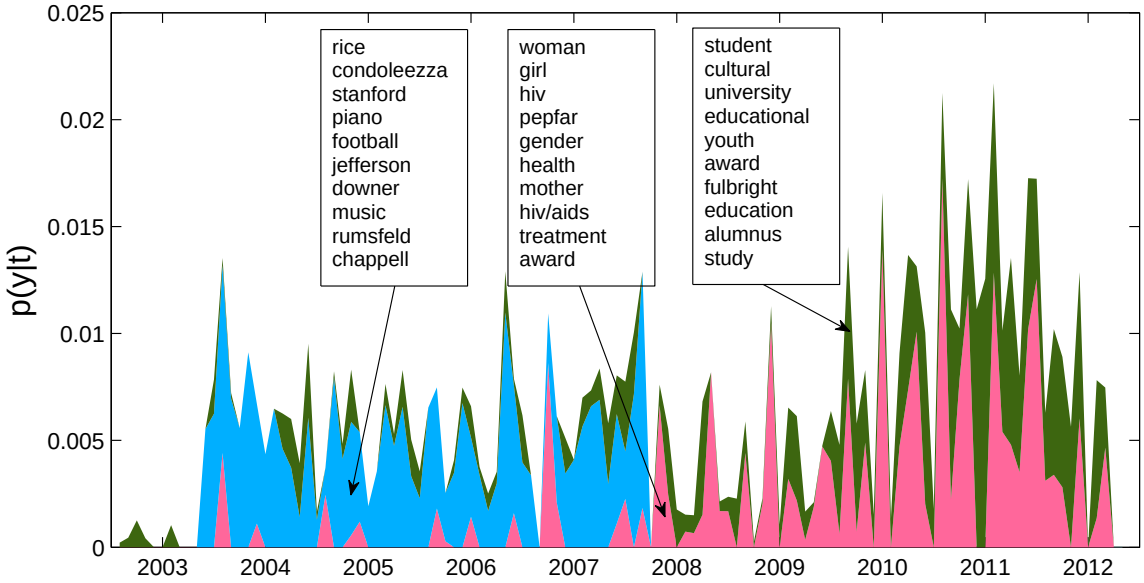


**Рис. 6: Визуализация набора тем для всей коллекции.**
Для каждой темы приведен список самых вероятных слов в ней и распределения моментов времени $p(y|t)$.




p(y|t)
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
sudan
darfur
sudanese
khartoum
rebel
kozak
cpa
frazer
bashir
garang
abuja
kiir
syrian
syria
opposition
annan
assad
league
kofi
sri
damascus
lanka
brahimi
sar
russian
russia
federation
lavrov
sergey
moscow
mfa
bogdanov
sphere
putin
mass
interaction


p(y|t)
0.025
0.02
0.015
0.01
0.005
0
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
rice
condoleezza
stanford
piano
football
jefferson
downer
music
rumsfeld
chappell
woman
girl
hiv
pepfar
gender
health
mother
hiv/aids
treatment
award
student
cultural
university
educational
youth
award
fulbright
education
alumnus
study

# 5 Нормировка коэффициентов регуляризации

При фиксированных коэффициентах регуляризации $\{\tau_i\}$ задача максимизации регуляризованного правдоподобия

$$\log \mathscr{L}(\Phi, \Theta) \;+\; \sum_i \tau_i R_i(\Phi, \Theta) \longrightarrow \max_{\Phi,\Theta}$$

решается итерационным методом — EM-алгоритмом.

Естественным вопросом встает выбор коэффиентов $\{\tau_i\}$ и интерпретация их значений. В общем случае требуется введение критерия или набора критериев, требующих оптимизации, после чего коэффициенты подбираются таким образом, чтобы достичь оптимума.

Здесь мы обсудим возможную репараметризацию коэффициентов, придающую им интерпретируемое значение и упрощающую процесс настройки тематической модели.

## 5.1 Регуляризаторы сглаживания и разреживания

Рассмотрим формулу M-шага (10), сглаживающую распределение $\phi_{wt}$:

$$\phi_{wt} \;\propto\; n_{wt} \;+\; \tau\beta_w \quad \Longleftrightarrow$$

$$\Longleftrightarrow \quad \phi_{wt} \;=\; \frac{n_{wt} + \tau\beta_w}{\sum_{w \in W} (n_{wt} + \tau\beta_w)} \;=\; \frac{n_{wt} + \tau\beta_w}{n_t + \tau}.$$

Интуитивно, мы хотим притянуть распределение $n_{wt}/n_t$, полученное как оценка максимального правдоподобия, к априорному распределению $\beta_w$ с некоторым весом $\lambda$.

Попробуем представим $\phi_{wt}$ как выпуклую комбинацию этих двух распределений:

$$\phi_{wt} \;=\; (1-\lambda)\frac{n_{wt}}{n_t} \;+\; \lambda\beta_w, \qquad \text{где } 0 \leqslant \lambda \leqslant 1.$$

Крайнему значению $\lambda = 0$ соответствует отсутствие регуляризации, а при $\lambda = 1$ достигаем максимального сглаживания: распределение $\phi_{wt}$ становится равным априорному распределению $\beta_w$.

Приравняем теперь оба представления $\phi_{wt}$ и выразим коэффициент $\tau$ через значение $\lambda$:

$$\phi_{wt} \;=\; \frac{n_{wt} + \tau\beta_w}{n_t + \tau} \;=\; (1-\lambda)\frac{n_{wt}}{n_t} + \lambda\beta_w; \qquad \Rightarrow \qquad \tau \;=\; \frac{n_t\lambda}{1-\lambda}.$$

Таким образом мы получили новое представление M-шага:

$$\phi_{wt} \;\propto\; n_{wt} \;+\; n_t\frac{\lambda}{1-\lambda}\beta_w,$$

счётчик $n_t$ оценивает число слов в коллекции, принадлежащих теме $t$, а величина $\frac{\lambda}{1-\lambda}$ определяет *во сколько раз* регуляризатор сглаживания влияет на оценку $\phi_{wt}$ больше, чем коллекция.

Новый коэффициент регуляризации зависит от темы: чем больше значение $n_t$, тем сильнее будет происходить сглаживание этой темы.

Возможная альтернатива здесь заключается в том, чтобы усреднить коэффициент регуляризации по всем темам. Получим следующее выражение:

$$\phi_{wt} \;\propto\; n_{wt} \;+\; \frac{n}{|T|}\frac{\lambda}{1-\lambda}\beta_w,$$

в котором коэффициент регуляризации уже не зависит от темы $t$.

В случае регуляризатора разреживания применимы те же рассуждения.
Оценку для $p(w|t)$ М-шага:

$$\phi_{wt} \;\propto\; \Big(n_{wt} \;-\; \tau\beta_w\Big)_+,$$

мы представим в виде барицентрической комбинации двух известных распределений:

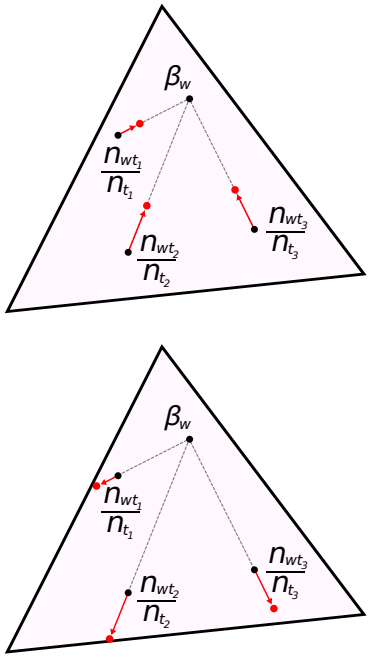
$$\phi_{wt} \;=\; (1-\lambda)\frac{n_{wt}}{n_t} \;+\; \lambda\beta_w, \qquad \lambda \in \mathbb{R}.$$

Получим следующую репараметризацию:

$$\phi_{wt} \;\propto\; \Big(n_{wt} \;-\; n_t\frac{\lambda}{1+\lambda}\;\beta_w\Big)_+, \quad \lambda > 0,$$

в которой величина $\frac{\lambda}{1+\lambda}$ имеет ту же ясную интерпретацию: *во сколько раз* регуляризатор разреживания влияет на оценку $\phi_{wt}$ больше, чем коллекция.

**Рис. 7: Иллюстрация работы регуляризаторов.**
Сверху: сглаживание, снизу: разреживание.




На **рис. 7** показана геометрическая иллюстрация работы регуляризаторов разреживания и сглаживания.

Множество всех распределений над словарем $W$ образует *симплекс тем* в $|W|$-мерном пространстве:

$$\triangle^{|W|-1} \;=\; \Big\{x \in \mathbb{R}^{|W|} \;\Big|\; x_k \geqslant 0, \sum_k x_k = 1\Big\}.$$

Каждая точка этого симплекса соответствует некоторому распределению.

Регуляризатор сглаживания на М-шаге смещает оценку максимального правдоподобия $\frac{n_{wt}}{n_t}$ в сторону априорного распределения $\beta_w$ на вектор, длина которого пропорциональна расстоянию между распределениями, с коэффициентом пропорциональности $\lambda$.

Регуляризатор разреживания, наоборот, отдаляет оценку максимального правдоподобия от заданного априорного распределения. В случае, когда вектор сдвига слишком большой, оценка остается на границе симплекса благодаря операции положительной срезки, при этом некоторые компоненты распределения зануляются.

## 5.2 Общий случай

Интерпретируемая репараметризация возможна и в общем случае $r$ гладких регуляризаторов:

$$\phi_{wt} \;\propto\; \left(n_{wt} \;+\; \sum_{i=1}^{k} \tau_i \phi_{wt} \frac{\partial R_i}{\partial \phi_{wt}}\right)_+.$$

Для каждого регуляризатора $R_i(\Phi, \Theta)$, $1 \leqslant i \leqslant r$ определим следующие величины:

- $r_{it} \;\equiv\; \sum\limits_{w \in W} \left|\phi_{wt} \frac{\partial R_i}{\partial \phi_{wt}}\right|$ — воздействие регуляризатора на тему $t$;
- $r_i \;\equiv\; \sum\limits_{t \in T} r_{it}$ — суммарное воздействие регуляризатора на коллекцию.

Формулу М-шага запишем в следующем виде:

$$\phi_{wt} \;\propto\; \left(n_{wt} \;+\; \sum_{i=1}^{k} \boldsymbol{\tau_i} \Big(\gamma_i \frac{n_t}{r_{it}} \;+\; (1-\gamma_i)\frac{n}{r_i}\Big) \phi_{wt} \frac{\partial R_i}{\partial \phi_{wt}}\right)_+,$$

где $\boldsymbol{\tau_i}$ — новый коэффициент регуляризации, показывающий *во сколько раз* соответствующий регуляризатор влияет на оценку $p(w|t)$ больше, чем коллекция;

$\gamma_i \in [0, 1]$ — *степень индивидуализации*, задающая наше предпочтение между равномерной регуляризацией по всем темам ($\gamma_i = 0$) и индивидуальным подходом к каждой теме ($\gamma_i = 1$).

Аналогичный подход возможен и для оценок $p(t|d)$:

$$\theta_{td} \;\propto\; \left(n_{td} \;+\; \sum_{i=1}^{k} \boldsymbol{\tau_i} \Big(\gamma_i \frac{n_t}{r_{id}} \;+\; (1-\gamma_i)\frac{n}{r_i}\Big) \theta_{td} \frac{\partial R_i}{\partial \theta_{td}}\right)_+,$$

где $r_{id} \;\equiv\; \sum\limits_{t \in T} \left|\theta_{td} \frac{\partial R_i}{\partial \theta_{td}}\right|$ — воздействие регуляризатора на документ $d$,
$r_i \;\equiv\; \sum\limits_{t \in D} r_{id}$ — суммарное воздействие регуляризатора на коллекцию.

## 5.3 Эксперименты

Эксперименты с отнормированными коэффициентами регуляризации проводились на коллекции пресс-релизов, из которой было взято первых 1000 документов.

Были зафиксированы подобранные коэффициенты регуляризаторов сглаживания распределений $p(w|t)$ и $p(t|d)$ для фоновых тем и разреживания $p(t|d)$ предметных тем.

Значение коэффициента разреживания распределений $p(w|t)$ для предметных тем варьировалось по сетке.

Для каждого набора значений коэффициентов регуляризации строилась тематическая модель и подсчитывалась *когерентность* найденных тем.

Число тем задавалось: $|T| = 50, 100, 150, 200$.

На **рис. 8** показано сравнение моделей до и после нормировки коэффициента.

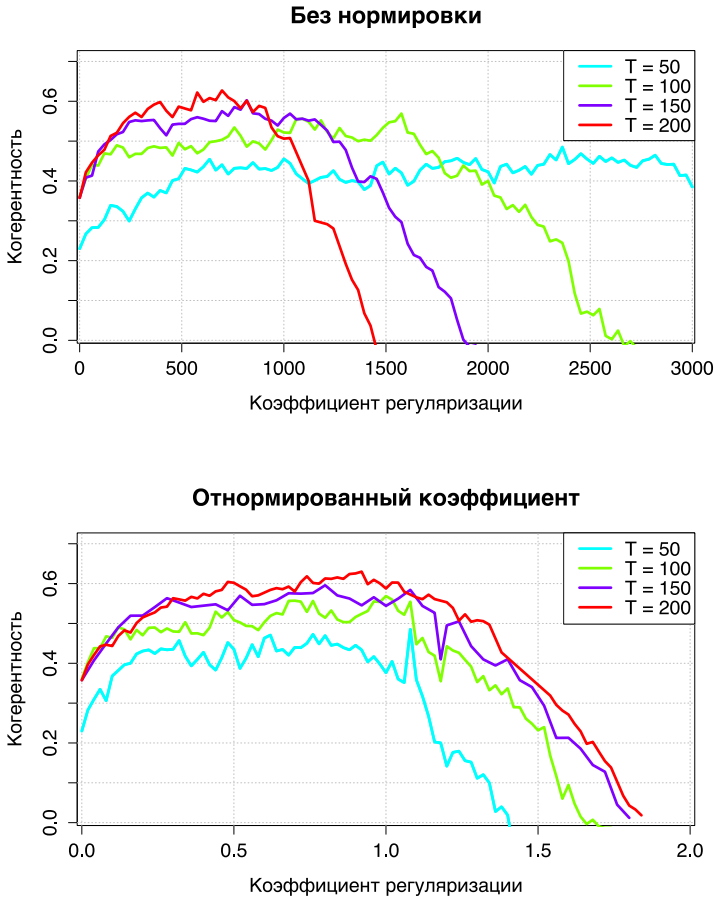



**Рис. 8:** Сверху: ненормированный коэффициент регуляризатора разреживания, снизу: после репараметризации.

Видно, что в общем случае коэффициент регуляризации может принимать большие значения, плохо поддающиеся интерпретации. При изменении числа тем в модели, область оптимальных значений коэффициентов изменяется.

После предложенной нормировки величина коэффициента показывает *во сколько раз* регуляризатор влияет на оценку сильнее, чем коллекция, а область оптимальных значений становится более устойчивой к изменению остальных параметров.

На **рис. 9** показана зависимость качества модели от коэффициента регуляризации для разных значений параметра $\gamma$.

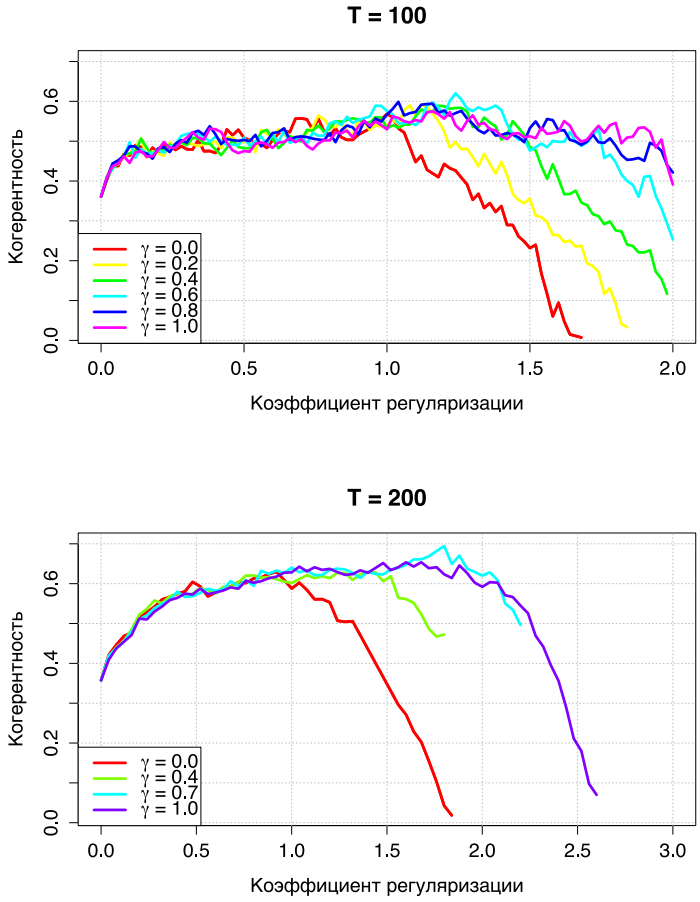



**Рис. 9:** Влияние *степени индивидуализации* $\gamma$. Сверху: 100 тем, снизу: 200 тем.

При увеличении $\gamma$ область оптимальных значений коэффициента регуляризации становится больше, увеличивается стабильность в настройке модели.

# 6 Заключение

В данной работе предлагается аддитивно регуляризованная темпоральная тематическая модель, учитывающая метки времени документов.

Использование дополнительной априорной информации позволяет улучшить качество модели, что демонстрируется на коллекции пресс-релизов министерств иностранных дел ряда стран за промежуток времени длиной более десяти лет.

Основным теоретическим результатом работы являются формулы регуляризованного EM-алгоритма для негладкой задачи оптимизации.

Также в работе предложен адаптивный способ репараметризации коэффициентов регуляризации, позволяющий сделать процесс настройки модели более устойчивым, и сбалансировать воздействия регуляризаторов на отдельные темы и документы.

# Список литературы


[1] *Ahmed A., Xing E. P.* Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream // *arXiv preprint arXiv:1203.3463.* — 2012.

[2] *AlSumait L., Barbará D., Domeniconi C.* On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking // Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on / IEEE. — 2008. — Pp. 3–12.

[3] *Beykikhoshk A., Arandjelović O., Venkatesh S., Phung D.* Hierarchical dirichlet process for tracking complex topical structure evolution and its application to autism research literature // Advances in Knowledge Discovery and Data Mining. — Springer, 2015. — Pp. 550–562.

[4] *Blei D. M., Lafferty J. D.* Dynamic topic models // Proceedings of the 23rd international conference on Machine learning / ACM. — 2006. — Pp. 113–120.

[5] *Blei D. M., Ng A., Jordan M.* Latent dirichlet allocation // *JMLR.* — 2003. — Vol. 3. — Pp. 993–1022.

[6] *Boyd S., Vandenberghe L.* Convex optimization. — Cambridge university press, 2004.

[7] *Chuang J., Gupta S., Manning C., Heer J.* Topic model diagnostics: Assessing domain relevance via topical alignment // Proceedings of the 30th International Conference on Machine Learning (ICML-13). — 2013. — Pp. 612–620.

[8] *Cichocki A., Zdunek R., Amari S.-i.* Hierarchical als algorithms for nonnegative matrix and 3d tensor factorization // Independent Component Analysis and Signal Separation. — Springer, 2007. — Pp. 169–176.

[9] *Dubey A., Hefny A., Williamson S., Xing E. P.* A nonparametric mixture model for topic modeling over time. // SDM / SIAM. — 2013. — Pp. 530–538.

[10] *Gohr A., Hinneburg A., Schult R., Spiliopoulou M.* Topic evolution in a stream of documents. — 2009.

[11] *Griffiths T. L., Steyvers M.* Finding scientific topics // *Proceedings of the National Academy of Sciences.* — 2004. — Vol. 101, no. suppl 1. — Pp. 5228–5235.

[12] *Hall D., Jurafsky D., Manning C. D.* Studying the history of ideas using topic models // Proceedings of the conference on empirical methods in natural language processing / Association for Computational Linguistics. — 2008. — Pp. 363–371.

[13] *Hofmann T.* Probabilistic latent semantic analysis // Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence / Morgan Kaufmann Publishers Inc. — 1999. — Pp. 289–296.

[14] *Kim S.-J., Koh K., Boyd S., Gorinevsky D.* L1 trend filtering // *SIAM review.* — 2009. — Vol. 51, no. 2. — Pp. 339–360.

[15] *Kuhn H. W.* The hungarian method for the assignment problem // *Naval research logistics quarterly.* — 1955. — Vol. 2, no. 1-2. — Pp. 83–97.

[16] *Lee D. D., Seung H. S.* Algorithms for non-negative matrix factorization // Advances in neural information processing systems. — 2001. — Pp. 556–562.

[17] *Li W., Wang X., McCallum A.* A continuous-time model of topic co-occurrence trends. — Defense Technical Information Center, 2006.

[18] *Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A.* Optimizing semantic coherence in topic models // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — 2011. — Pp. 262–272.

[19] *Murphy K. P.* Machine learning: a probabilistic perspective. — MIT press, 2012.

[20] *Naveed N., Sizov S., Staab S.* Att: Analyzing temporal dynamics of topics and authors in social media // Proceedings of the 3rd International Web Science Conference / ACM. — 2011. — P. 1.

[21] *Newman D., Bonilla E. V., Buntine W.* Improving topic coherence with regularized topic models // Advances in neural information processing systems. — 2011. — Pp. 496–504.

[22] *Newman D., Lau J. H., Grieser K., Baldwin T.* Automatic evaluation of topic coherence // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — 2010. — Pp. 100–108.

[23] *Ren L., Dunson D. B., Carin L.* The dynamic hierarchical dirichlet process // Proceedings of the 25th international conference on Machine learning / ACM. — 2008. — Pp. 824–831.

[24] *Schmidt M., Fung G., Rosales R.* Fast optimization methods for l1 regularization: A comparative study and two new approaches // Machine Learning: ECML 2007. — Springer, 2007. — Pp. 286–297.

[25] *Stevens K., Kegelmeyer P., Andrzejewski D., Buttler D.* Exploring topic coherence over many models and many topics // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning / Association for Computational Linguistics. — 2012. — Pp. 952–961.

[26] *Vorontsov K. V.* Additive regularization for topic models of text collections // Doklady Mathematics / Pleiades Publishing. — Vol. 89. — 2014. — Pp. 301–304.

[27] *Vorontsov K. V., Potapenko A. A.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // Analysis of Images, Social Networks and Texts. — Springer, 2014. — Pp. 29–46.

[28] *Wang C., Blei D., Heckerman D.* Continuous time dynamic topic models // *arXiv preprint arXiv:1206.3298*. — 2012.

[29] *Wang X., McCallum A.* Topics over time: a non-markov continuous-time model of topical trends // Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining / ACM. — 2006. — Pp. 424–433.

[30] *Wei X., Sun J., Wang X.* Dynamic mixture models for multiple time-series. // IJCAI. — Vol. 7. — 2007. — Pp. 2909–2914.

[31] *Wright S. J., Nocedal J.* Numerical optimization. — Springer New York, 1999. — Vol. 2.

[32] *Zhang J., Song Y., Zhang C., Liu S.* Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora // Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining / ACM. — 2010. — Pp. 1079–1088.