• Вероятностные языковые модели • Лекция 5. Конструирование регуляризаторов и устойчивость тематических моделей

Константин Вячеславович Воронцов
k.vorontsov@iai.msu.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные языковые модели (курс лекций, К.В.Воронцов)»

ВМК МГУ • 30 марта 2026

Содержание

Напоминание. Тематическая модель «мешка термов»

Дано: коллекция текстовых документов D, словарь W;
n_{dw} — частота терма $w \in W$ в документе $d \in D$.

Найти: вероятностную языковую модель $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$
с параметрами $\phi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Напоминание. Тематическая модель локальных контекстов

Дано: последовательность $w_1, \ldots, w_n$ термов словаря W;
$C_i \subset \{1, \ldots, n\}$ — локальный контекст терма w_i, $1, \ldots, n$;
α_{ci} — коэффициент внимания, вес терма w_c из C_i для w_i.

Найти: вер. языковую модель $p(w|C_i) = \sum_{t \in T} \phi_{tw} \frac{p(w)}{p(t)} p(t|C_i)$
с параметрами $\phi_{tw} = p(t|w)$

Критерий: $\sum_{i=1}^{n} \ln \sum_{t \in T} \phi_{tw_i} \frac{p(w_i)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c} + R(\Phi) \to \max_{\Phi}$

EM-алгоритм (после некоторых насильственных упрощений):

$$\text{E-шаг:} \begin{cases} p_{ti} = \underset{t \in T}{\text{norm}} \left(\frac{\phi_{tw_i}}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c} \right), & p(t) = \sum_{w \in W} \phi_{tw} p(w) \end{cases}$$

$$\text{M-шаг:} \begin{cases} \phi_{tw} = \underset{t \in T}{\text{norm}} \left(n_{tw} + \phi_{tw} \frac{\partial R}{\partial \phi_{tw}} \right), & n_{tw} = \sum_{i=1}^{n} p_{ti} [w_i = w] \end{cases}$$

Напоминание. Обобщённая модель LDA

Сглаживание ($\beta_{wt} > 0$, $\alpha_{td} > 0$) и разреживание ($\beta_{wt} < 0$, $\alpha_{td} < 0$):

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td}$$

Сглаживание фоновой темы t_Φ с общей лексикой языка:

- $\beta_{wt_\Phi} = \beta_0 p_\Phi(w)$ — тема t_Φ похожа на заданное $p_\Phi(w)$
- $\alpha_{t_\Phi d} = \alpha_0$ — общая лексика есть в каждом документе d

Сглаживание по «белым спискам» (seed words, seed topics):

- $\beta_{wt} = \beta_0 [w \in W_t]$ — термы из W_t должны быть в t
- $\alpha_{td} = \alpha_0 [t \in T_d]$ — темы из T_d должны быть в d

Разреживание по «чёрным спискам»:

- $\beta_{wt} = -\beta_0 [w \in W_t]$ — термов из W_t не должно быть в t
- $\alpha_{td} = -\alpha_0 [t \in T_d]$ — тем из T_d не должно быть в d

Есть ли проблема $\ln 0$ при разреживании распределений?

В регуляризаторе сглаживания/разреживания

$$R(\Phi) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} \to \max$$

не возникает ли проблема с $\ln \phi_{wt}$ при $\phi_{wt} = 0$ или $\phi_{wt} \to 0$?

Подправим регуляризатор, при сколь угодно малом ε:

$$R(\Phi) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln(\phi_{wt} + \varepsilon) \to \max .$$

Подставив в формулу M-шага, получим для всех $t \in T$:

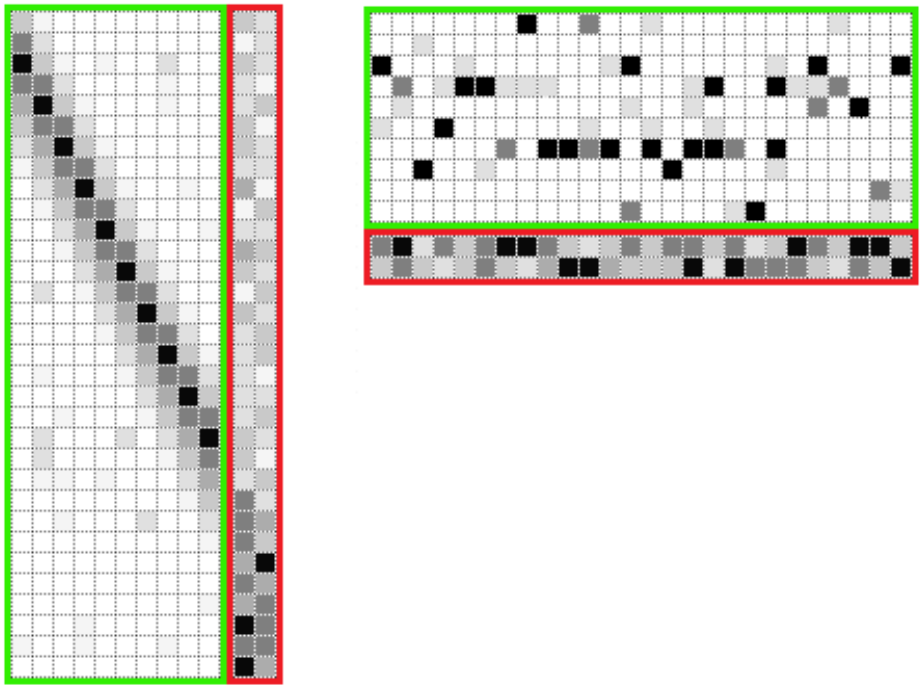
$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}} \left(n_{wt} + \beta_0 \beta_{wt} \frac{\phi_{wt}}{\phi_{wt} + \varepsilon} \right) \xrightarrow{\varepsilon \to 0} \underset{w \in W}{\operatorname{norm}} \left(n_{wt} + \beta_0 \beta_{wt} [\phi_{wt} \neq 0] \right),$$

Если $\phi_{wt} = 0$, то и на последующих итерациях $n_{wt} = \phi_{wt} = 0$.

Разделение тем на предметные и фоновые

Предметные темы S содержат термины предметной области,
$p(w|t)$, $p(t|d)$, $t \in S$ — разреженные, существенно различные
Фоновые темы B содержат слова общей лексики,
$p(w|t)$, $p(t|d)$, $t \in B$ — существенно отличные от нуля

$\Phi_{W \times T}$ $\Theta_{T \times D}$



Регуляризатор декоррелирования тем

Цели: усилить различность тем; выделить в каждой теме лексическое ядро, отличающее её от других тем; способствовать переходу общей лексики в фоновые темы

Критерий: минимум ковариаций между вектор-столбцами ϕ_t

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

Подставляем в формулы M-шага, получаем ещё один вариант разреживания — контрастирование строк матрицы Φ (малые вероятности ϕ_{wt} в строке становятся ещё меньше):

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

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation. 2010

Разреживающий регуляризатор для отбора тем

Цель: избавиться от незначимых тем (topic selection)

Разреживаем распределение $p(t) = \sum_d p(d)\,\theta_{td}$, максимизируя кросс-энтропию между $p(t)$ и равномерным распределением:

$$R(\Theta) = -\tau \sum_{t \in T} \ln \sum_{d \in D} p(d)\,\theta_{td} \to \max$$

Подставляем в M-шаг R, затем несмещённую оценку $\theta^*_{td} = \frac{n_d}{n_t}$:

$$\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} - \tau \frac{n_d}{n_t}\theta_{td}\right) = \underset{t \in T}{\text{norm}}\left(n_{td}\left(1 - \frac{\tau}{n_t}\right)\right)$$

Эффект: обнуляются строки матрицы Θ с малыми n_t, заодно удаляются зависимые и расщеплённые темы (см. далее)

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization. SLDS 2015.

Дивергенция Кульбака–Лейблера и её свойства

Два дискретных распределения: $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$:

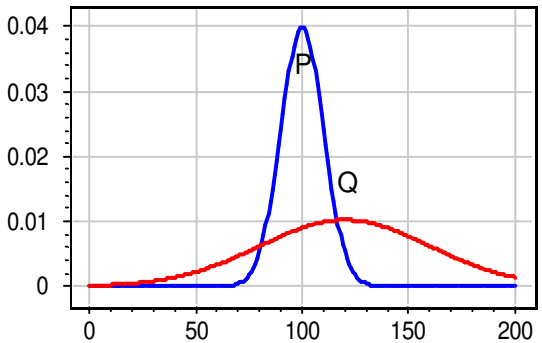
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}$$

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

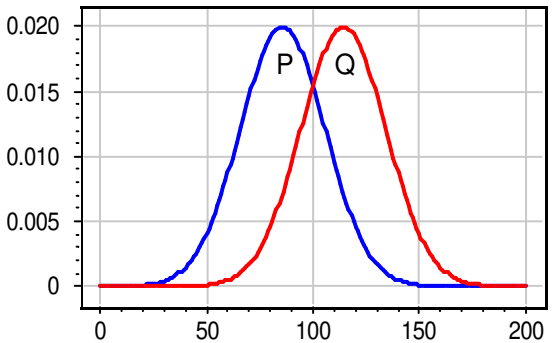
2. min KL $\Leftrightarrow$ min кросс-энтропии $\Leftrightarrow$ max правдоподобия:

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

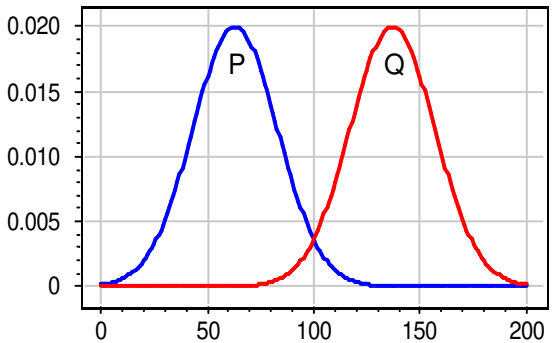
3. $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P) \;\Rightarrow$ скорее P вложено в Q, чем Q в P:



$\mathrm{KL}(P\|Q) = 0.44$
$\mathrm{KL}(Q\|P) = 2.97$



$\mathrm{KL}(P\|Q) = 0.44$
$\mathrm{KL}(Q\|P) = 0.44$



$\mathrm{KL}(P\|Q) = 2.97$
$\mathrm{KL}(Q\|P) = 2.97$

Меры различия дискретных вероятностных распределений

Расстояния — симметричные функции, $\rho \in [0,1]$:

- $\rho(P,Q) = 1 - \dfrac{\sum_i p_i q_i}{\left(\sum_i p_i^2\right)^{1/2}\left(\sum_i q_i^2\right)^{1/2}}$ — косинусное расст.
- $\rho^2(P,Q) = 1 - \sum_i \sqrt{p_i q_i}$ — расст. Хеллингера
- $\rho(P,Q) = 1 - |S_P \cap S_Q| : |S_P \cup S_Q|$ — расст. Жаккара,
  где $S_X = \{i : x_i \geqslant \delta\}$ — множество существенных исходов X

Дивергенции — несимметричные меры вложенности P в Q:

- $\mathrm{KL}(P\|Q) = \sum_i p_i \ln\left(\frac{p_i}{q_i}\right)$ — див. Кульбака–Лейблера
- $D_\lambda(P\|Q) = \frac{1}{\lambda(\lambda+1)} \sum_i p_i \left(\left(\frac{p_i}{q_i}\right)^\lambda - 1\right)$ — див. Кресси–Рида
- $D_\alpha(P\|Q) = \frac{1}{\alpha-1} \log \sum_i p_i \left(\frac{p_i}{q_i}\right)^{\alpha-1}$ — див. Рéньи,

при увеличении λ, α игнорируются малые вероятности
при $\lambda \to 0$, $\alpha \to 1$ монотонной функцией связаны с KL

Расстояния и дивергенции для построения регуляризаторов

- сходство или различие распределений — одно из самых частых требований при конструировании регуляризаторов
- *KL-дивергенция* — наиболее частый выбор благодаря связи с принципом максимума правдоподобия; запомним: $\mathrm{KL}\big(P \text{ эмпирическое} \,\big\|\, Q(\alpha) \text{ параметрическое}\big)$
- *косинусное расстояние* — лучшее для векторного поиска
- *дивергенции Реньи* и *Кресси–Рида* обобщают KL, вводят управляющий параметр, который можно оптимизировать
- *евклидово расстояние* не рекомендуется использовать для сравнения дискретных распределений
- *расстояние Хеллингера* похоже на него, но корректнее сравнивает малые вероятности
- *расстояние Жаккара* — простое, но не дифференцируемое
- окончательный выбор — по результатам экспериментов

Включение и отключение регуляризаторов

- регуляризация ведёт итерационный процесс к матричному разложению с требуемыми свойствами, но даёт смещённые оценки Φ, Θ. На последней итерации можно отключать регуляризатор, возвращая несмещённые PLSA-оценки:

$$\phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt}\right) \qquad \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td}\right)$$

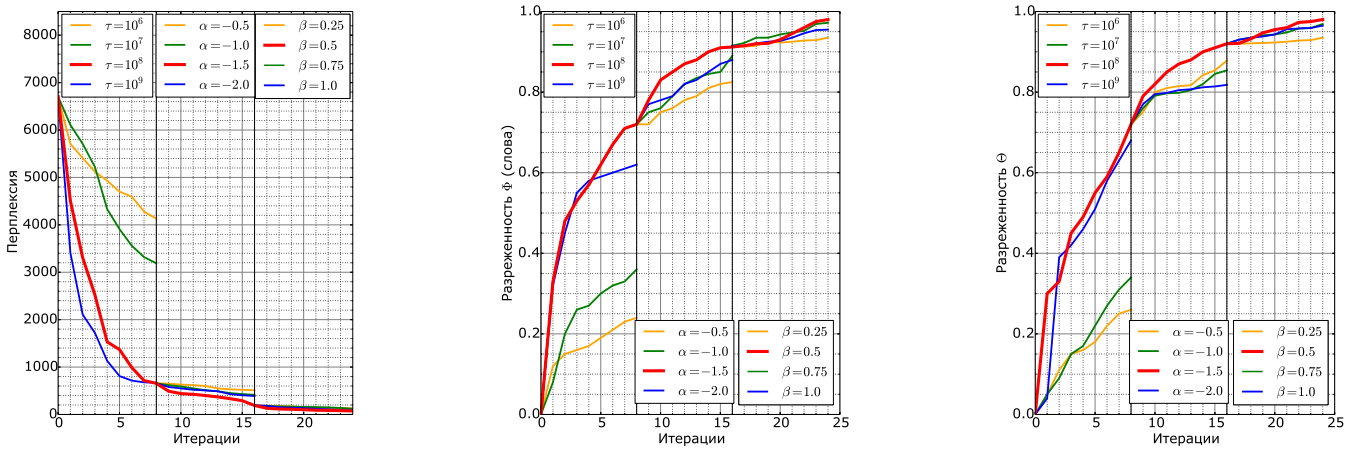
- коэффициенты регуляризации можно менять в итерациях
- регуляризаторы можно включать не сразу или по очереди
- регуляризаторы можно отключать по достижению эффекта
- одни регуляризаторы могут выполнять подготовительную работу для применения следующих регуляризаторов
- **открытая проблема:** выбор стратегии регуляризации

А.Кузьмин. Адаптивный выбор траектории регуляризации. МФТИ, 2017.

M.Khodorchenko, S.Teryoshkin, T.Sokhin, N.Butakov. Optimization of learning strategies for ARTM-based topic models. 2020.

Управление траекторией регуляризации

1. задать диапазон и сетку значений каждого τ_k (удобно использовать относительные коэффициенты $\tilde{\tau}_k$)
2. задать последовательность подключения регуляризаторов (имеются эмпирические рекомендации — см. далее)
3. визуализировать несколько критериев качества (спойлер):



A.O.Янина, K.B.Воронцов. Мультимодальные тематические модели для разведочного поиска в коллективном блоге. JMLDA 2016.

V.Bulatov, E.Egorov, E.Veselova, D.Polyudova, V.Alekseev, A.Goncharov, K.Vorontsov. TopicNet: making additive regularisation for topic modelling accessible. LREC-2020

Относительные коэффициенты регуляризации

Цель: привести все τ_k к рабочему диапазону $\approx [0.05, 1]$

Формула M-шага со взвешенной суммой регуляризаторов R_k:

$$\phi_{wt} = \underset{w \in W}{\text{norm}} \left(n_{wt} + \sum_k \tau_k \phi_{wt} \frac{\partial R_k}{\partial \phi_{wt}} \right)$$

Суммарное воздействие r_{kt} регуляризатора R_k на тему t и *суммарное воздействие* r_k регуляризатора R_k на все темы:

$$r_{kt} = \sum_{w \in W} \left| \phi_{wt} \frac{\partial R_k}{\partial \phi_{wt}} \right|, \qquad r_k = \sum_{t \in T} r_{kt}$$

Относительный коэффициент регуляризации $\tilde{\tau}_k$:

$$\tau_k = \tilde{\tau}_k \frac{n}{r_k} \quad \text{или} \quad \tau_k = \tilde{\tau}_k \frac{n_t}{r_{kt}} \quad \text{или} \quad \tau_k = \tilde{\tau}_k \left(\gamma_k \frac{n_t}{r_{kt}} + (1 - \gamma_k) \frac{n}{r_k} \right),$$

где γ_k — *индивидуализация* воздействия R_k на темы

Правдоподобие и перплексия (perplexity)

Правдоподобие языковой модели $p(w|d)$ (чем выше, тем лучше):

$$\mathscr{L}(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w|d), \qquad p(w|d) = \sum_{t\in T} \phi_{wt}\theta_{td}$$

Перплексия языковой модели $p(w|d)$ (чем меньше, тем лучше):

$$\mathcal{P}(D) = \exp\Big(-\frac{1}{n}\mathscr{L}(\Phi,\Theta)\Big), \quad n = \sum_{d\in D}\sum_{w\in d} n_{dw}$$

Интерпретация перплексии:

- если распределение $p(w|d) = \frac{1}{|W|}$ равномерное, то $\mathcal{P} = |W|$
- мера «удивлённости» модели словам в тексте
- коэффициент ветвления (branching factor) текста
- известные оценки человеческой перплексии: 8–12

Когерентность — численная оценка интерпретируемости,

наиболее сильно коррелирующая с экспертными оценками.

Когерентность (согласованность) темы t по k топовым словам:

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

где w_i — i-е слово в порядке убывания ϕ_{wt},

$\text{PMI}(u, v) = \ln \frac{P_{uv}}{P_u P_v}$ — *поточечная взаимная информация* (pointwise mutual information),

P_{uv} — доля документов, в которых слова u, v хотя бы один раз встречаются рядом (в одном предложении или в окне 10 слов),

P_u — доля документов, в которых u встретился хотя бы 1 раз,

Когерентность модели = средняя когерентность всех тем.

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

Лексическое ядро, чистота и контрастность темы

Лексическое ядро W_t темы t, варианты определения:

- W_t — top-k термов с наибольшими значениями $p(w|t)$
- $W_t = \{w\colon p(w|t) > p(w)\}$
- $W_t = \{w\colon p(w|t) > \frac{1}{|W|}\}$ [Кольцов и др., 2014]
- $W_t = \{w\colon p(t|w) > 0.25\}$ [Воронцов, Потапенко, 2014]

Характеристики лексического ядра темы:

- $|W_t|$ — *размер ядра* темы, для ориентировки: $|W_t| \sim \frac{|W|}{|T|}$
- $\sum\limits_{w\in W_t} p(w|t)$ — *чистота* темы, из $[0,1]$, больше⇒лучше
- $\frac{1}{|W_t|}\sum\limits_{w\in W_t} p(t|w)$ — *контрастность* темы, $[0,1]$, больше⇒лучше

Koltcov S., Koltsova O., Nikolenko S. Latent Dirichlet Allocation: Stability and applications to studies of user-generated content. ACM WebSci, 2014.

Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. AIST, 2014.

Разреживание, сглаживание, декоррелирование, отбор тем

M-шаг при комбинировании 6 регуляризаторов:

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

$$\theta_{td} = \underset{t}{\text{norm}}\Big(n_{td} + \tau_4 \underbrace{\alpha_t[t \in B]}_{\substack{\text{сглаживание}\\ \text{фоновых}\\ \text{тем}}} - \tau_5 \underbrace{\alpha_t[t \in S]}_{\substack{\text{разреживание}\\ \text{предметных}\\ \text{тем}}} - \tau_6 \underbrace{\frac{n_d}{n_t}\theta_{td}}_{\substack{\text{удаление}\\ \text{малых тем}}} \Big)$$

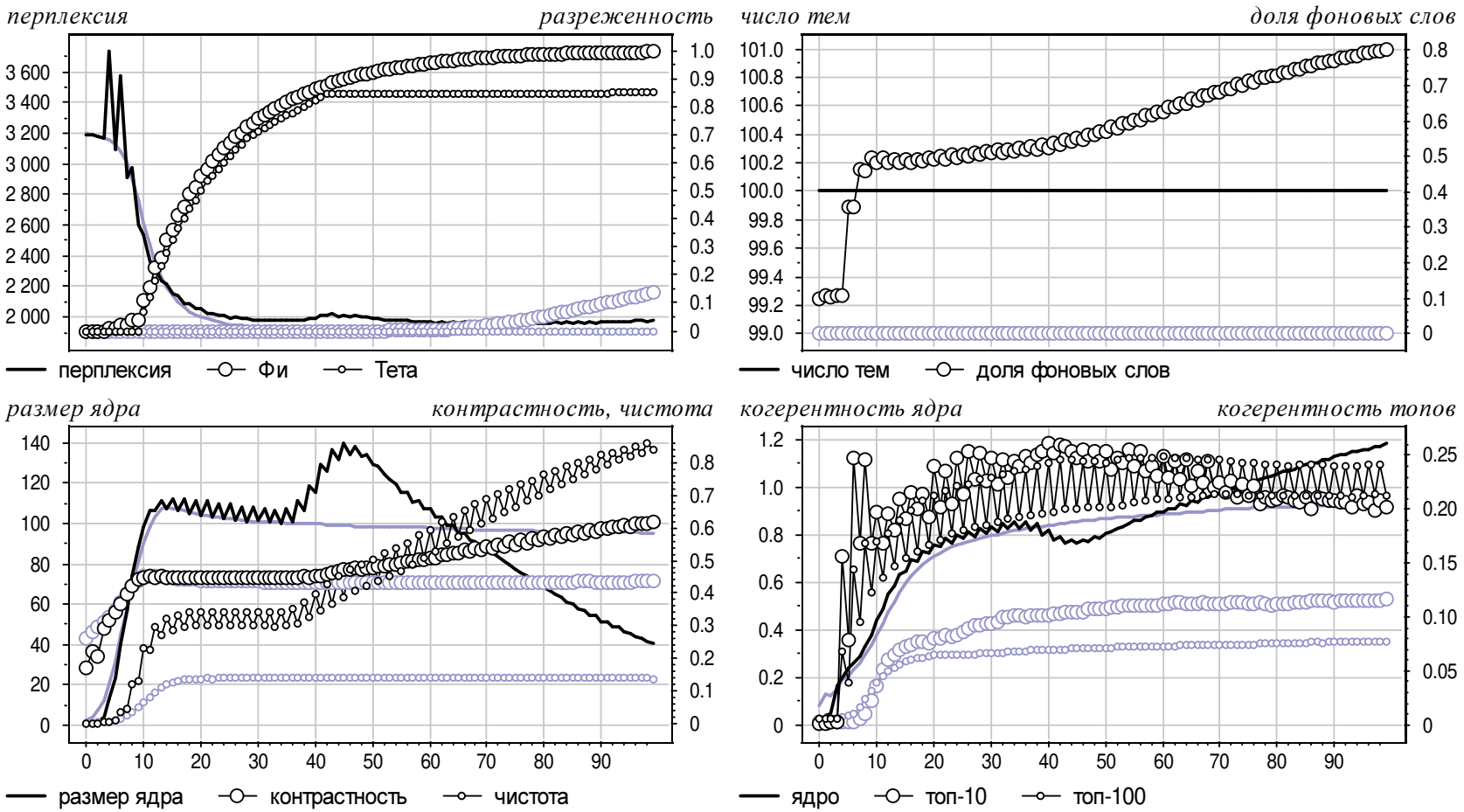
Каждому регуляризатору R_k надо подбирать коэффициент τ_k.

Данные: статьи NIPS (Neural Information Processing System)
$|D| = 1566$ статей, $n = 2.3\,\text{M}$, $|W| = 13\,\text{K}$,
контрольная коллекция: $|D'| = 174$.

Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. AIST'2014.

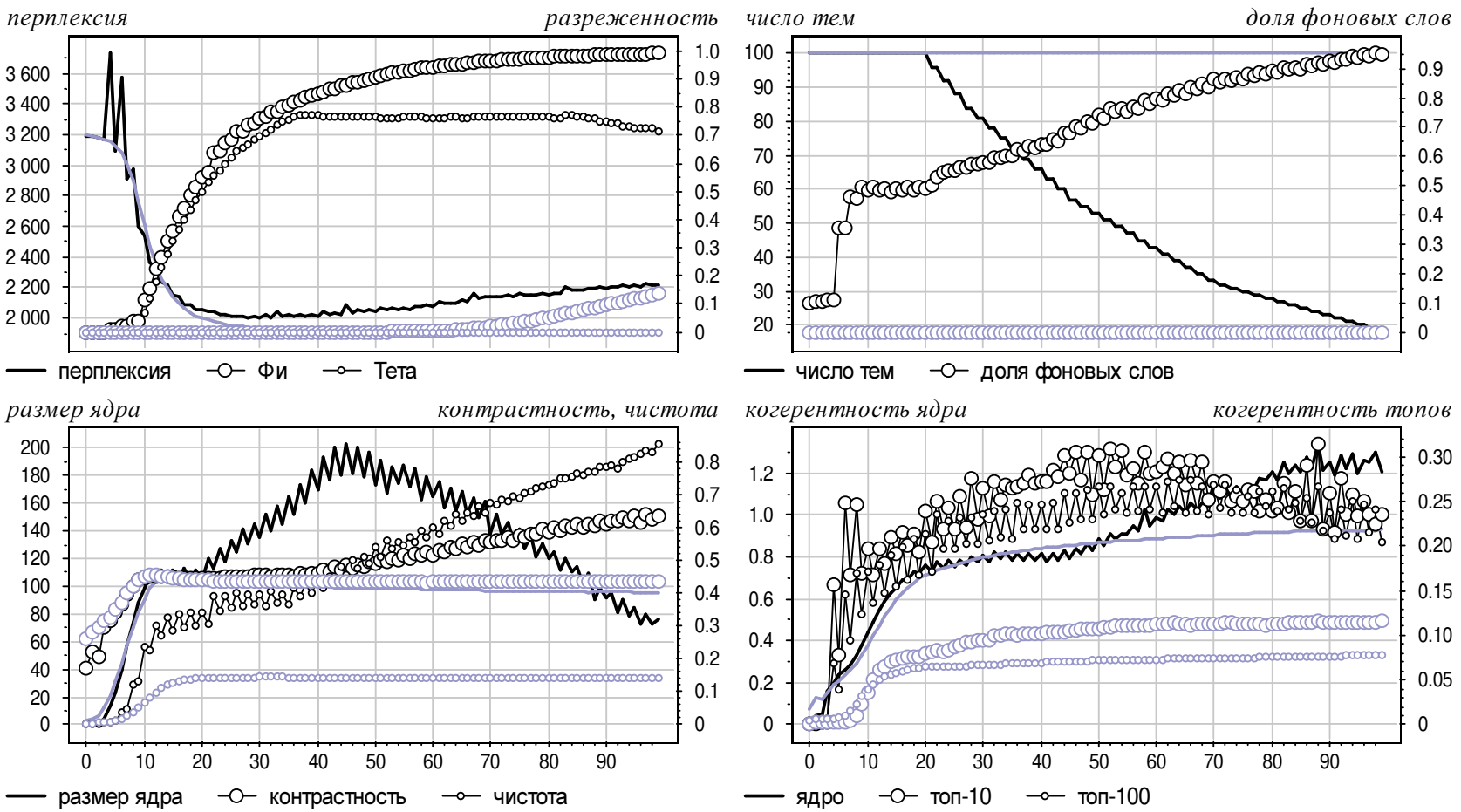
Разреживание, сглаживание, декоррелирование

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



Те же регуляризаторы, плюс отбор тем

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



Выводы из экспериментов

Одновременное улучшение многих критериев качества
при незначительной деградации *перплексии* (правдоподобия):

- *разреженность* выросла от 0 до 95%–98%
- *когерентность тем* выросла от 0.1 до 0.3
- *чистота тем* выросла от 0.15 до 0.8
- *контрастность тем* выросла от 0.4 до 0.6

Рекомендации по выбору *траектории регуляризации*:

- разреживание включать постепенно после 10-20 итераций
- сглаживание можно включать сразу
- декоррелирование лучше включать сразу и сильно
- отбор тем включать постепенно,
- не совмещая с декоррелированием на одной итерации

Способны ли PLSA и LDA восстановить истинные темы?

Матрицы Φ_0 и Θ_0 порождаются распределением Дирихле.
Синтетическая коллекция порождается матрицами Φ_0 и Θ_0.
Размеры: $|D| = 500$, $|W| = 1000$, $|T| = 30$, $n_d \in [100, 600]$.

Цель — сравнить восстановленные распределения $p(i|j)$
с исходными синтетическими распределениями $p_0(i|j)$
по среднему расстоянию Хеллингера:

$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{p(i|j)} - \sqrt{p_0(i|j)} \right)^2},$$

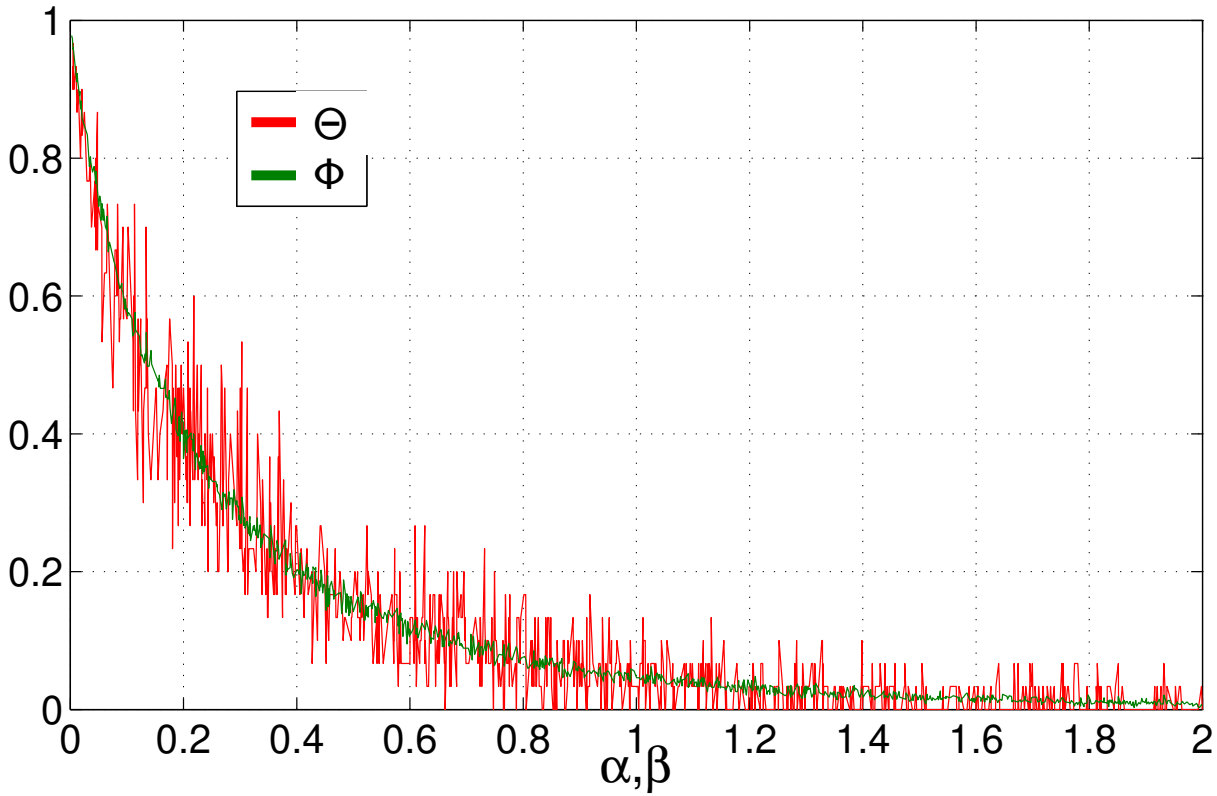
как для самих матриц Φ и Θ, так и для их произведения:

$$D_\Phi = H(\Phi, \Phi_0); \quad D_\Theta = H(\Theta, \Theta_0); \quad D_{\Phi\Theta} = H(\Phi\Theta, \Phi_0\Theta_0).$$

Соответствие между темами находится *венгерским алгоритмом*
(который решает *задачу о назначениях*)

Разреженность векторов, порождаемых распределением Dir

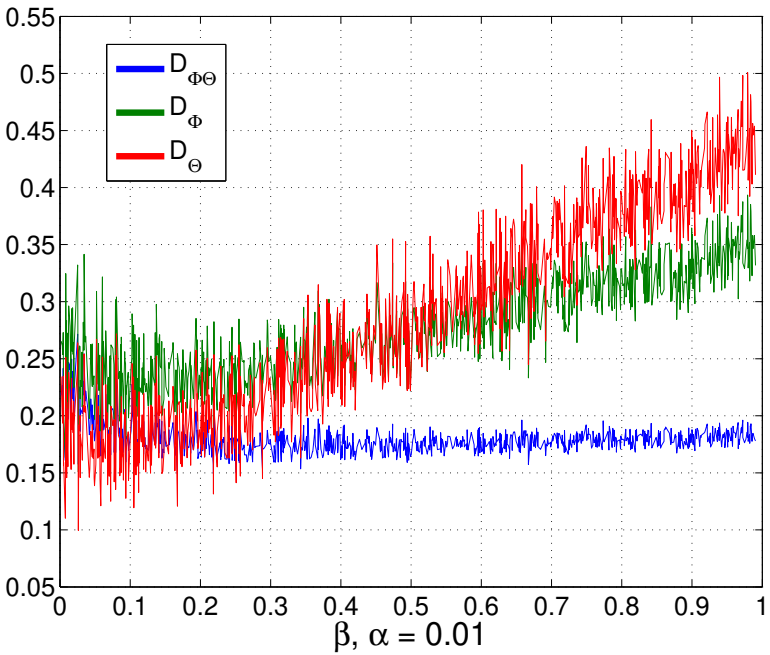
Зависимость разреженности (доли почти нулевых элементов) распределений $\theta_d^0 \sim \mathrm{Dir}(\alpha)$ и $\phi_t^0 \sim \mathrm{Dir}(\beta)$ от параметров α и β симметричного распределения Дирихле:



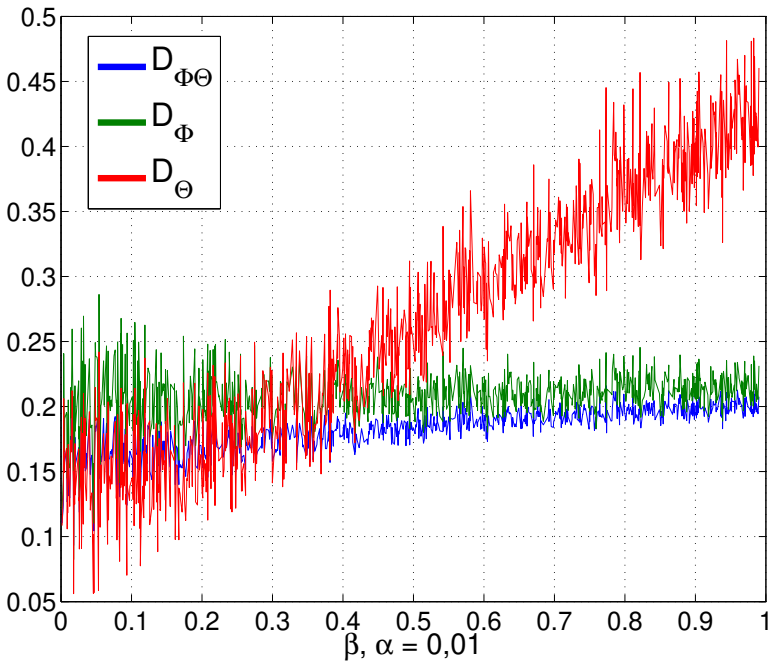
Неустойчивость восстановления матриц Φ и Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от параметра β при фиксированном $\alpha = 0.01$

PLSA



LDA

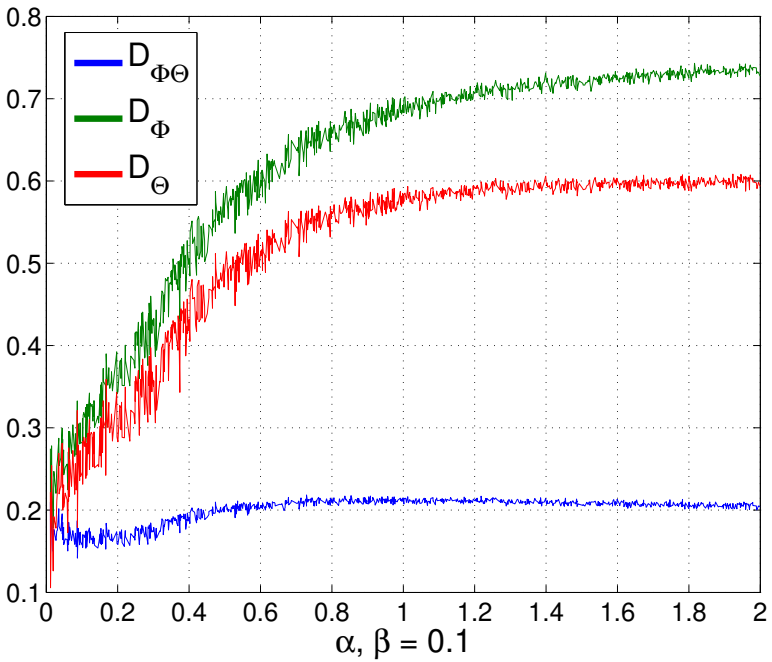


Виталий Глушаченков. Устойчивость матричных разложений в задачах тематического моделирования. Магистерская диссертация, МФТИ, 2013.

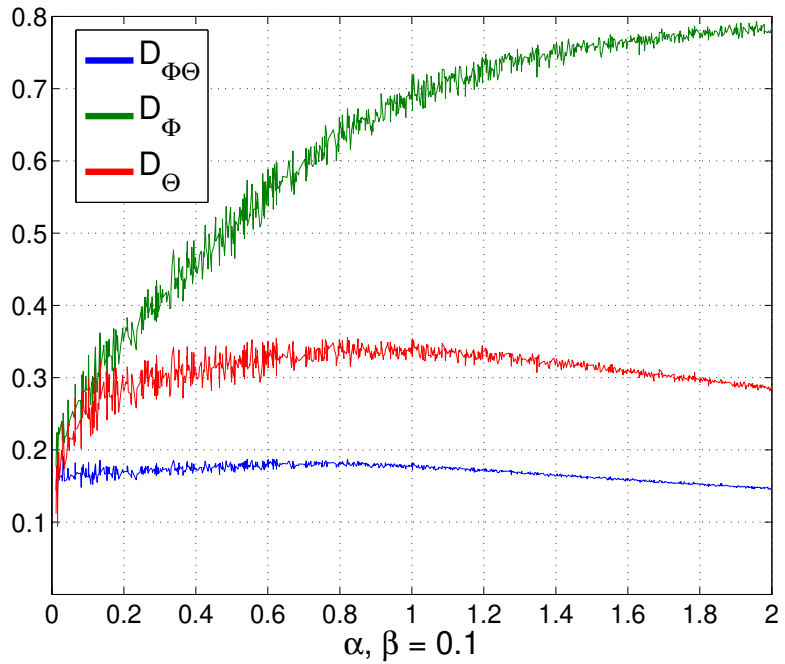
Неустойчивость восстановления матриц Φ и Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от параметра α при фиксированном $\beta = 0.1$

PLSA



LDA



Виталий Глушаченков. Устойчивость матричных разложений в задачах тематического моделирования. Магистерская диссертация, МФТИ, 2013.

Второй эксперимент — на реальных данных

Посты ЖЖ: $|D| = 300\,\text{K}$, $|W| = 154\,\text{K}$, $n = 35\,\text{M}$, $|T| = 120$.
LDA: симметричное распределение Дирихле, $\beta = 0.1$, $\alpha = 0.5$.

Цель эксперимента — оценить различность тем, получаемых в нескольких запусках алгоритма LDA Gibbs Sampling.

Проблема «проклятия размерности»:
длинные хвосты мешают сравнивать распределения.

Доля существенных слов в темах (word ratio):

$$WR = \frac{1}{|W|}\frac{1}{|T|}\sum_{w\in W}\sum_{t\in T}\left[\phi_{wt} > \frac{1}{|W|}\right] \quad \text{(в эксперименте } \sim 3.5\%\text{)}$$

Доля существенных тем в документах (document ratio):

$$DR = \frac{1}{|D|}\frac{1}{|T|}\sum_{d\in D}\sum_{t\in T}\left[\theta_{td} > \frac{1}{|T|}\right] \quad \text{(в эксперименте } \sim 11.5\%\text{)}$$

Koltcov S., Koltsova O., Nikolenko S. Latent Dirichlet Allocation: Stability and applications to studies of user-generated content. ACM WebSci, 2014.

Методика эксперимента

Оставлены слова w, имеющие $\phi_{wt} > \frac{1}{|W|}$ хотя бы в одной теме
Сокращение словаря (vocabulary reduction): 154 K $\to$ 8 K.

Дивергенция Кульбака–Лейблера между темами t и s:

$$\mathrm{KL}(t,s) = \sum_{w \in W} p(w|t) \ln \frac{p(w|t)}{p(w|s)}$$

Нормированная KL-близость пар тем t и s:

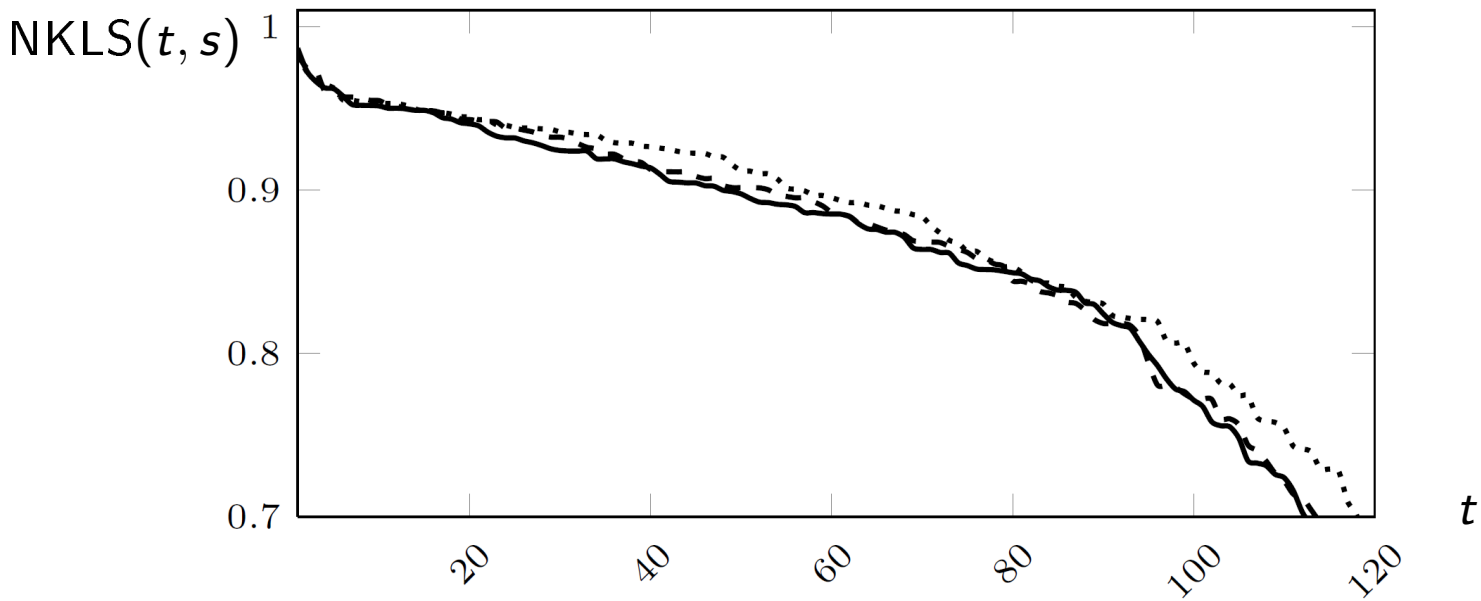
$$\mathrm{NKLS}(t,s) = \left(1 - \frac{\mathrm{KL}(t,s)}{\max\limits_{t',s'} \mathrm{KL}(t',s')}\right)$$

При $\mathrm{NKLS}(t,s) > 0.9$ в темах совпадают 30–50 топовых слов, и эксперты-социологи признают такие темы одинаковыми.

Koltcov S., Koltsova O., Nikolenko S. Latent Dirichlet Allocation: Stability and applications to studies of user-generated content. ACM WebSci, 2014.

Неустойчивость LDA в разных запусках

Результат эксперимента: нормированная KL-близость NKLS между темой t и ближайшей к ней s в другом запуске.



1. Менее 50% тем воспроизводятся от запуска к запуску.
2. Плохо воспроизводятся как мусорные темы, так и хорошие.

Koltcov S., Koltsova O., Nikolenko S. Latent Dirichlet Allocation: Stability and applications to studies of user-generated content. ACM WebSci, 2014.

Выводы из экспериментов

- Матрицы Φ, Θ устойчиво восстанавливаются только при сильной разреженности Φ_0, Θ_0 (более 90% нулей)
- Произведение $\Phi\Theta$ восстанавливается устойчиво, независимо от разреженности исходных Φ_0, Θ_0
- В разных запусках со случайной инициализацией или сэмплированием строятся существенно различные темы
- Распределение Дирихле — слишком слабый регуляризатор
- **Открытые проблемы:** насколько повышает устойчивость
  — декорреляция, сглаживание фона, другие регуляризаторы?
  — тематическая модель битермов?
  — тематическая модель локальных контекстов?

Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models. 2015.

Koltcov S., Koltsova O., Nikolenko S. Latent Dirichlet Allocation: Stability and applications to studies of user-generated content. ACM WebSci, 2014.

Эксперименты с отбором тем на синтетических данных

Коллекция статей NIPS (Neural Information Processing System)

- $|D| = 1566$ обучающих документов; $|D'| = 174$ тестовых
- $|W| = 13\,\text{K}$ — мощность словаря

Синтетическая коллекция:

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

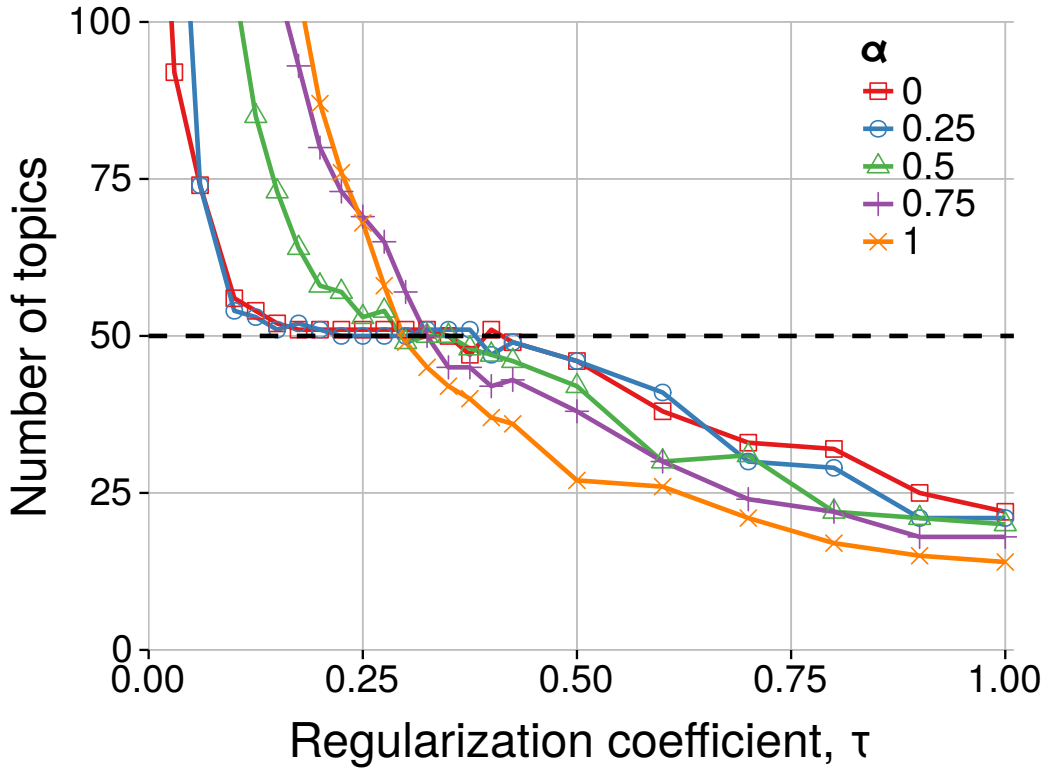
$$n^0_{dw} = n_d \sum_{t \in T_0} \phi_{wt}\theta_{td}$$

Параметрическое семейство полусинтетических данных:

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

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

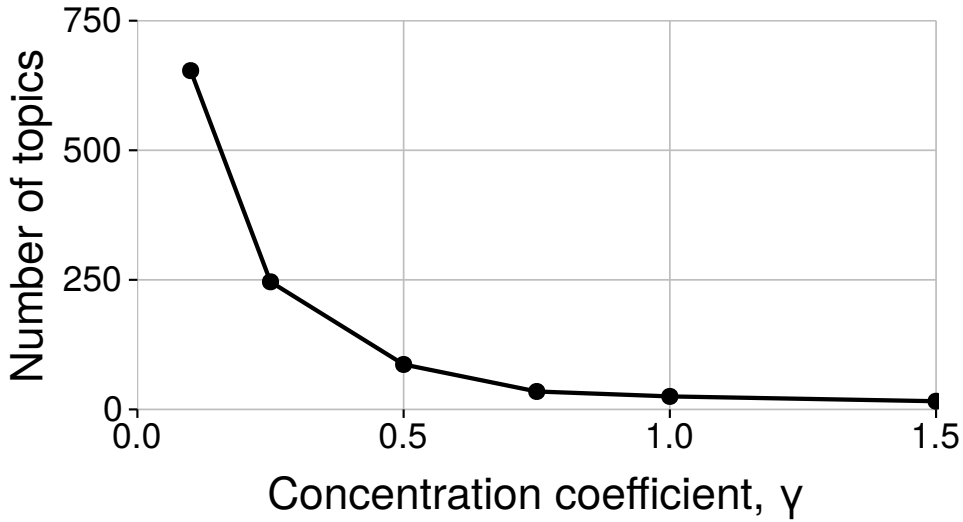
Попытка определения числа тем



- на синтетических данных надёжно находим $|T| = 50$
- причём в широком интервале значений коэффициента τ
- однако на реальных данных чёткого интервала нет

Сравнение с байесовской тематической моделью HDP

HDP, Hierarchical Dirichlet Process [Tech et.al, 2006] — «state-of-the-art» байесовский подход к определению числа тем

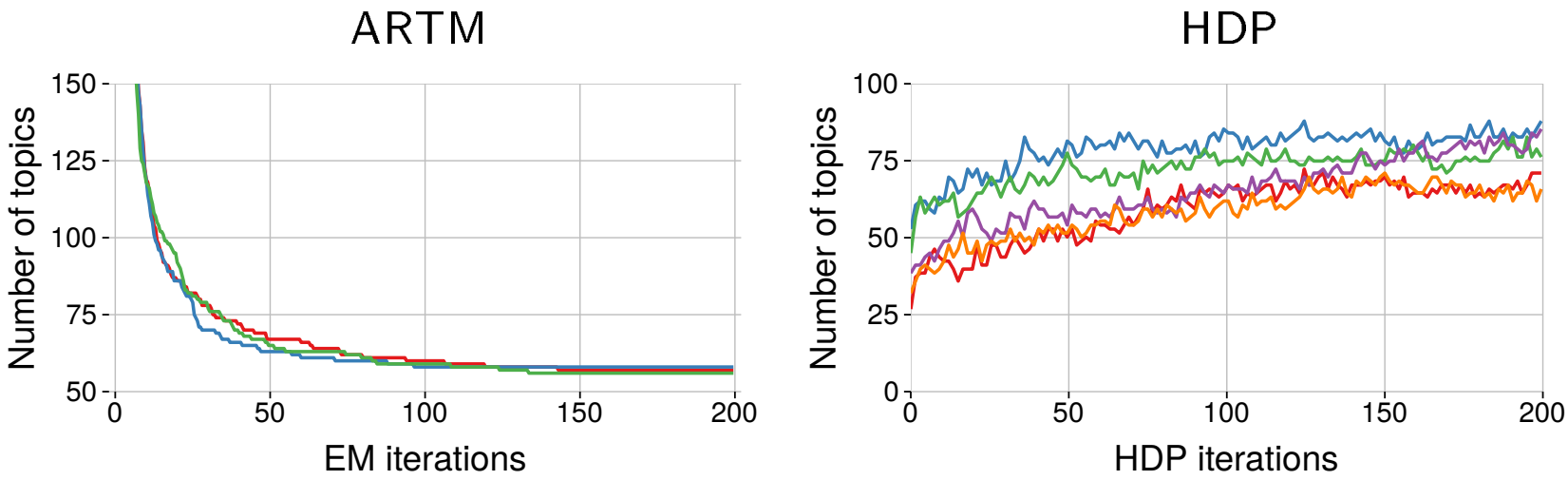


- *Коэффициент концентрации* γ в HDP влияет на $|T|$ столь же сильно, как выбор коэффициента τ в ARTM.

Y.W.Teh, M.Jordan, M.Beal, D.Blei. Hierarchical Dirichlet Processes. 2005.

Сравнение ARTM и HDP по устойчивости

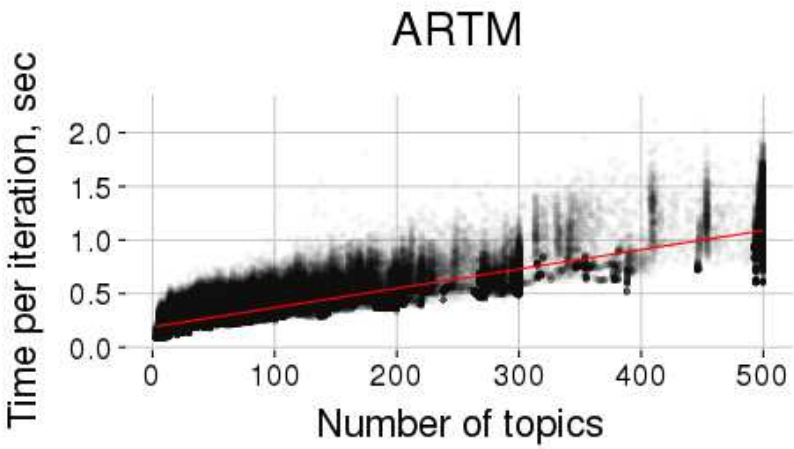
Запуск ARTM и HDP много раз из случайных инициализаций:

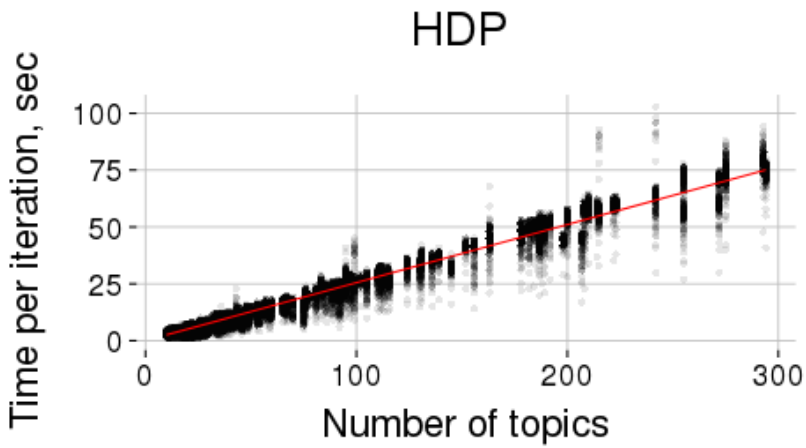


- HDP менее устойчив, причём в двух смыслах:
  - число тем сильнее флуктуирует от итерации к итерации;
  - результаты нескольких запусков различаются сильнее.
- «Рекомендуемые» значения параметров γ в HDP и τ в ARTM дают примерно равное число тем $|T| \approx 60$

Сравнение ARTM и HDP по времени вычислений

Сравнение времени одного прохода коллекции (sec)





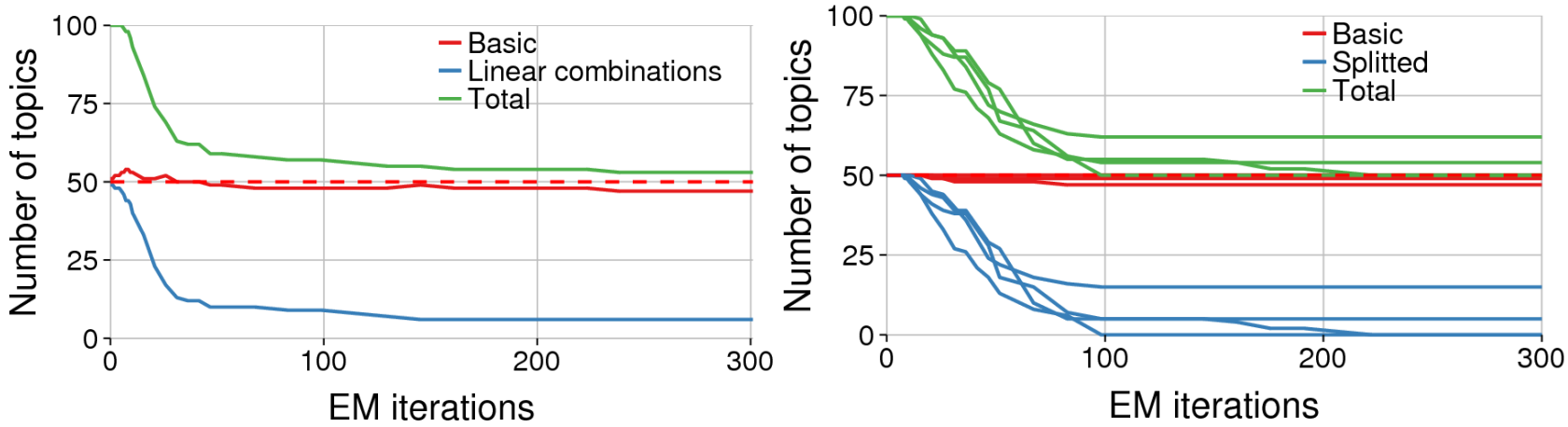
- ARTM в 100 раз быстрее!

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.

Александр Плавин. Отбор тем в задачах тематического моделирования. ВКР бакалавра, МФТИ. 2015.

Удаление линейно зависимых и расщеплённых тем

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



Полезные побочные эффекты регуляризатора отбора тем (на синтетических данных, при оптимальном выборе τ):

- удаляются линейно зависимые и расщеплённые темы
- остаются наиболее различные темы исходной модели.

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.

Выводы из экспериментов

- Регуляризатор отбора тем удаляет незначимые темы и определяет оптимальное число тем, если оно существует
- Увы, в реальных данных его не существует!
  Оптимизировать $|T|$ возможно, но не по модели ΦΘ,
  а по внешним критериям (поиск, классификация и т. п.)
- Выход есть — иерархически дробить темы на подтемы,
  и пусть пользователь выбирает нужную ему детализацию
- Есть простой метод для удаления лишних тем, но
  как обнаруживать новые темы в потоке или в батчах
  и добавлять их в ARTM — пока **открытая проблема**
- Регуляризатор отбора тем имеет полезный побочный эффект, удаляя линейно зависимые и расщеплённые темы
- Почему это происходит — **открытая проблема**

Резюме

- Регуляризация — стандартный приём для решения некорректно поставленных задач
- ARTM позволяет комбинировать регуляризаторы и строить тематические модели с требуемыми свойствами
- Декоррелирование — наиболее полезный регуляризатор
- Сглаживание + разреживание + декоррелирование — часто используемая комбинация регуляризаторов
- Оптимального числа тем, похоже, не существует
- Про другие регуляризаторы — в следующих лекциях
- **Открытые проблемы** модели локальных контекстов:
  — насколько она устойчивее модели «мешка слов»?
  — улучшает ли она интерпретируемость тем?
  — какие регуляризаторы ей нужны?

Задания по курсу

Задача-минимум: научиться решать задачи анализа текстов с использованием тематического моделирования

Задача-максимум: получить новый научный результат

виды деятельности	оценка
теоретическая задача	X
теоретическая задача*	$2X$
теоретическая задача**	$3X$
решение прикладной задачи	$10X$
обзор по последним PTM/NTM	$10X$
участие в проекте	$20X$
работа над открытой проблемой	$25X$

где X — оценка за вид деятельности по 5-балльной шкале.
score — суммарная оценка по всем видам деятельности.

Итоговая оценка: $\min\left(5, \lfloor \text{score}/20 \rfloor\right)$ по 5-балльной шкале.

Задания к лекции 1

Упражнения на принцип максимума правдоподобия:

1. Биграммная модель коллекции: $p(w|v) = \xi_{wv}$,
где v — слово, идущее в тексте перед w.
Найти параметры модели ξ_{wv}.

2. Биграммная модель документов: $p(w|v,d) = \xi_{dwv}$.
Найти параметры модели ξ_{dwv}.

Подсказка: применить условия ККТ или основную лемму.

3*. Творческое задание (возможны разные решения).
Предложите модель, разделяющую роли слов в текстах:
— тематические слова
— специфичные слова документа (шум)
— слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов $p(r|w)$, $r \in \{\text{т}, \text{ш}, \text{ф}\}$.
Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.
Подсказка 3: можно использовать документную частоту слов.

Задания к лекции 2

4. Пользуясь основной леммой, докажите, что регуляризатор битермов эквивалентен добавлению псевдодокументов d_u в исходную коллекцию (см. слайд 13)

Прикладная исследовательская задача:
автоматическое выделение научных терминов (ATE)

- Дано:
  коллекция размеченных текстов конкурса ruTermEval;
  неразмеченная коллекция текстов той же тематики
- Найти:
  метод ATE на основе комбинирования ARTM и TopMine;
  обоснование, что синтаксический анализ не нужен;
  зависимость качества ATE от объёма коллекции
- Критерий:
  качество ATE (Prec, Rec, F1) на размеченных данных

Теоретическое задание к лекции 3

Выведете EM-алгоритм для тематической языковой модели:

5. $p(w|d) = \sum_t \phi_{wt}\theta_{td}$, используя в качестве исходных данных последовательность $(d_i, w_i)_{i=1}^n$ вместо счётчиков n_{dw}.
Докажите эквивалентность обычному EM-алгоритму ARTM.

6. $p(w|d) = \sum_t \phi_{tw}\frac{p(w)}{p(t)}\theta_{td}$, где $p(t)$ фиксировано,
$\phi_{tw} = p(t|w)$, $\theta_{td} = p(t|d)$ — параметры модели.

7. $p(w|d) = \sum_t \phi_{tw}\frac{p(w)}{p(t)}\theta_{td}$, где $p(t)$ фиксировано,
$\phi_{tw} = p(t|w)$ — параметры модели, $\theta_{td} = \sum_w \frac{n_{dw}}{n_d}\phi_{tw}$.

8*. Фиксация $p(t)$ как внешнего параметра упрощает выкладки, но может нарушать условия целостности модели:

$$p(t) = \sum_t \phi_{tw}p(w), \qquad p(t) = \sum_t \theta_{td}p(d).$$

Как обеспечить выполнение этих условий в EM-алгоритме?

Теоретическое задание к лекции 4

9. Докажите, что необходимым условием максимума

$$\sum_{i=1}^{n} \ln \sum_{t \in T} p(w_i, t|i, \Omega) \;\to\; \max_{\Omega}$$

для языковой модели со скрытыми переменными $t \in T$ (не обязательно темами) и параметрами $\Omega = (\omega_{kj})$ — набором неотрицательных нормированных векторов, является система

$$\text{E-шаг:} \quad p(t|w_i, i) = \operatorname*{norm}_{t \in T} p(w_i, t|i, \Omega)$$

$$\text{M-шаг:} \quad \omega_{kj} = \operatorname*{norm}_{k} \left(\sum_{i=1}^{n} \sum_{t \in T} p(t|w_i, i)\, \omega_{kj} \frac{\partial}{\partial \omega_{kj}} \ln p(w_i, t|i, \Omega) \right)$$

10. Выведете отсюда EM-алгоритм для частных случаев:

1) $p(w, t|i, \Omega) = \phi_{wt}\theta_{td_i}$;
2) $p(w, t|i, \Omega) = \phi_{tw} \frac{p(w)}{p(t)} \sum_{w \in d_i} \frac{n_{d_i w}}{n_{d_i}} \phi_{tw}$;
3) $p(w, t|i, \Omega) = \phi_{tw} \frac{p(w)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c}$.

11. Творческое задание.** Предложите способ ввести обучаемые параметры в тематическую модель внимания.

Исследовательская задача к лекции 4

Реализуйте EM-алгоритм для модели локального контекста (или воспользуйтесь готовой реализацией)

Исследуйте зависимость метрик качества модели

- перплексия: $\mathscr{P} = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} p(w|C_i)\right)$
- разреженность, различность, когерентность тем
- дефекты целостности модели:
  $\left\|p(t) - \frac{n_t}{n}\right\|$, $\left\|p(t) - \sum_t \phi_{tw} p(w)\right\|$, $\left\|p(t) - \sum_t \theta_{td} p(d)\right\|$

от номера итерации и от параметров модели:

- $|T|$ — число тем
- L — число проходов
- τ — вес N_{tw} в формуле M-шага, особый случай $\tau = 0$
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$, β — баланс левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, секций
- опция «$i \in C_i$ или $i \notin C_i$»

Теоретическое задание к лекции 5

12. Найдите дискретное распределение $P = (p_i)_{i=1}^n$ в задаче $\sum_i n_i \mu(p_i) \to \max$ с гладкой монотонно возрастающей $\mu(p)$.
Отдельно рассмотрите случаи $\mu(p) = p^s$, $s = 1$, $s \to 0$.

13. Выведете EM-алгоритм в случае, когда ln заменён гладкой монотонно возрастающей функцией μ:

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

Подумайте, какие замены логарифма полезны, и почему.

14. Простейшая идея разреживания — обнуление малых вероятностей. Чтобы обосновать эту эвристику, найдите, какому регуляризатору соответствует формула M-шага

$$\phi_{wt} = \operatorname*{norm}_w \left(n_{wt} [n_{wt} > \gamma n_t] \right)$$

Подсказка: с учётом подстановки несмещённой оценки ϕ^*_{wt}

Теоретическое задание к лекции 5

Проект «Тематизатор». Аналитик построил модель $\Phi^0\Theta^0$ и отметил среди столбцов матрицы Φ^0 темы двух типов: удачные $T_+ \subset T$ и неудачные $T_- \subset T$.

Теперь он хочет построить модель ещё раз так, чтобы

- удачные темы остались в матрице Φ;
- остальные темы построились по-другому и были не похожи на каждую из неудачных тем $t \in T_-$.

15. Предложите регуляризаторы для этого.

16. Не получится ли так, что новые темы будут отдаляться от суммы неудачных тем $\sum_{t \in T_-} \phi^0_{wt}$ вместо того, чтобы отдаляться от каждой из неудачных тем по отдельности? Почему это плохо и как этого избежать?

17. Предложите способ инициализации Φ для новой модели.

Исследовательская задача к лекции 5

Продолжение исследования по автоматическому выделению научных терминов (Automatic Term Extraction, ATE)

- Дано:
  коллекция размеченных текстов конкурса ruTermEval;
  неразмеченная коллекция текстов той же тематики
- Найти:
  — оптимальную стратегию регуляризации на основе декоррелирования и сглаживания фоновых тем
  — рекомендации по управлению относительными коэффициентами регуляризации
  — критерий тематичности терминов по расстоянию между распределениями $p(t|w)$ и $p_0(t) = \frac{1}{|T|}$, позволяющий наиболее чётко отличать термины от фоновой лексики
- Критерий:
  — максимум доли терминов в предметных темах
  — минимум доли терминов в фоновых темах

Исследовательская задача к лекции 5

Продолжение исследования модели локального контекста (можно воспользоваться готовой реализацией EM-алгоритма)

Исследуйте устойчивость модели в сравнении с ARTM

- без регуляризации
- с регуляризатором декоррелирования, при различных значениях относительного коэффициента регуляризации

Как на устойчивость модели влияют её параметры:

- $|T|$ — число тем
- L — число проходов
- τ — вес N_{tw} в формуле M-шага, особый случай $\tau = 0$
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$, β — баланс левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, секций
- опция «$i \in C_i$ или $i \notin C_i$»

Примеры датасетов для практических заданий по курсу

1. Открытые датасеты (английский): 20NG, NIPS, KOS
2. Ранжированные результаты поиска научных статей (по данным eLibrary, arXiv, PubMed)
3. Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
4. Техноблоги: Хабр (русский), TechCrunch (английский)
5. Данные социальных сетей: VK, Twitter, Telegram,...
6. Статьи по Complexity Sciences (для хронокарты науки)

- Википедия
- Новостной поток (20 источников на русском языке)
- Данные кадровых агентств: резюме + вакансии
- Транзакции клиентов Sberbank DSD 2016
- Акты арбитражных судов РФ

---

http://bigartm.org
http://drive.google.com/drive/folders/1PPnw6aZOJAJoLRYuwdGm437RssV-XQx0

Проекты

- «Тематизатор» для социо-гуманитарных исследований:
  — пользователь задаёт грубый фильтр текстового потока;
  — задача: «классифицировать иголки в стоге сена»,
  — разделив темы на информативные и мусорные,
  — выделив аспекты и тональности в каждой теме;
  — конечная цель: кол./кач. анализ предметной области,
  — реализация данного сценария как модуля в среде Orange

- «Мастерская знаний» для научного поиска:
  — пользователь строит тематические подборки статей,
  — поисковая выдача формируется моделью SciRus;
  — задача: показать пользователю тематику подборки;
  — понадобится: автоматическое выделение терминов,
  — выделение тематических фраз из документов,
  — автоматическое именование и суммаризация тем;
  — конечная цель: помочь в понимании предметной области

Открытые проблемы тематического моделирования

1. Тематические модели внимания последовательного текста
2. Проблема несбалансированности тем в коллекции
3. Измерение интерпретируемости тем (когерентность)
4. Обеспечение 100%-й интерпретируемости тем
5. Автоматическое именование и суммаризация тем
6. Калибровка моделей тематической фильтрации
7. Согласование тем с предобученными эбмедингами LLM
8. Статистические оценки состоятельности тем
9. Обнаружение новых тем или трендов в потоке текстов
10. Обеспечение устойчивости и полноты множества тем
11. Автоматический подбор гиперпараметров, AutoML
12. Гиперграфовые тематические модели для RecSys