Многокритериальные и многомодальные вероятностные тематические модели коллекций текстовых документов

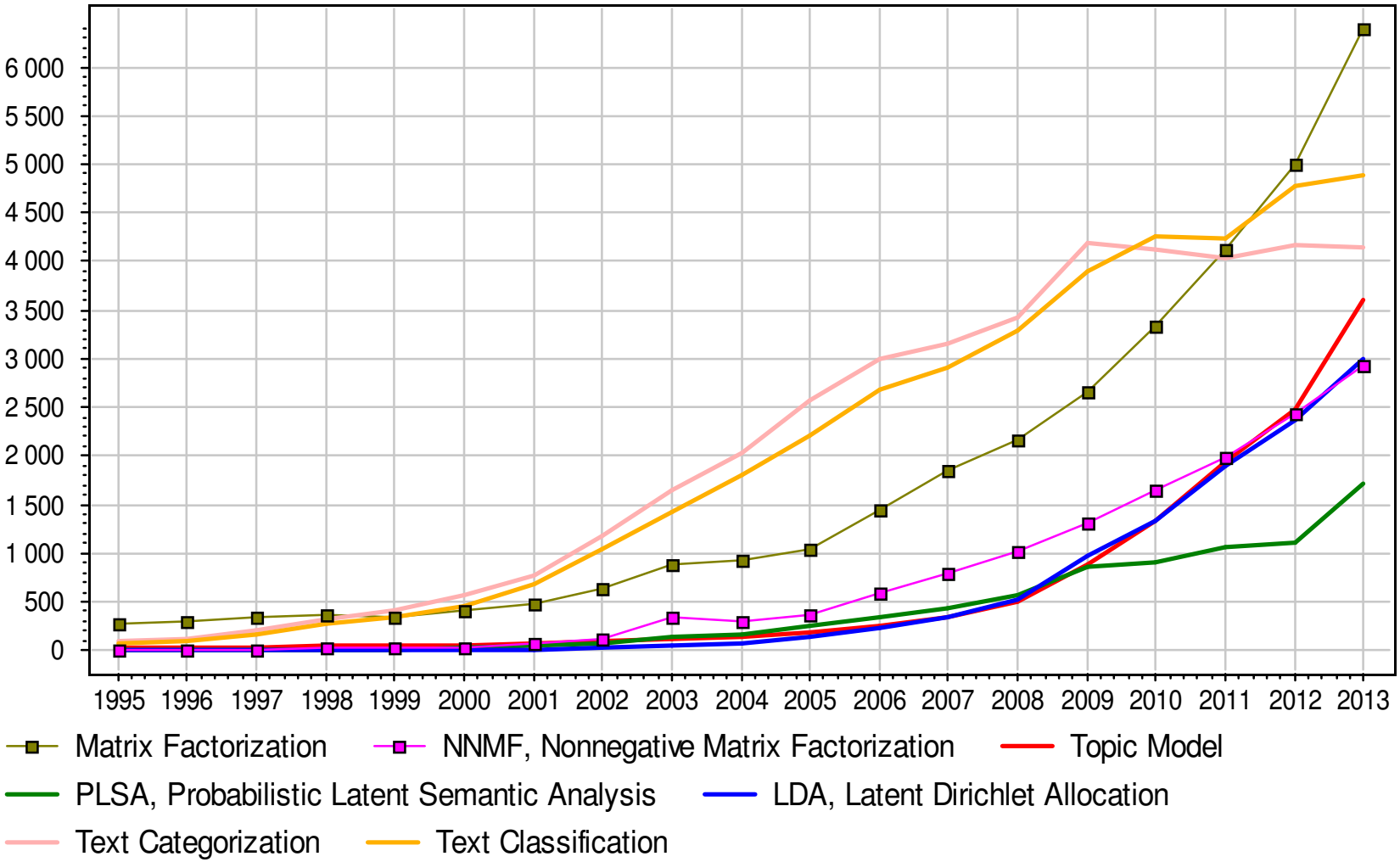
Воронцов Константин Вячеславович
ВЦ РАН • МФТИ • FORECSYS

Интеллектуализация Обработки Информации, ИОИ-10
6–10 октября 2014, Греция, о. Крит

Содержание

Тематическое моделирование и близкие области исследований

Динамика цитирования, по данным Google Scholar:



Понятие «латентной темы»

Основная цель тематического моделирования — автоматическое выявление тематики текстовых документов, поиск информации по смыслу, а не по ключевым словам

Документ d — это последовательность терминов $w_1, \ldots, w_{n_d}$,
$p(w|d) = \frac{n_{dw}}{n_d}$ — известная частота термина w в документе d

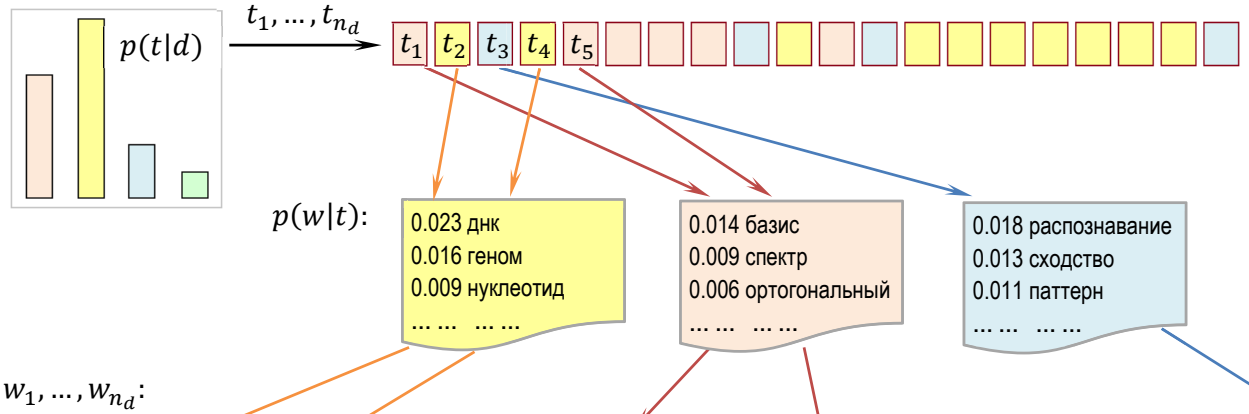
Коллекция документов — i.i.d. выборка (d_i, w_i, t_i), $i = 1, \ldots, n$ из дискретного вероятностного пространства $D \times W \times T$

- *Тема* — это специальные термины предметной области, совместно часто встречающиеся в документах,
  $p(w|t)$ — латентное распределение терминов w в теме t
- *Тематика документа* — это доли тем в документе:
  $p(t|d)$ — латентное распределение тем t в документе d

Прямая задача — порождение коллекции по $p(w|t)$ и $p(t|d)$

Вероятностная тематическая модель коллекции документов D:

$$p(w|d) = \sum_t p(w|t)\, p(t|d), \quad d \in D$$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Обратная задача — восстановление $p(w|t)$ и $p(t|d)$ по коллекции

Дана коллекция текстовых документов (мешков слов):
n_{dw} — сколько раз термин w встречается в документе d

Найти модель $p(w|d) = \sum_t \phi_{wt}\theta_{td}$ с параметрами ϕ, θ:
$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Задача максимизации логарифма правдоподобия:

$$\mathscr{L}(\Phi,\Theta) = \sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta},$$

при ограничениях нормировки и неотрицательности

$$\phi_{wt} \geqslant 0; \quad \sum_w \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_t \theta_{td} = 1$$

PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Теорема

Точка максимума $\mathscr{L}(\Phi, \Theta)$ удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} \equiv p(t|d,w)$, n_{wt}, n_{td}:

$$\text{E-шаг:} \quad p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{t'} \phi_{wt'}\theta_{t'd}}$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \dfrac{n_{wt}}{n_t}; \quad n_{wt} = \sum\limits_{d \in D} n_{dw} p_{tdw}; \quad n_t = \sum\limits_{w \in W} n_{wt} \\ \theta_{td} = \dfrac{n_{td}}{n_d}; \quad n_{td} = \sum\limits_{w \in d} n_{dw} p_{tdw}; \quad n_d = \sum\limits_{t \in T} n_{td} \end{cases}$$

EM-алгоритм — чередование E- и M-шага до сходимости, т. е. решение системы уравнений методом простых итераций.

✓ *Идея на будущее: можно использовать и другие методы!*

LDA — Latent Dirichlet Allocation [Blei 2003]

Оценки условных вероятностей $\phi_{wt} \equiv p(w|t)$, $\theta_{td} \equiv p(t|d)$:

- в PLSA — несмещённые оценки максимума правдоподобия:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad \theta_{td} = \frac{n_{td}}{n_d}$$

- в LDA — сглаженные байесовские оценки:

$$\phi_{wt} = \frac{n_{wt} + \beta_w}{n_t + \beta_0}, \qquad \theta_{td} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}$$

Различие проявляется только при малых n_{wt}, n_{td}

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.

Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models. Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.

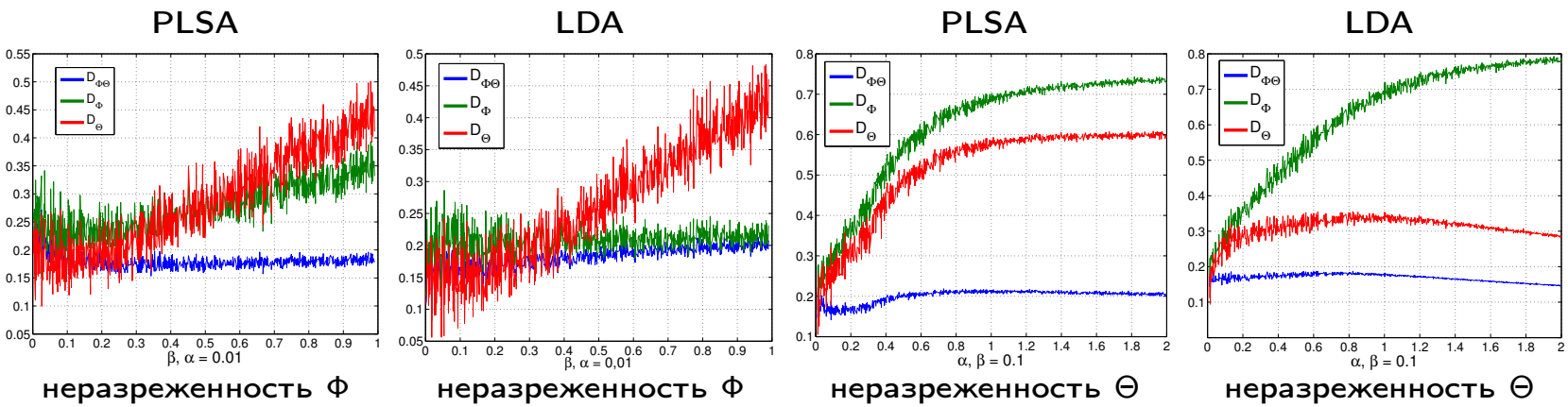
Задача тематического моделирования некорректно поставлена

Неединственность стохастического матричного разложения:

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

для невырожденных $S_{T\times T}$ таких, что Φ', Θ' — стохастические.

Эксперимент. Произведение $\Phi\Theta$ восстанавливается устойчиво, матрица Φ и матрица Θ — только когда сильно разрежены:



неразреженность Φ | неразреженность Φ | неразреженность Θ | неразреженность Θ

Вывод 1: нужны дополнительные требования к модели.

Вывод 2: требований сглаживания в LDA не достаточно.

ARTM — Аддитивная регуляризация тематической модели

Пусть, наряду с правдоподобием, требуется максимизировать ещё n критериев — *регуляризаторов* $R_i(\Phi, \Theta)$, $i = 1, \ldots, n$.

Метод многокритериальной оптимизации — *скаляризация*.

Задача максимизации регуляризованного правдоподобия:

$$\underbrace{\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}}_{\mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w \in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t \in T} \theta_{td} = 1$$

где $\tau_i > 0$ — *коэффициенты регуляризации*.

EM-алгоритм с регуляризацией M-шага

Теорема

Точка максимума $\mathscr{L}(\Phi,\Theta)+R(\Phi,\Theta)$ удовлетворяет системе уравнений со вспомогательными переменными p_{tdw}, n_{wt}, n_{td}:

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_{t'}\phi_{wt'}\theta_{t'd}}; \\ \text{M-шаг:} & \phi_{wt} \propto \left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\right)_+; \quad n_{wt} = \sum\limits_{d\in D} n_{dw}p_{tdw}; \\ & \theta_{td} \propto \left(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\right)_+; \quad n_{td} = \sum\limits_{w\in d} n_{dw}p_{tdw}; \end{cases}$$

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

PLSA: $R(\Phi,\Theta) = 0$

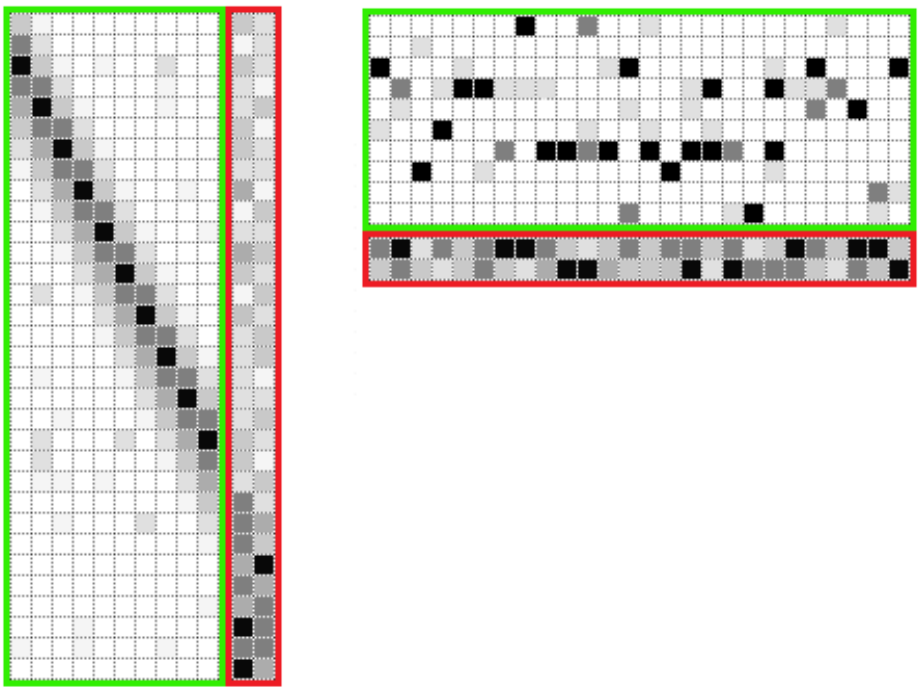
LDA: $R(\Phi,\Theta) = \sum\limits_{t,w}\beta_w \ln\phi_{wt} + \sum\limits_{d,t}\alpha_t \ln\theta_{td}$

Требования интерпретируемости и гипотеза о структуре тем

Предметные темы S содержат термины предметной области, распределения $p(w|t)$ разреженные, существенно различные

Фоновые темы B содержат слова общей лексики, распределения $p(w|t)$ и $p(t|d)$ не разреженные

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



Справочные сведения. Дивергенция Кульбака–Лейблера

Функция расстояния между распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$:

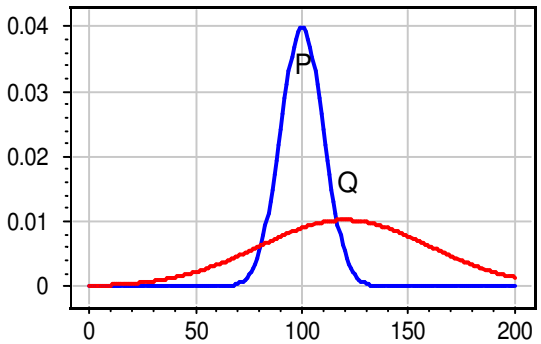
$$\mathrm{KL}(P\|Q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}.$$

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

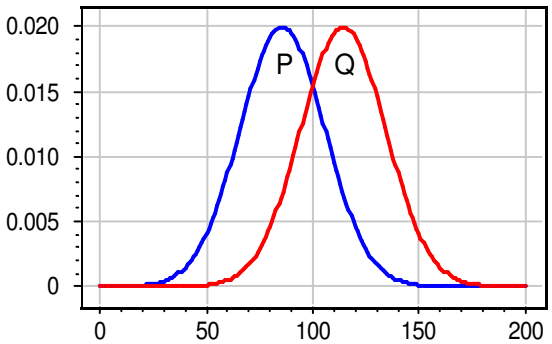
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

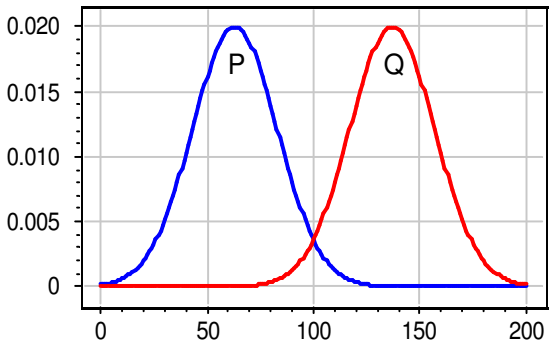
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



$\mathrm{KL}(P\|Q) = 0.442$
$\mathrm{KL}(Q\|P) = 2.966$



$\mathrm{KL}(P\|Q) = 0.444$
$\mathrm{KL}(Q\|P) = 0.444$



$\mathrm{KL}(P\|Q) = 2.969$
$\mathrm{KL}(Q\|P) = 2.969$

Сглаживание + разреживание + декорреляция + отбор тем

1. разреживание предметных тем $S \subset T$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in S} \mathrm{KL}(\beta_w \| \phi_{wt}) + \alpha_0 \sum_{d \in D} \mathrm{KL}(\alpha_t \| \theta_{td}) \to \max$$

2. сглаживание фоновых тем $B \subset T$, аналог LDA:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in B} \mathrm{KL}(\beta_w \| \phi_{wt}) - \alpha_0 \sum_{d \in D} \mathrm{KL}(\alpha_t \| \theta_{td}) \to \max$$

3. декоррелирование тем как столбцов Φ:

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

4. отбор тем путём разреживания $p(t)$:

$$R(\Theta) = \tau \mathrm{KL}\Big(\tfrac{1}{|T|} \| p(t)\Big) \to \max, \quad p(t) = \sum_{d \in D} \theta_{td} p(d)$$

Эксперимент по комбинированию регуляризаторов

Задача: улучшить интерпретируемость, не ухудшив перплексию

Набор регуляризаторов:

1. сглаживание фоновых тем — столбцов Φ, строк Θ
2. разреживание предметных тем — столбцов Φ, строк Θ
3. декоррелирование предметных тем — столбцов Φ
4. удаление незначимых тем — строк Θ

Данные: NIPS (Neural Information Processing System)

- $|D| = 1566$ статей конференции NIPS на английском языке;
- суммарной длины $n \approx 2.3 \cdot 10^6$,
- словарь $|W| \approx 1.3 \cdot 10^4$.
- контрольная коллекция: $|D'| = 174$.

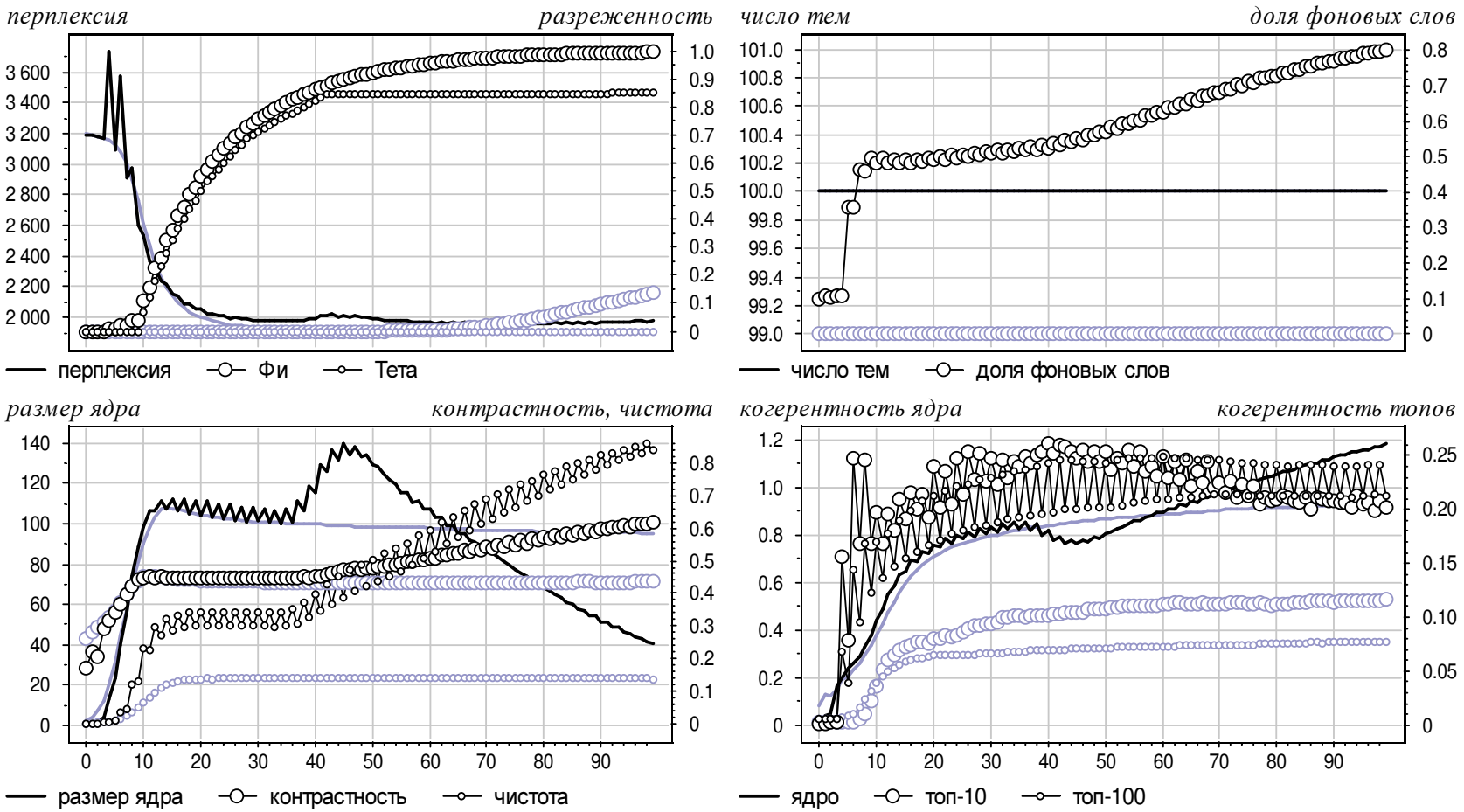
Критерии качества модели

Построение BTM — многокритериальная оптимизация.
Поэтому критериев для контроля качества модели тоже много.

- Перплексия контрольной коллекции: $\mathcal{P} = \exp\left(-\frac{1}{n'}\mathscr{L}(D')\right)$
- Разреженность — доля нулевых элементов в Φ и Θ
- Характеристики интерпретируемости тем:
  - когерентность темы [Newman, 2010]
  - размер ядра темы: $|W_t|$, ядро $W_t = \{w\colon p(t|w) > 0.25\}$
  - чистота темы: $\sum\limits_{w \in W_t} p(w|t)$
  - контрастность темы: $\frac{1}{|W_t|}\sum\limits_{w \in W_t} p(t|w)$
- Вырожденность тематической модели:
  - число тем $|T|$
  - доля фоновых слов: $\frac{1}{n}\sum\limits_{d \in D}\sum\limits_{w \in d}\sum\limits_{t \in B} p(t|d,w)$

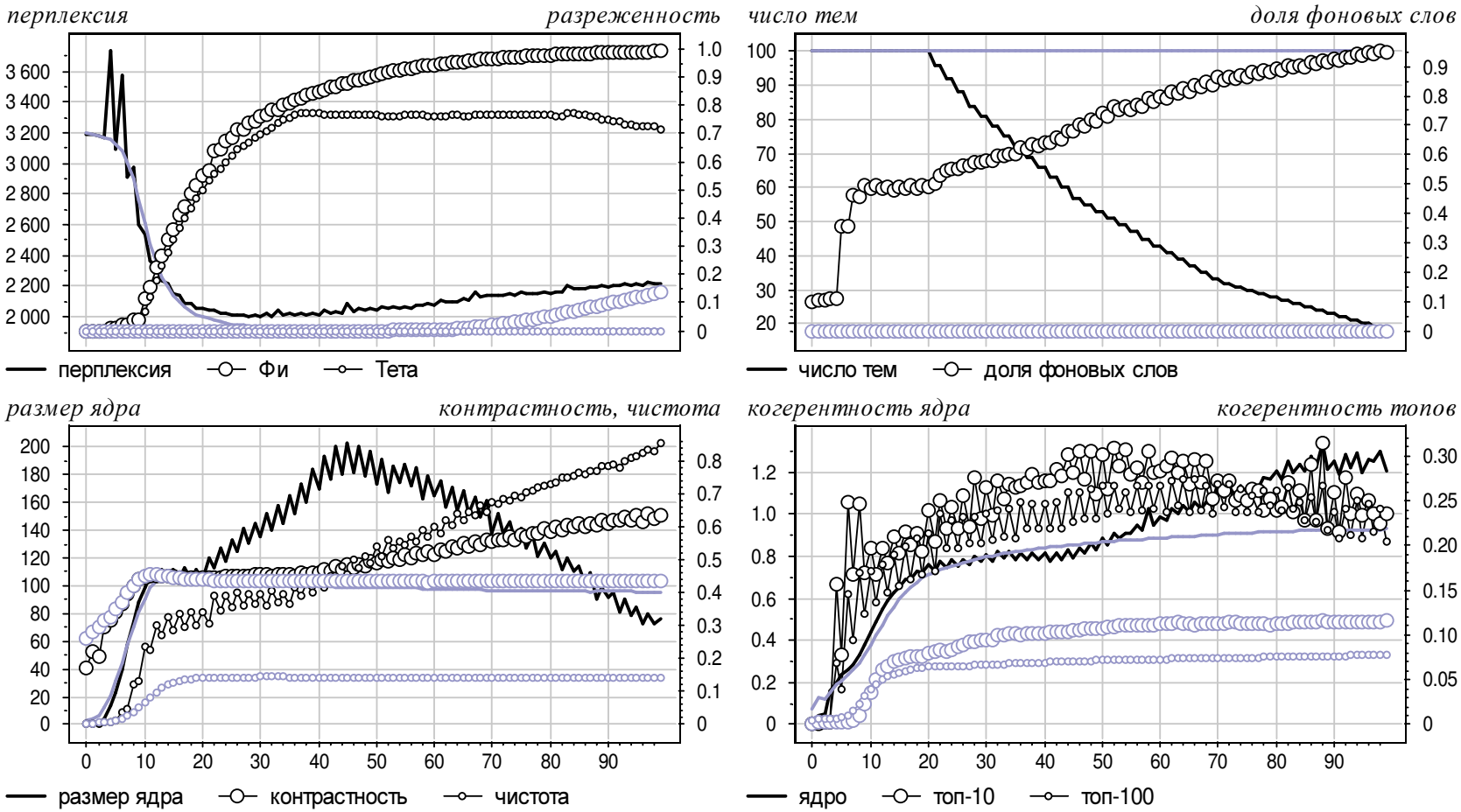
Мониторинг качества модели в EM-алгоритме

Сравнение PLSA (серый) и ARTM со сглаживанием, разреживанием и декоррелированием (чёрный)



Мониторинг качества модели в EM-алгоритме

Сравнение PLSA (серый) и ARTM со сглаживанием, разреживанием, декоррелированием и отбором тем (чёрный)



Выводы

Одновременное улучшение многих показателей:

- *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
- *размер ядер тем* вырос от 0 до 150 терминов
- почти без потери *перплексии* (правдоподобия) модели

Рекомендации по подбору траектории регуляризации:

- сглаживание включать сразу
- декорреляцию включать сразу и как можно сильнее
- разреживание включать постепенно после 10-20 итераций
- только после этого включать отбор тем
- декорреляцию и отбор тем делать на разных итерациях

Задача анализа потока пресс-релизов

Дано: коллекция официальных пресс-релизов внешнеполитических ведомств ряда стран, на английском языке. Более 20 тыс. сообщений за 10 лет, 180Мб текста.

Найти:

- какие темы перманентные?
- какие темы привязаны к событиям?
- какие темы и в какие моменты коррелируют?

Регуляризаторы:

- разреживание, сглаживание, декоррелирование
- разреживание тем $p(t|y)$ в каждый момент времени y
- сглаживание тем $p(y|t)$ в соседние моменты времени

Регуляризаторы для динамической тематической модели

Y — моменты времени (например, годы публикаций),
$y(d)$ — метка времени документа d,
$D_y \subset D$ — все документы, относящиеся к моменту $y \in Y$.

Гипотеза 1: распределение $p(t|y) = \sum_{d \in D_y} \theta_{td} p(d)$ разрежено:

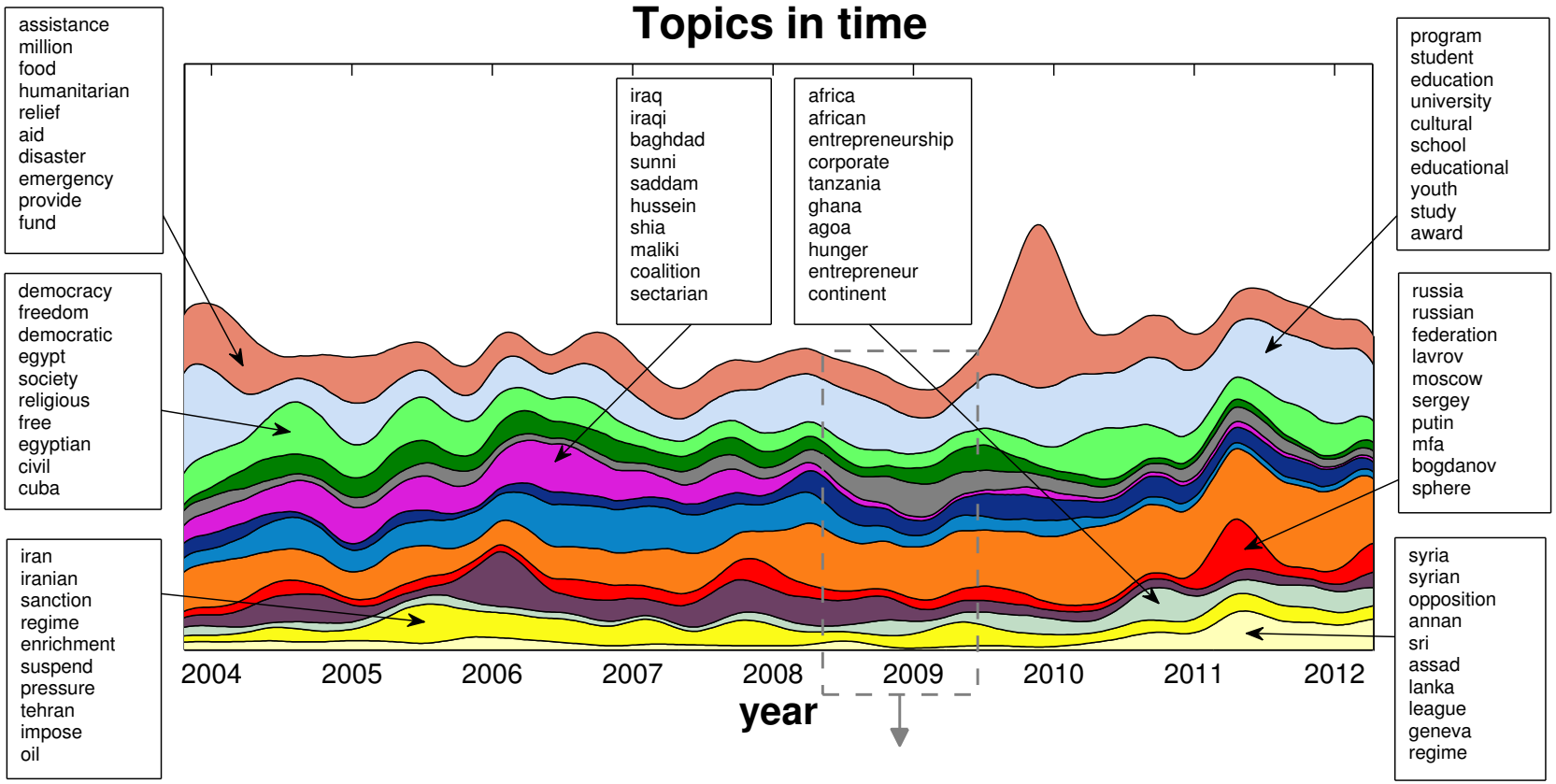
$$R_1(\Theta) = -\tau_1 \sum_{y \in Y} \mathrm{KL}\left(\tfrac{1}{|T|} \| p(t|y)\right) \to \max.$$

Гипотеза 2: $p(y|t)$ меняются плавно, с редкими скачками:

$$R_2(\Theta) = -\tau_2 \sum_{y \in Y} \sum_{t \in T} \left|p(y|t) - p(y-1|t)\right| \to \max.$$

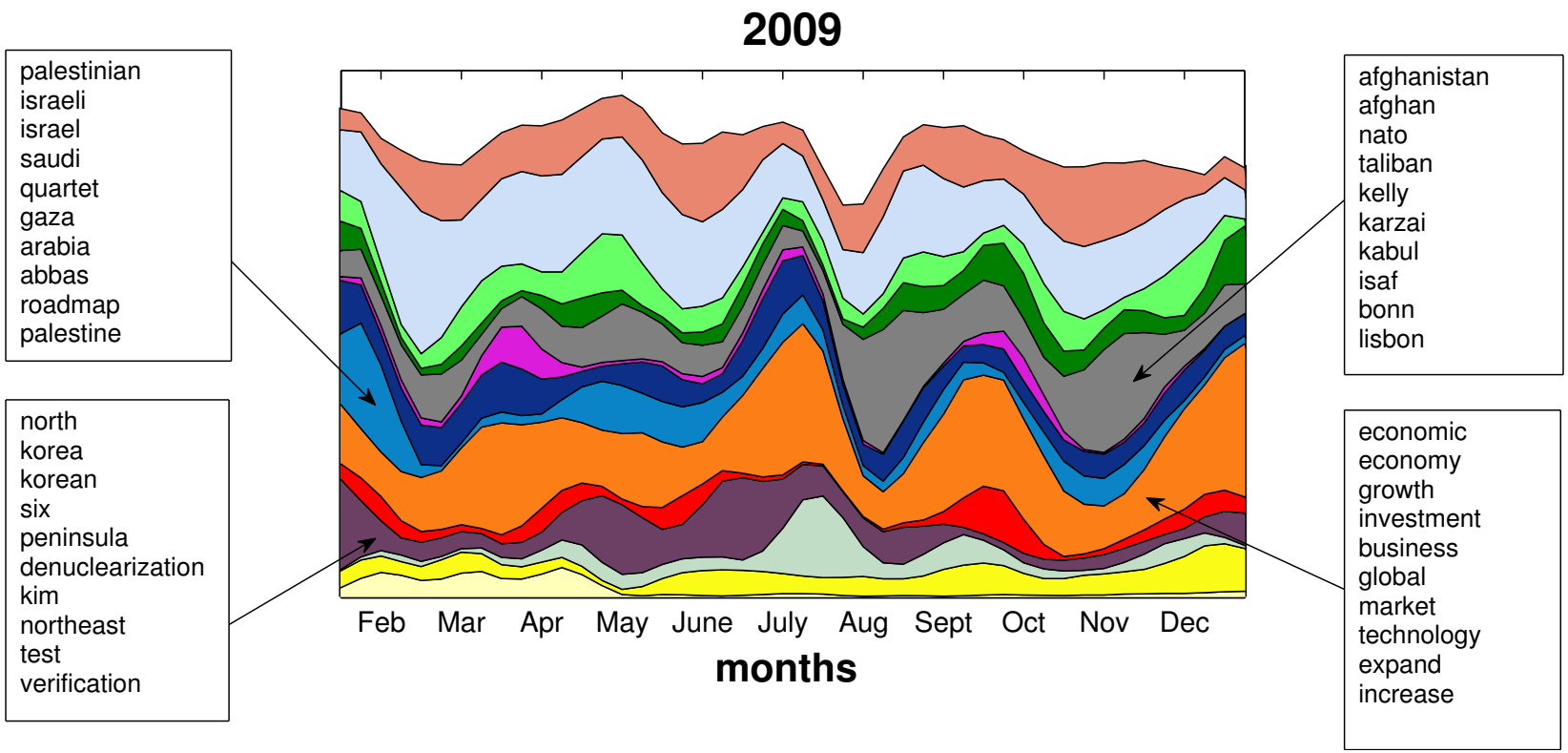
Эксперименты с динамической тематической моделью

Примеры хорошо интерпретируемых тем



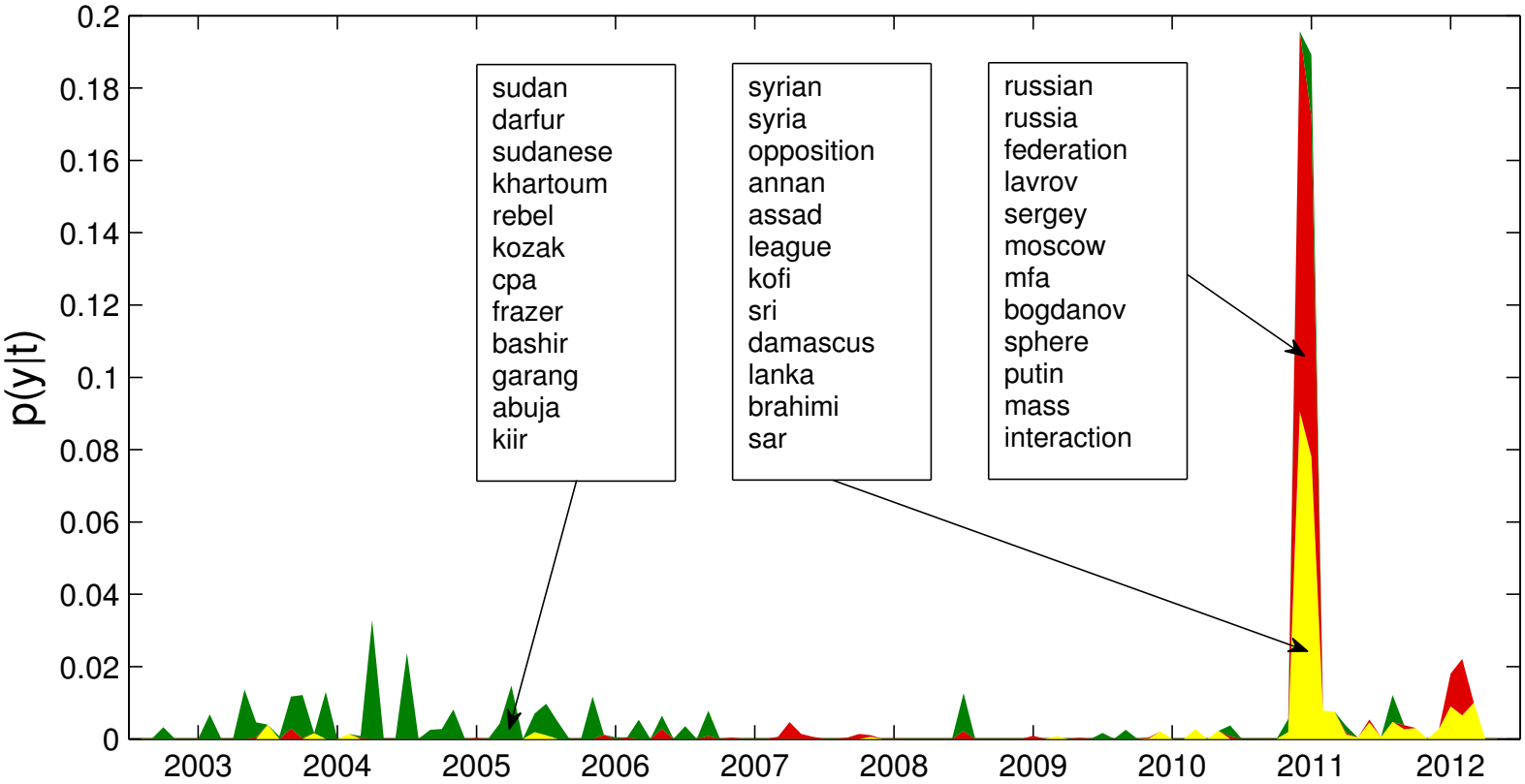
Эксперименты с динамической тематической моделью

Укрупнение масштаба времени



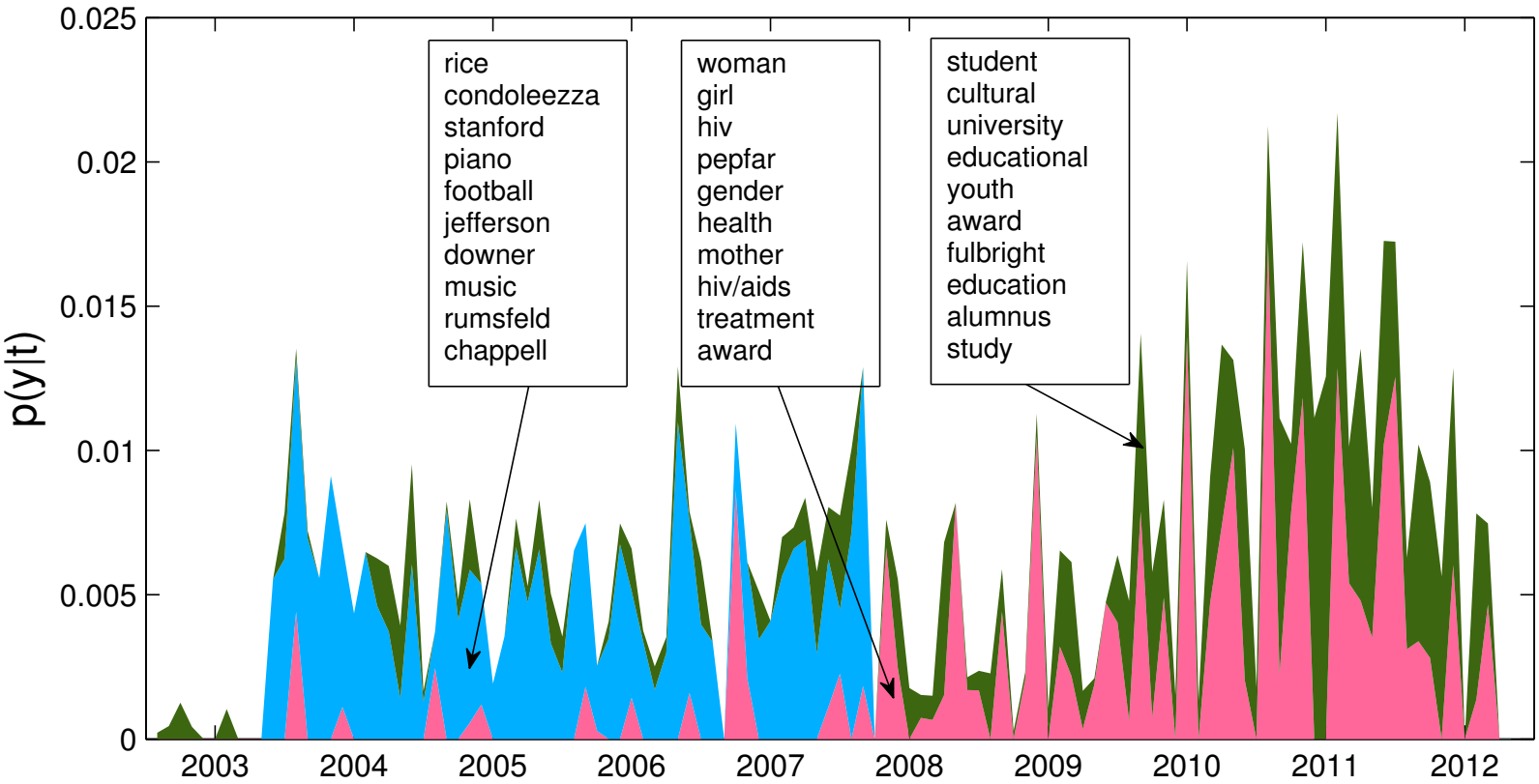
Эксперименты с динамической тематической моделью

Примеры событийных тем и момента их совместного всплеска



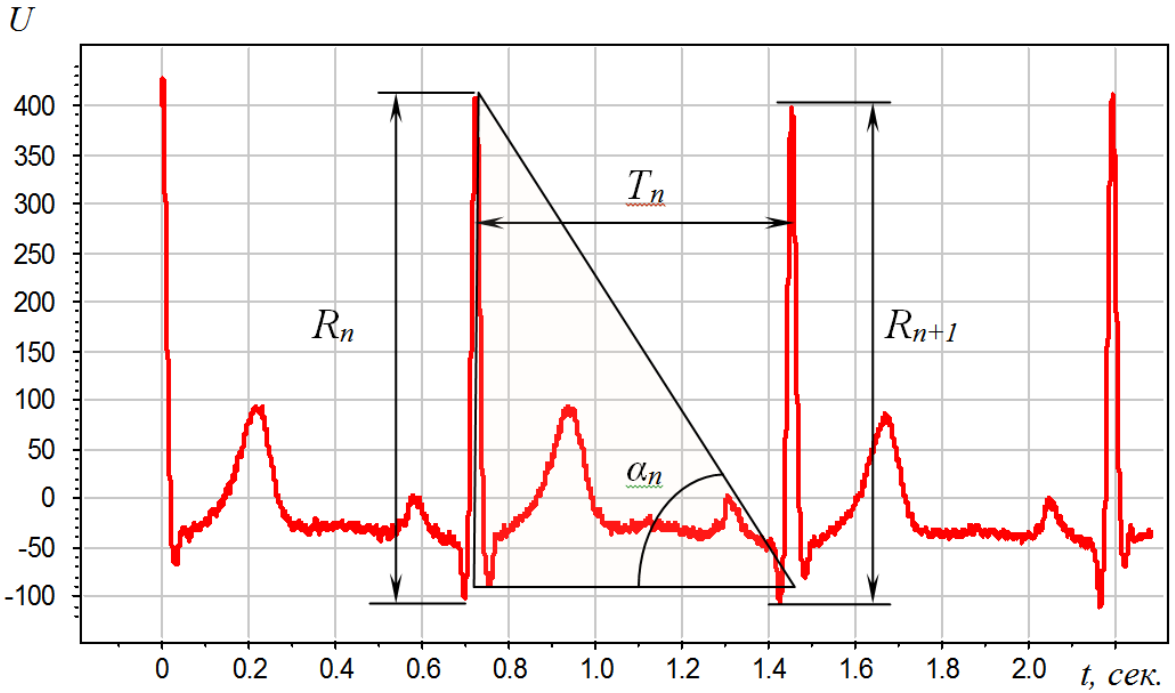
Эксперименты с динамической тематической моделью

Примеры перманентных тем



Информационный анализ электрокардиосигналов

Теория информационной функции сердца В.М.Успенского: амплитуды R_n и интервалы T_n кардиоциклов несут информацию о многих заболеваниях внутренних органов.



$$\alpha_n = \text{arctg}\,\frac{R_n}{T_n}$$

Диагностическая система «Скринфакс» (2-е поколение)

- более 10 лет эксплуатации (начало исследований: 1978)
- более 20 тысяч прецедентов (кардиограмма + диагноз)
- более 30 заболеваний внутренних органов (+ туберкулёз)

Дискретизация ЭКГ-сигнала

Вход: последовательность интервалов и амплитуд $(T_n, R_n)_{n=1}^N$

Правила кодирования:

если	$R_n < R_{n+1}$,	$T_n < T_{n+1}$,	$\alpha_n < \alpha_{n+1}$	то	$s_n = \mathtt{A}$
если	$R_n \geqslant R_{n+1}$,	$T_n \geqslant T_{n+1}$,	$\alpha_n < \alpha_{n+1}$	то	$s_n = \mathtt{B}$
если	$R_n < R_{n+1}$,	$T_n \geqslant T_{n+1}$,	$\alpha_n < \alpha_{n+1}$	то	$s_n = \mathtt{C}$
если	$R_n \geqslant R_{n+1}$,	$T_n < T_{n+1}$,	$\alpha_n \geqslant \alpha_{n+1}$	то	$s_n = \mathtt{D}$
если	$R_n < R_{n+1}$,	$T_n < T_{n+1}$,	$\alpha_n \geqslant \alpha_{n+1}$	то	$s_n = \mathtt{E}$
если	$R_n \geqslant R_{n+1}$,	$T_n \geqslant T_{n+1}$,	$\alpha_n \geqslant \alpha_{n+1}$	то	$s_n = \mathtt{F}$

Выход: *кодограмма* $d = (s_n)_{n=1}^{N-1}$ — последовательность символов алфавита $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$:

```
DBFEACFDAAFBABDDAADFAAFFEACFEACFBAEFFAABFFAAFFAAFFAAFFAAEBFAEBFEAAFCAAFFAAD
FCAFFAADFCADFCCDFDACFFACDFAEFFACFFEADFCAFBCADFFECFFAAFFAAFFAEFFCACFCAEFFCAD
DAADBFAAFFAEBFAABFACDFFAAFBAADFAADFDAAFCECFCEDFCEEFCAEFBECBBBAADBAACFFAAFFA
CFFCECFDAABDAEFFAAFFCEDBFAAFFAEFFAEFBACFBAEDFEAAFFCAFFDAAFFAEBDAADBBADFDAFF
EABFCCAFDEEBDECFFACFFAABFAADFBAAFFACFFFAEFFACFFACFFCECFBAAFFFAAFFFAAFFAADFB
AABFACDFDAEFFAADBAAEFFEAFBCECFDECCFBAAFFAADFDACDFAAFFAADFCAADFAEFBAAFFCADFE
AFFCECFCECFFAAFFABCFDAAAFFADBFCAEFFAABFACBFAAEBFAEBFCAFFBAAFFAAFFDACFDAABFB
CAFFAECFFACFFACDFCADFDAABFAAEDDABBFCACDBAAFFAAFFCADFAADFDACFFAEDFCACFCAEBCE
```

Векторизация кодограммы ЭКГ-сигнала

Вход: *кодограмма* $d = (s_1, \ldots, s_{N-1})$ как текстовая строка

Выход: частоты триграмм n_{dw} — сколько раз триграмма w появилась в кодограмме d, $w \in W$, словарь $|W| = 6^3 = 216$

1. FFA - 42	17. EFF - 10	33. CEC - 6	49. EAC - 3
2. FAA - 33	18. DAA - 10	34. ADB - 5	50. DDA - 3
3. AFF - 32	19. ECF - 9	35. FFE - 5	51. CAC - 3
4. AAF - 30	20. FFC - 9	36. EBF - 5	52. EDF - 3
5. ADF - 18	21. FEA - 9	37. CFD - 5	53. EFB - 3
6. FCA - 18	22. DFC - 8	38. AFB - 4	54. DBA - 3
7. ACF - 17	23. ABF - 8	39. AAE - 4	55. FCC - 2
8. AAD - 15	24. AAB - 8	40. CFC - 4	56. AFC - 2
9. CFF - 14	25. FCE - 8	41. CAE - 4	57. EAA - 2
10. AEF - 13	26. AEB - 7	42. DAC - 4	58. CED - 2
11. FDA - 13	27. DFD - 7	43. DBF - 4	59. CAA - 2
12. FAE - 12	28. ACD - 6	44. BFC - 4	60. BCA - 2
13. FAC - 12	29. CDF - 6	45. CFB - 4	61. BBA - 2
14. FBA - 11	30. DFA - 6	46. AED - 3	62. DFF - 2
15. BFA - 11	31. CAF - 6	47. FFF - 3	63. BDA - 2
16. BAA - 11	32. CAD - 6	48. FBC - 3	64. DAE - 2

Тематическая модель классификации документов

Соответствие тематического моделирования и диагностики:

- документ $\leftrightarrow$ кодограмма ЭКГ
- слово $\leftrightarrow$ триграмма
- класс или категория $\leftrightarrow$ заболевание
- тема $\leftrightarrow$ диагностический эталон класса

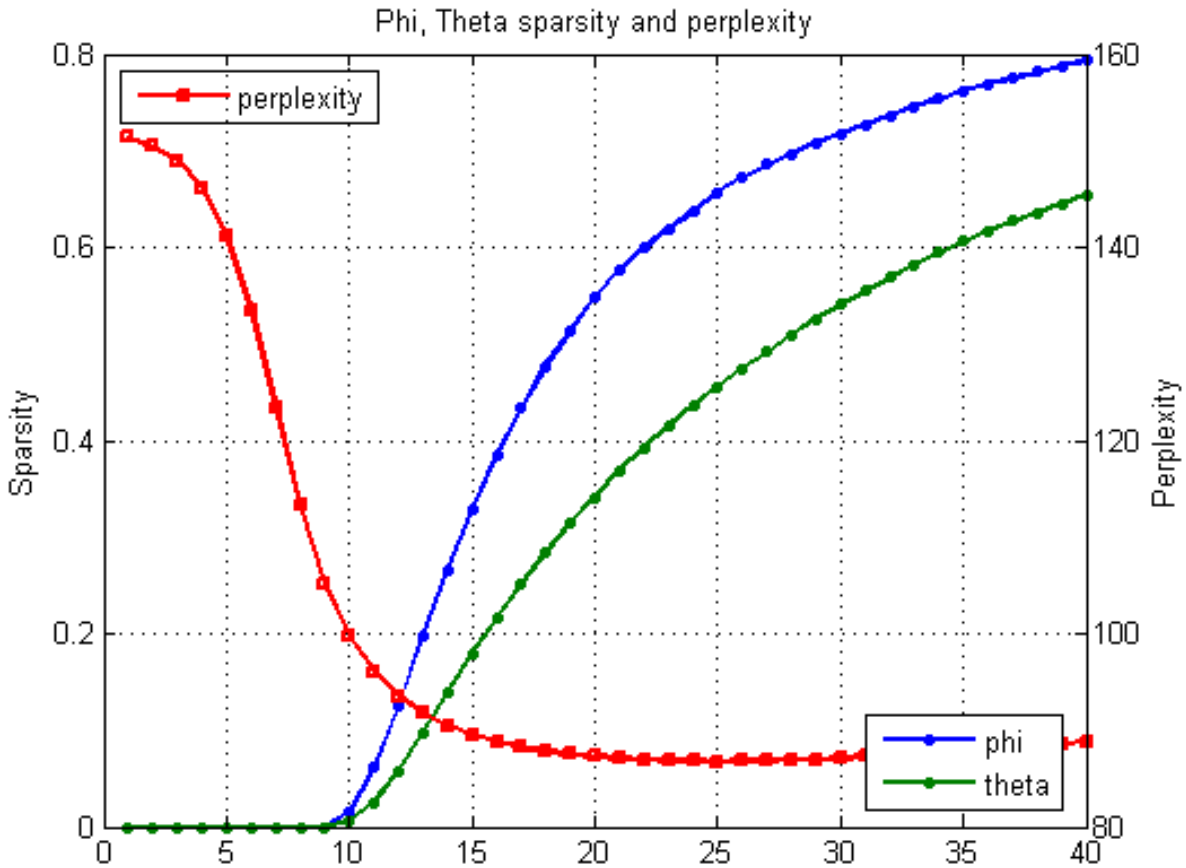
Классификация документов по классам $c \in C$:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct} \theta_{td} \to \max$$

где m_{dc} — индикатор принадлежности документа d классу c,
$\psi_{ct} = p(c|t)$ — ещё одна матрица параметров модели

Мониторинг качества модели в EM-алгоритме

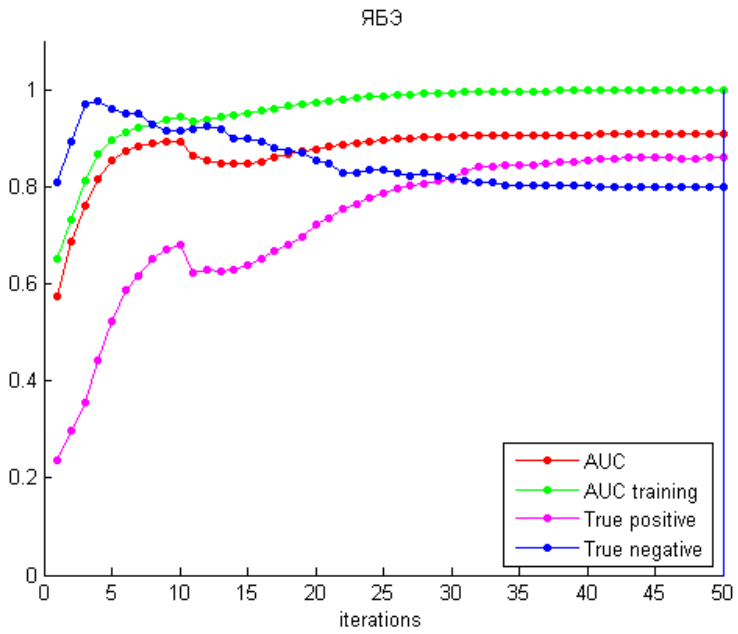
Разреженность матриц Φ, Θ и перплексия модели, при разреживании с 10й итерации (язвенная болезнь)



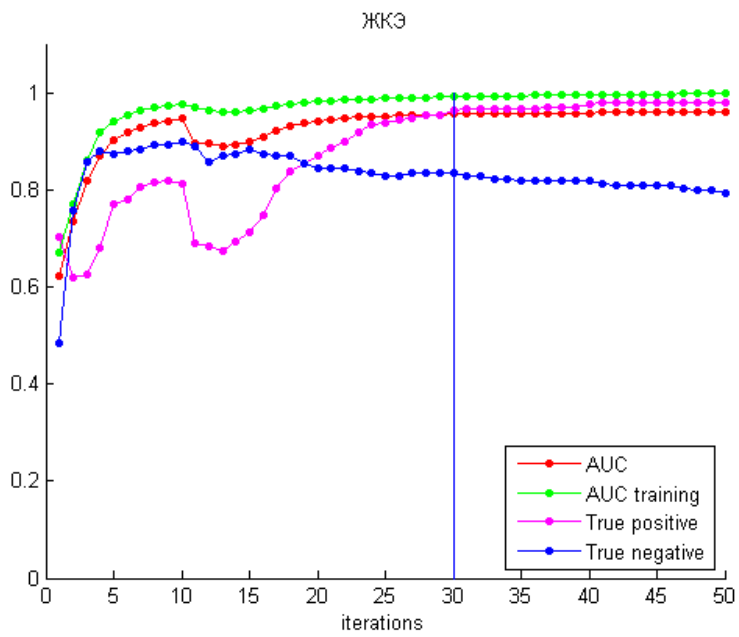
Мониторинг качества модели в EM-алгоритме

Включение разреживающего регуляризатора с 10й итерации

Язвенная болезнь



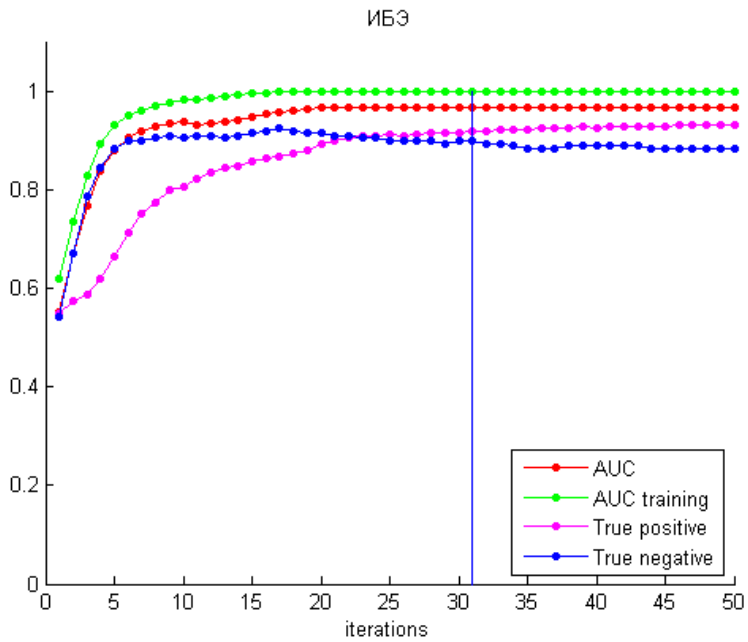
Желчнокаменная болезнь



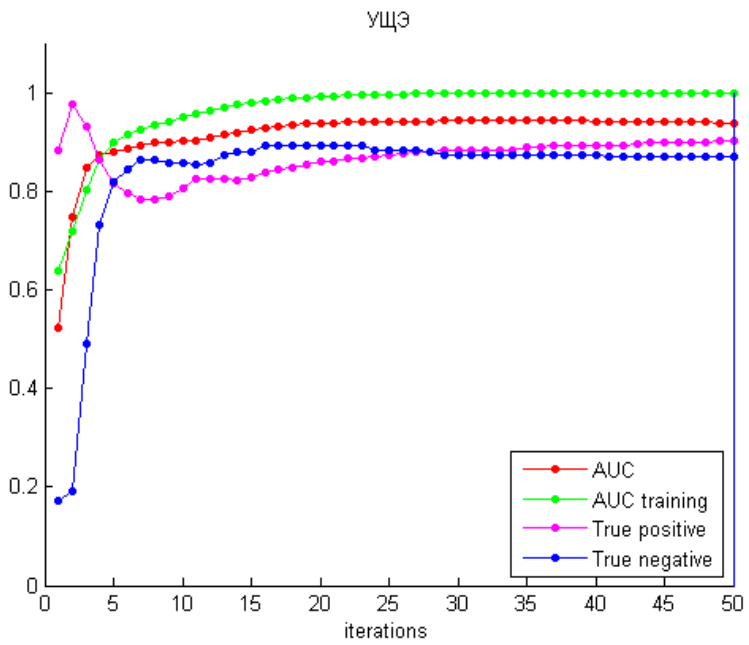
Мониторинг качества модели в EM-алгоритме

Включение разреживающего регуляризатора с 10й итерации

Ишемическая болезнь сердца



Узловой зоб щитовидной железы



Результаты кросс-валидации

Обучающая выборка — для оптимизации параметров модели
Тестовая выборка — для оценивания чувс., спец., AUC
40×10-fold cross-validation — для доверительного оценивания

болезнь	выборка	AUC, %	C% при Ч=95%
некроз головки бедренной кости	327	99.19 ± 0.10	96.6 ± 1.76
желчнокаменная болезнь	277	98.98 ± 0.23	94.4 ± 1.54
ишемическая болезнь сердца	1262	97.98 ± 0.14	91.1 ± 1.86
гастрит	321	97.76 ± 0.11	88.3 ± 2.64
гипертоническая болезнь	1891	96.76 ± 0.09	84.7 ± 1.99
сахарный диабет	868	96.75 ± 0.19	85.3 ± 2.18
аденома простаты	257	96.49 ± 0.13	80.1 ± 3.19
рак	525	96.49 ± 0.28	82.2 ± 2.38
узловой зоб щитовидной железы	750	95.57 ± 0.16	73.5 ± 3.41
холецистит хронический	336	95.35 ± 0.12	74.8 ± 2.46
дискинезия ЖВП	714	94.99 ± 0.16	70.3 ± 4.67
мочекаменная болезнь	649	94.99 ± 0.11	69.3 ± 2.14
язвенная болезнь	779	94.62 ± 0.10	63.6 ± 2.55

Выводы

- Неожиданное применение методов вычислительной лингвистики к анализу электрокардиосигналов
- Удивительно высокая точность диагностики *многих* заболеваний *по одной* электрокардиограмме
- Совокупность регуляризаторов, позволяющая находить свой набор диагностических эталонов для каждой болезни:

  + Разреживание диагностических эталонов
  + Сглаживание фоновых тем, не относящимся к болезням
  + Декоррелирование диагностических эталонов
  + Регуляризатор для оптимизации числа тем
  + Явная оптимизация AUC:

$$R(\Psi,\Theta) = -\tau \sum_{c\in C} \sum_{d\in D_c} \sum_{d'\notin D_c} \mathscr{L}\left(\sum_{t\in T} \psi_{ct}(\theta_{td} - \theta_{td'})\right) \to \max$$

От задач с классами к задачам со многими модальностями

Примеры модальностей в текстовых документах:

- слова — основная модальность
- слова каждого языка — отдельная модальность
- категории рубрикатора
- авторы документов
- время создания документа
- документы, ссылающиеся на данный
- документы, на которые ссылается данный
- сущности (entity), упоминаемые в текстах
- признаки на изображениях, связанных с текстом
- пользователи, смотревшие документ
- рекламные баннеры, просмотренные вместе с документом

Мультимодальные тематические модели

Произвольное число модальностей X_j, $j = 1, \ldots, m$.
Вероятностное пространство $D \times T \times X$, $X = X_1 \sqcup \cdots \sqcup X_m$.
Каждый документ d состоит из токенов $x_1, \ldots, x_{n_d} \in X$.
Тематическая модель j-й модальности:

$$p(x|d) = \sum_{t \in T} p(x|t)\, p(t|d) = \sum_{t \in T} \phi_{xt}\theta_{td}, \quad x \in X_j, \quad d \in D$$

Задача максимизации взвешенного правдоподобия:

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

при ограничениях нормировки и неотрицательности

$$\phi_{xt} \geqslant 0; \quad \sum_{x \in X_j} \phi_{xt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t} \theta_{td} = 1$$

Регуляризованые мультимодальные тематические модели

Теорема (EM-алгоритм)

Точка максимума $\mathscr{L}(\Phi,\Theta)+R(\Phi,\Theta)$ удовлетворяет системе уравнений со вспомогательными переменными p_{tdx}, n_{xt}, n_{tdj}:

$$\text{E-шаг:}\quad p_{tdx}=\frac{\phi_{xt}\theta_{td}}{\sum\limits_{s\in T}\phi_{xs}\theta_{sd}};$$

$$\text{M-шаг:}\quad \phi_{xt}\propto\left(n_{xt}+\phi_{xt}\frac{\partial R}{\partial\phi_{xt}}\right)_+;\quad n_{xt}=\sum_{d\in D}n_{dx}p_{tdx};$$

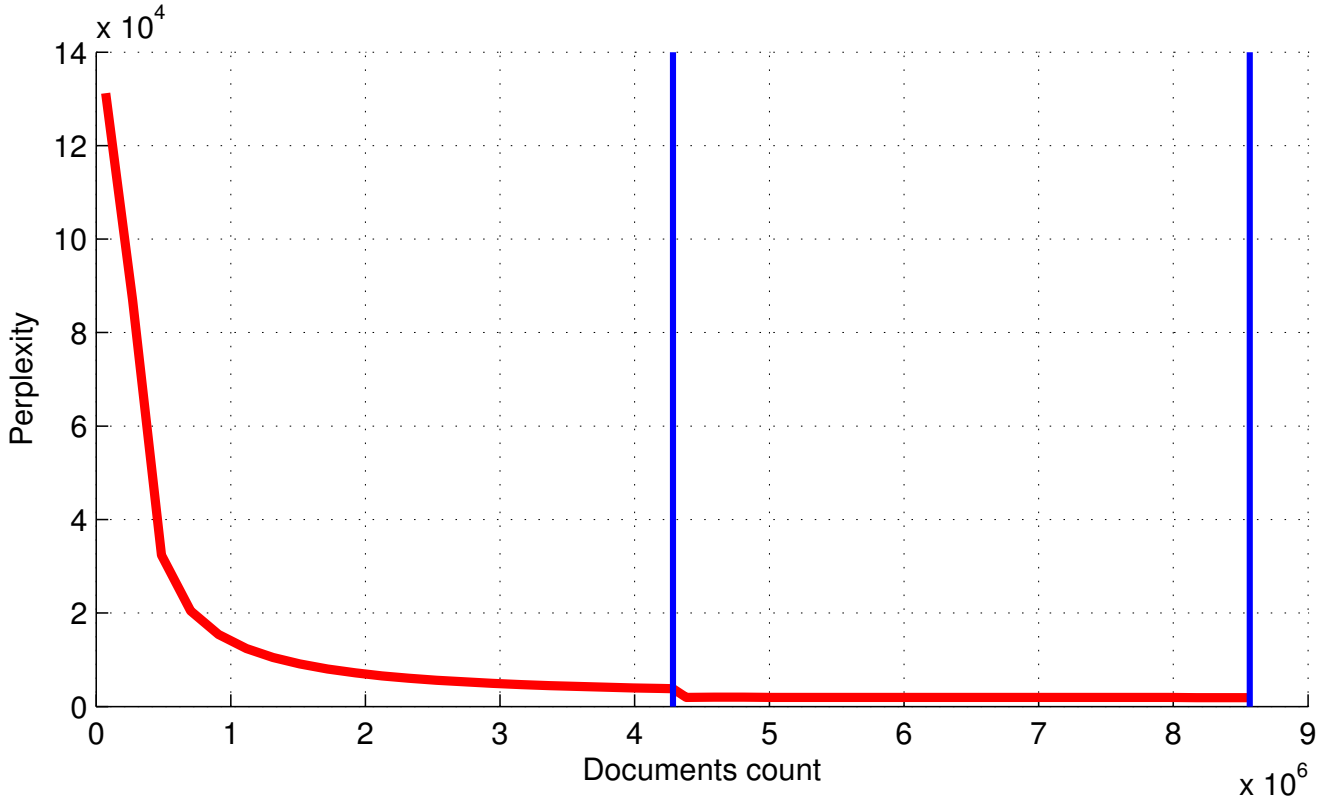
$$\theta_{td}\propto\left(\sum_{j=1}^{m}\tau_j n_{tdj}+\theta_{td}\frac{\partial R}{\partial\theta_{td}}\right)_+;\quad n_{tdj}=\sum_{x\in X_j}n_{dx}p_{tdx}.$$

BigARTM — библиотека тематического моделирования

- параллельная, распределённая, кроссплатформенная
- быстрый онлайновый алгоритм тематического моделирования
- поддержка мультимодальных моделей
- поддержка plugin-библиотеки регуляризаторов ARTM
- поддержка plugin-библиотеки метрик качества
- http://bigartm.org
  — проект с открытым кодом
- https://github.com/bigartm/bigartm
  — открытый репозиторий
- Разработчики: **Александр Фрей**, Мурат Апишев, Антон Марценюк, Андрей Шадриков, Никита Шаповалов, Никита Дойков, Анна Потапенко, Марина Дударенко

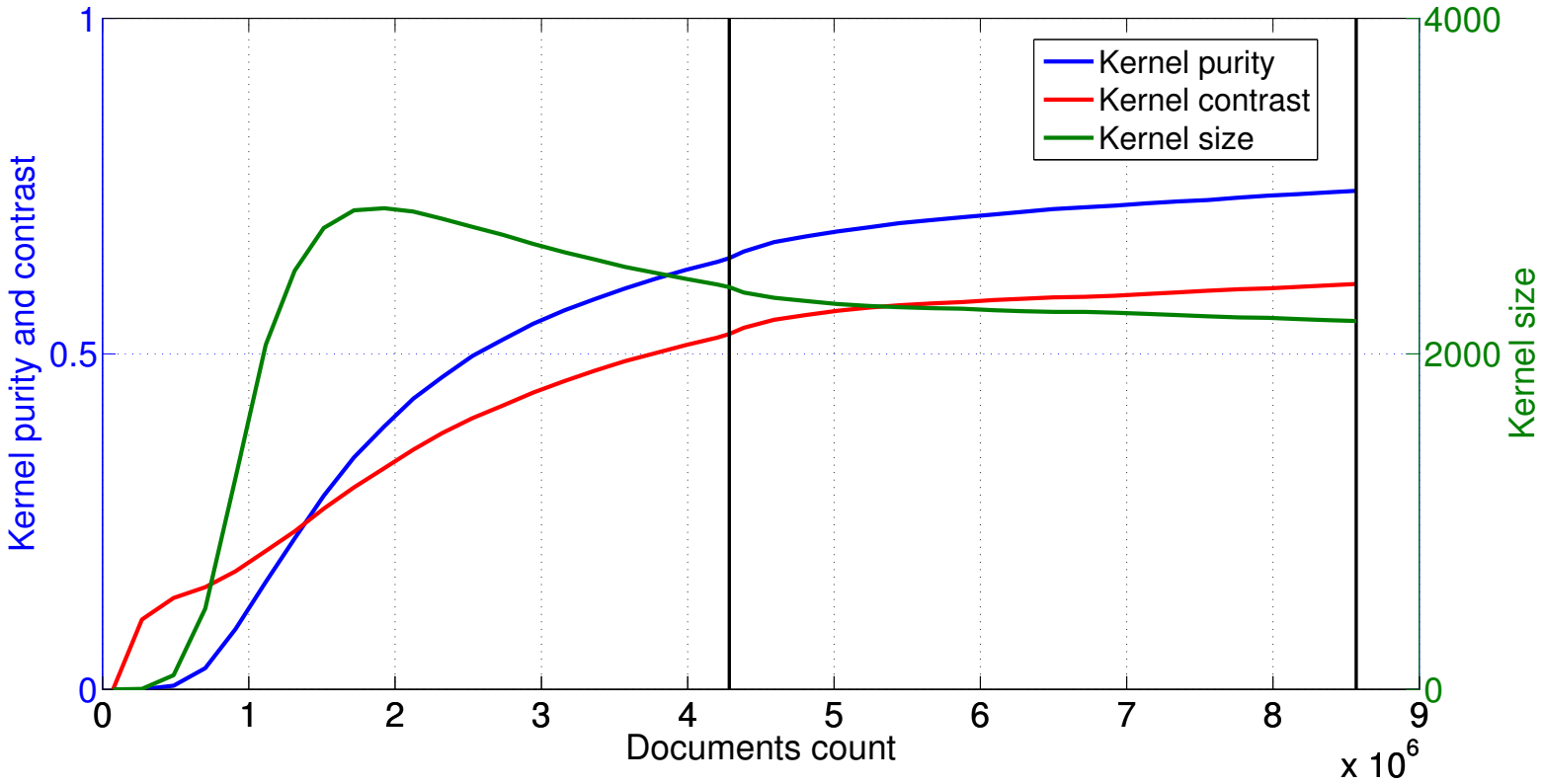
Эксперименты на PubMed

Коллекция PubMed: 7.6 Гб, $|D| = 4.8 \cdot 10^6$ документов общей длиной $n = 7 \cdot 10^8$ слов, объём словаря $|W| = 1.4 \cdot 10^5$ слов
100 тем. 8 ядер, 15 Гб оперативной памяти.
Сходимость перплексии за 2 прохода коллекции (1300 секунд):



Эксперименты на Pubmed

Сходимость других метрик качества за 2 прохода коллекции



Текущие исследования

- Мультиязычные модели [в докладе М.А.Дударенко]
- Иерархические тематические модели
- Автоматическое определение числа тем
- Автоматическое выделение терминов-словосочетаний
- Оптимизация траектории регуляризации
- Построение лексических цепочек и выявление тематической структуры последовательного текста

Прикладные задачи:

- от тематической модели ИОИ + ММРО + JMLDA + ...
- ...до систематизации научного контента Рунета

Литература

- *Hofmann T.* Probabilistic Latent Semantic Indexing // SIGIR, 1999.
- *Blei D., Ng A., Jordan M.* Latent Dirichlet Allocation // Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.
- *Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models // Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.
- *Newman D., Lau J. H., Grieser K., Baldwin T.* Automatic Evaluation of Topic Coherence // Human Language Technologies, HLT-2010, Pp. 100–108.
- *Rubin T. N., Chambers A., Smyth P., Steyvers M.* Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 1–2.

- *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. — Т. 455., № 3. С. 268–271.
- *Vorontsov K. V., Potapenko A. A.* Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // Analysis of Images, Social Networks and Texts. Ekaterinburg, 10–12 April 2014. Springer.
- *Vorontsov K. V., Potapenko A. A.* Additive Regularization of Topic Models // Machine Learning Journal. Springer (to appear).

Воронцов Константин Вячеславович
voron@forecsys.ru

Страницы на www.MachineLearning.ru:

- Участник:Vokov
- Вероятностные тематические модели (курс лекций, К. В. Воронцов)
- Тематическое моделирование