# Вероятностные тематические модели коллекций текстовых документов

К. В. Воронцов
vokov@forecsys.ru

Просеминар кафедры
«Математические методы прогнозирования» ВМК МГУ
http://www.MachineLearning.ru/wiki «ММП»

26 февраля 2013

# Содержание

# Определения и обозначения

**Дано:**

$W$ — словарь, множество слов (терминов)

$D$ — множество (коллекция, корпус) текстовых документов

$n_{dw}$ — сколько раз термин $w \in W$ встретился в документе $d \in D$

**Найти:**

— к каким темам относится каждый документ

— какими терминами определяется каждая тема

**Дополнительно найти**:

— сколько тем содержится в коллекции

— как темы делятся на подтемы, образуя иерархию

— как темы развиваются во времени

— тематику объектов, связанных с документами: рисунков, авторов, журналов, конференций, организаций, стран и т. д.

# Цели тематического моделирования (topic modeling)

- Тематический поиск документов и объектов по тексту любой длины или по любому объекту
- Категоризация, классификация, аннотирование, суммаризация текстовых документов

**Типичные приложения:**

- Поиск научной информации
- Поиск экспертов (expert search), рецензентов, проектов
- Выявление трендов и фронта исследований
- Анализ и агрегирование новостных потоков
- Рекомендательные сервисы (коллаборативная фильтрация)
- Рубрикация коллекций изображений, видео, музыки
- Аннотация генома и другие задачи биоинформатики

# Обобщения и модификации тематических моделей

- Hierarchical models — строят иерархическую структуру тем
- Temporal models — учитывают годы публикаций
- Author-topic models — оценивают распределение авторов $p(a|w,d)$ для каждого слова документа
- Entity-topic models — оценивают тематику связанных объектов (минералы, животные, растения, вещества, гены, страны, народы, личности, фирмы, изделия, и т. п.)
- Модели, учитывающие связь слов внутри документа
- Модели связей между документами (ссылки, цитирование)

*Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad.*
Knowledge discovery through directed probabilistic topic models: a survey.
Frontiers of Computer Science in China, Vol. 4, No. 2., 2010, Pp. 280–301.
(русский перевод на MachineLearning.ru)

**Topic Modeling Bibliography:**
`http://www.cs.princeton.edu/~mimno/topics.html`

# Этапы предварительной обработки текстов

- Распознавание класса документа (научный? реферат? художественный? публицистика?)
- Удаление переносов, чисел, колонтитулов, таблиц, остатков формул, оглавлений и т. д.
- Удаление опечаток и ошибок сканирования
- Выделение метаописания: название, авторы, год и т. д.
- Выделение библиографических ссылок
- Приведение всех слов к нормальной форме (лемматизация или стемминг)
- Выделение терминов (term extraction) и/или выделение словосочетаний (key phrase extraction) (сводятся к задачам классификации или ранжирования)
- Удаление общеупотребительных слов (стоп-слов)
- Удаление слишком редких специфических слов

# Формализация постановки задачи

**Гипотезы о вероятностной природе данных:**

1. каждое слово в документе связано с некоторой темой $t \in T$
2. гипотезы «мешка слов» и «мешка документов»:

   коллекция $D$ — это выборка независимых наблюдений $(d, w)$ из дискретного распределения $p(d, w, t)$ на $D \times W \times T$;
   тема $t$ — скрытая переменная


3. гипотеза условной независимости: $p(w|d,t) = p(w|t)$;

**Вероятностная модель порождения документа $d$:**

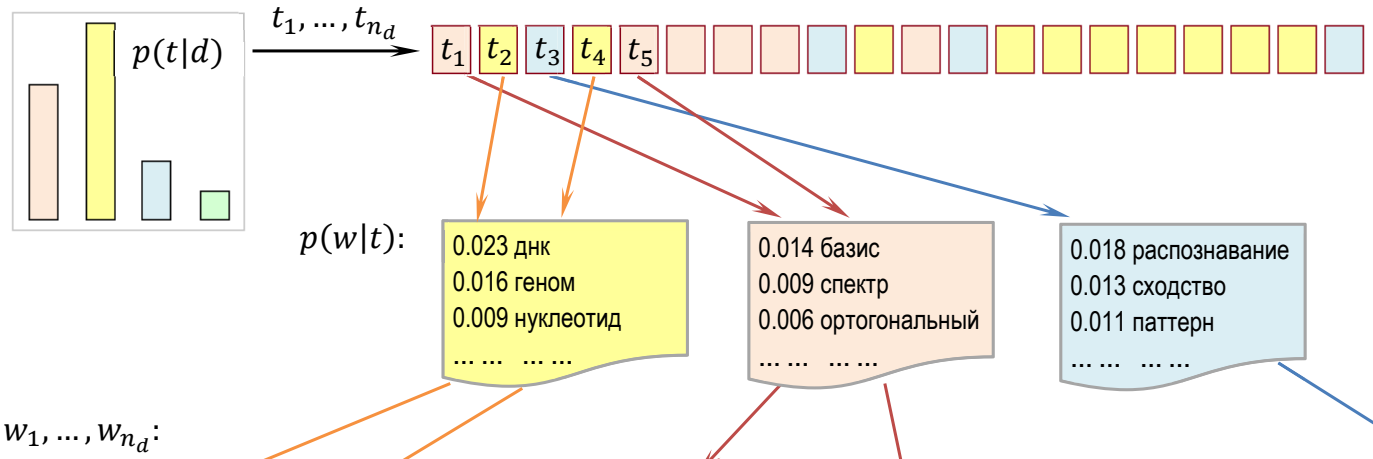
$$p(w|d) = \sum_{t \in T} p(w|t)\, p(t|d)$$

**По заданным $p(w|d)$ найти:**

- $p(w|t)$ — распределение терминов в каждой теме $t \in T$;
- $p(t|d)$ — распределение тем в каждом документе $d \in D$.

# Вероятностная модель порождения документа $d$

Вероятностная тематическая модель: $p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$




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).

# Модель PLSA — Probabilistic Latent Semantic Analysis (1999)

Метод максимума правдоподобия по $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$:

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

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w\in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{d\in D} \theta_{td} = 1.$$

**Интерпретация:** неотрицательное матричное разложение

$$\|F - \Phi\Theta\|_{\mathrm{KL}} \to \min_{\Phi,\Theta}$$

$F = \big(\hat{p}(w|d)\big)_{W\times D}$ — известная матрица исходных данных;
$\Phi = \big(\phi_{wt}\big)_{W\times T}$ — искомая матрица терминов тем $\phi_{wt} = p(w|t)$;
$\Theta = \big(\theta_{td}\big)_{T\times D}$ — искомая матрица тем документов $\theta_{td} = p(t|d)$.

# Частотные оценки условных вероятностей

Вероятностная тематическая модель: $p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$

Если рассматривать коллекцию как выборку троек $(d, w, t)$, то

$$\hat{p}(w|d) = \frac{n_{dw}}{n_d}, \qquad \hat{p}(w|t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(t|d) = \frac{n_{dt}}{n_d};$$

$n_{dwt}$ — число троек $(d, w, t)$ во всей коллекции;

$n_{dw} = \sum_{t \in T} n_{dwt}$ — число вхождений термина $w$ в документ $d$;

$n_{dt} = \sum_{w \in d} n_{dwt}$; $\qquad n_d = \sum_{w \in d} \sum_{t \in T} n_{dwt}$ — длина документа $d$;

$n_{wt} = \sum_{d \in D} n_{dwt}$; $\qquad n_t = \sum_{d \in D} \sum_{w \in d} n_{dwt}$ — «длина темы» $t$;

$n = \sum_{d \in D} \sum_{w \in d} \sum_{t \in T} n_{dwt}$ — длина всей коллекции;

# EM-алгоритм (Expectation–Maximization)

**E-шаг:** условные вероятности тем $p(t|d,w)$ для всех $t, d, w$ вычисляются через $\phi_{wt}$, $\theta_{td}$ по формуле Байеса:

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

**M-шаг:** решение задачи максимизации правдоподобия выражается аналитически через частотные оценки условных вероятностей, если положить $\hat{n}_{dwt} = n_{dw}p(t|d,w)$:

$$\phi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \qquad \hat{n}_{wt} = \sum_{d \in D} \hat{n}_{dwt}, \qquad \hat{n}_t = \sum_{w \in W} \hat{n}_{wt};$$

$$\theta_{td} = \frac{\hat{n}_{dt}}{\hat{n}_d}, \qquad \hat{n}_{dt} = \sum_{w \in d} \hat{n}_{dwt}, \qquad \hat{n}_d = \sum_{t \in T} \hat{n}_{dt}.$$

**EM-алгоритм** — это чередование E и M шагов до сходимости.

# Рационализация EM-алгоритма: E-шаг встроен внутрь M-шага

**Идея:** не хранить $p(t|d,w)$, а вычислять по мере необходимости.
Сложность алгоритма $O(|D| \cdot |W| \cdot |T|)$.

---

**Вход:** коллекция $D$, число тем $|T|$, начальные $\Phi$ и $\Theta$;
**Выход:** распределения $\Phi$ и $\Theta$;

---

**повторять**
  обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ для всех $d \in D$, $w \in W$, $t \in T$;
  **для всех** $d \in D$, $w \in d$
    $p(t|d,w) := \phi_{wt}\theta_{td} \,/\, \sum_\tau \phi_{w\tau}\theta_{\tau d}$, для всех $t \in T$;
    **для всех** $t \in T$ таких, что $\phi_{wt}\theta_{td} > 0$
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на $n_{dwt} = n_{dw}p(t|d,w)$;
  $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W$, $t \in T$;
  $\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D$, $t \in T$;
**пока** $\Phi$ и $\Theta$ не стабилизируются.

# Эвристики

**Эвристики** — это «разумные» оценки, формулы, правила, не имеющие строгих теоретических обоснований, но дающие хорошие результаты на практике.

- как инициализировать переменные $\phi_{wt}$, $\theta_{td}$?
- когда остановить итерационный процесс?
- как выбрать число тем $T$?
- как добиться разреженности $p(w|t)$, $p(t|d)$, $p(t|d,w)$?
- как ускорить сходимость без потери качества?
- как вообще оценивать качество тематической модели?
- как учесть массу лингвистических тонкостей?

# Робастная модель с фоном и шумом

**Гипотеза:** каждое употребление термина в документе объясняется либо темой, либо специфично для данного документа (шум), либо это общеупотребительный термин (фон).

Модель смеси тематической, шумовой и фоновой компонент:

$$p(w|d) = \gamma p_{\text{ш}}(w|d) + \varepsilon p_{\text{ф}}(w) + (1 - \gamma - \varepsilon) \sum_{t \in T} p(w|t)\, p(t|d)$$

**Найти:**

$p(w|t)$, $p(t|d)$, $p_{\text{ш}}(w|d)$, $p_{\text{ф}}(w)$, для всех $d \in D$, $w \in W$, $t \in T$.

*Chemudugunta C., Smyth P., Steyvers M.* Modeling general and specific aspects of documents with a probabilistic topic model // Advances in Neural Information Processing Systems, MIT Press, 2006. — Vol. 19. — Pp. 241–248.

# «Многие знания — многие печали»

Наша модель допускает произвольные сочетания эвристик:

1. частое обновление параметров $\phi_{wt}$, $\theta_{td}$
2. сэмплирование темы $t$ из $p(t|d,w)$ для каждого $(d,w)$
3. сглаживание частотных оценок условных вероятностей
4. оптимизация параметров сглаживания
5. робастность
6. разреживание $\phi_{wt}$, $\theta_{td}$

Всего более 300 вариантов! Какой же из них лучший?

Для получения ответа сделано несколько сотен экспериментов на реальных данных. Каждый эксперимент — около 30 минут.

*Воронцов К. В., Потапенко А. А.* Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование, 2012. — Т. 4, №12. — С. 693–706.

# Методы оценивания качества тематических моделей

- На размеченной тестовой коллекции $D'$:
  — число ошибок классификации (чем меньше, тем лучше).
- На неразмеченной тестовой коллекции $D'$:
  — перплексия, неопределённость (чем меньше, тем лучше):

$$\text{perplexity} = \exp\left(-\frac{\sum\limits_{d\in D'}\sum\limits_{w\in d} n_{dw}\ln p(d,w)}{\sum\limits_{d\in D'}\sum\limits_{w\in d} n_{dw}}\right)$$

- На обучающей коллекции $D$: насколько нарушается гипотеза условной независимости $p(w|d,t) = p(w|t)$

$$\text{KL}\Big(\hat{p}(d,w|t),\ \hat{p}(d|t)\cdot\hat{p}(w|t)\Big) = \sum_{d,w}\frac{n_{dwt}}{n_t}\log\frac{n_{dwt}\cdot n_t}{n_{td}\cdot n_{wt}}$$

*David Mimno, David Blei*. Bayesian Checking for Topic Models // Empirical Methods in Natural Language Processing, 2011.

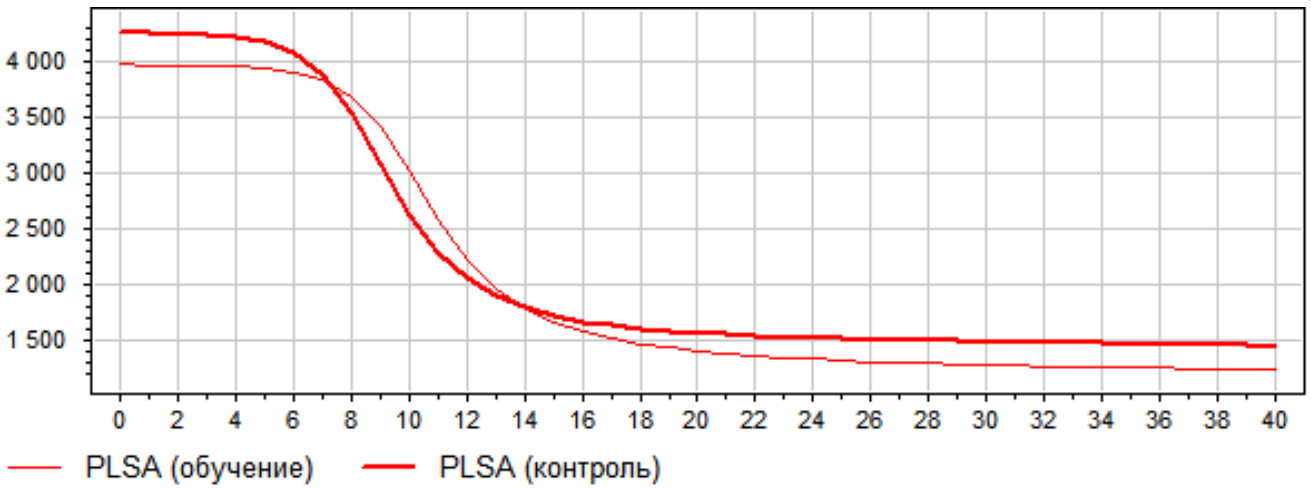
# Методика эксперимента

$D$ — коллекция 2000 авторефератов диссертаций на русском языке суммарной длины $n \approx 8.7 \cdot 10^6$, словарь $|W| \approx 3 \cdot 10^4$.

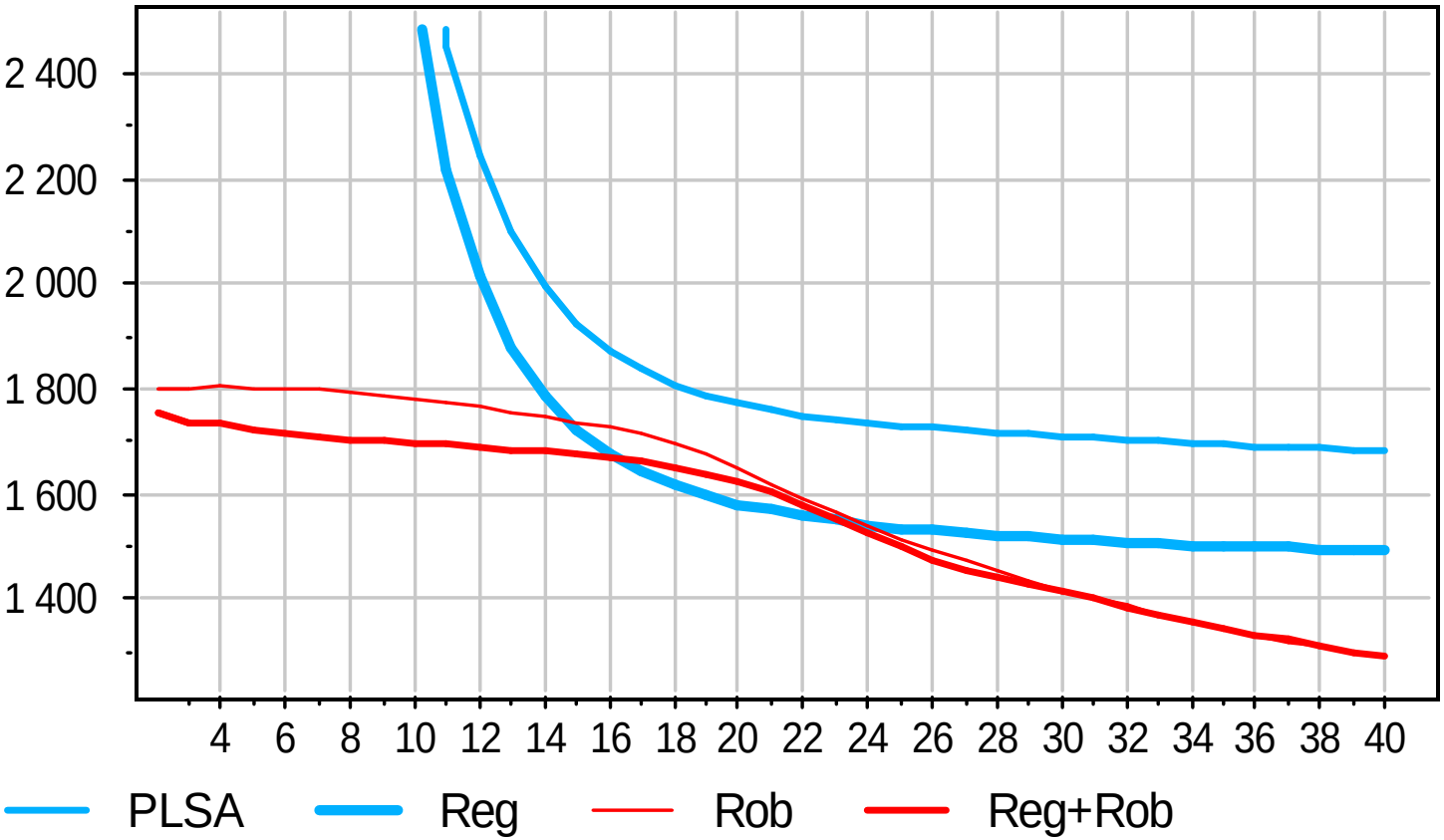
Предобработка: лемматизация, удаление стоп-слов.

$D'$ — коллекция 200 авторефератов, не включённых в $D$.

Строятся графики зависимости перплексии от числа итераций (проходов коллекции); число итераций 40; число тем $|T| = 100$;

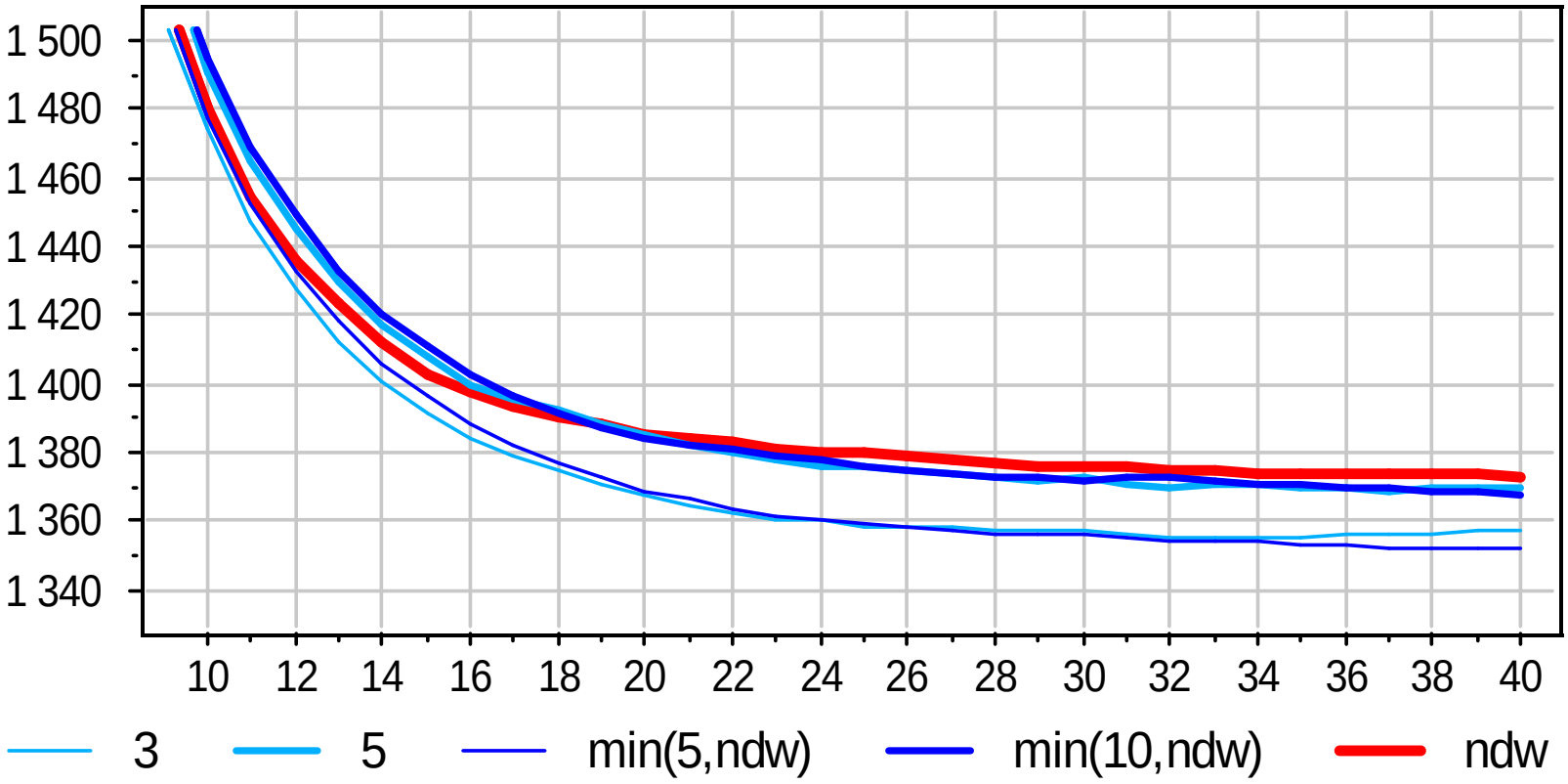


# Робастная модель не нуждается в регуляризации



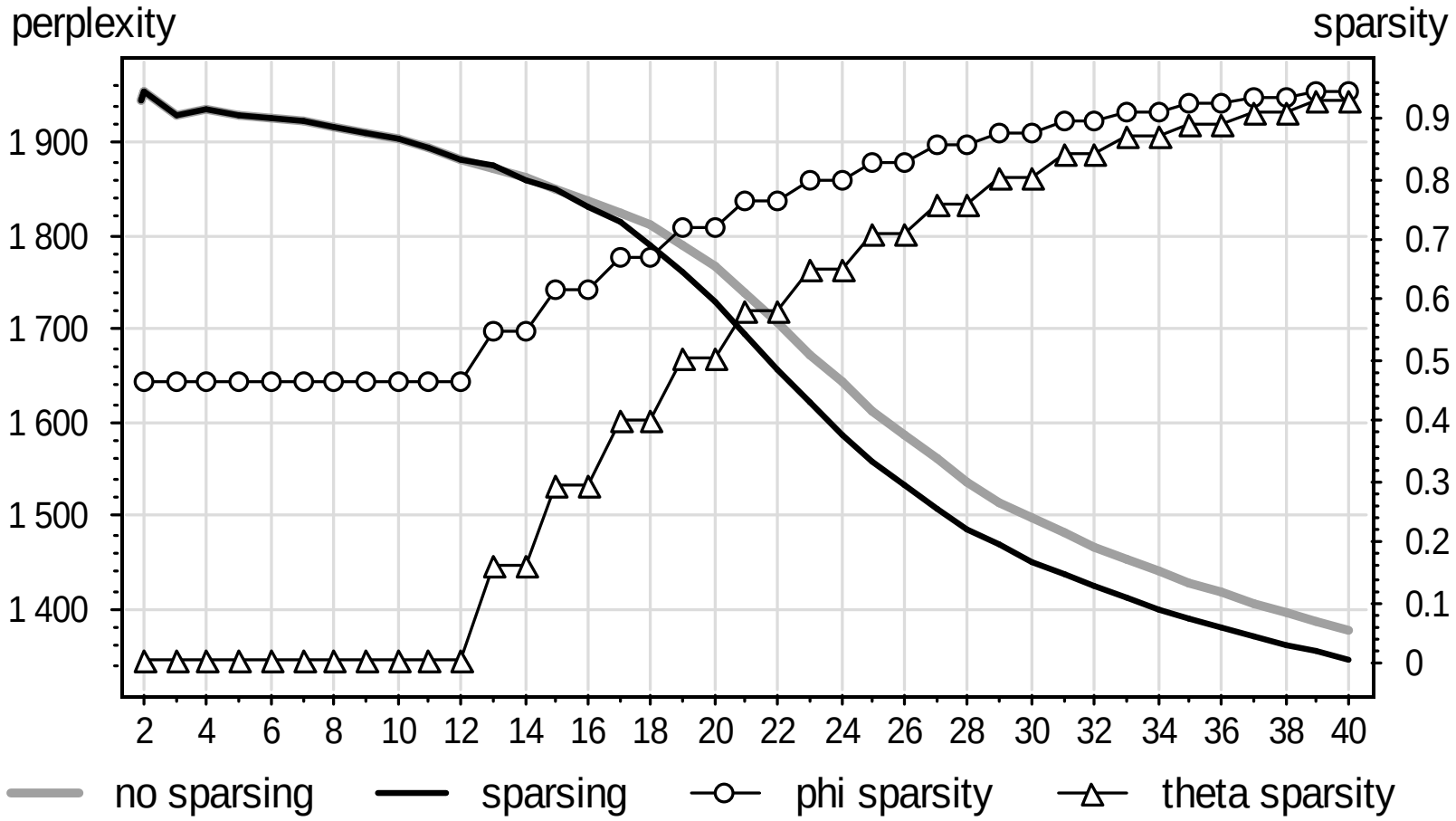

Робастность сильнее уменьшает перплексию PLSA, чем сглаживание. Сглаживание не улучает робастную модель.

# Экономное сэмплирование не ухудшает качество




Оптимально сэмплировать $\min\{5, n_{dw}\}$ тем.

# Сильное разреживание не ухудшает качество




Робастная модель допускает разреживание матриц $\Phi$, $\Theta$ на 90%.

# Улучшение качества и увеличение скорости

**Меры по улучшению качества тематических моделей:**

1. решение проблемы устойчивости: $F = \Phi\Theta = (\Phi R)(R^{-1}\Theta)$
2. учёт дополнительной информации (асессорских оценок, рубрикаторов) о связях документов и терминов с темами
3. оптимизация структуры разреживания
4. контроль требования условной независимости
5. учёт лингвистических знаний

**Меры по улучшению вычислительной эффективности:**

1. распараллеливание
2. разреживание $p(w|t)$, $p(t|d)$
3. разреживание $p(w|d)$

# Многофункциональная модель — открытая проблема

Для реализации научного поиска нужна тематическая модель, которая была бы одновременно:

- Иерархическая сетевая
- Сбалансированная и мелко гранулированная
- Линейно масштабируемая
- Робастная
- Разреженная
- Устойчивая
- Инкрементная
- Согласованная с экспертными оценками
- Согласованная с внешними рубрикаторами
- Многоязыковая

# НИР — научно-исследовательская работа

Общая информация и рекомендации для студентов:
www.MachineLearning.ru

- Математические методы прогнозирования (кафедра ВМиК МГУ)
- Научно-исследовательская работа (рекомендации)
- Написание отчётов и статей (рекомендации)
- Подготовка презентаций (рекомендации)
- Защита выпускной квалификационной работы (рекомендации)

# Примеры НИР: курсовые, дипломные, кандидатские

- Анна Потапенко, 4й курс ММП
  Многофункциональные вероятностные тематические модели
- Евгений Соколов, 5-й курс ММП
  Вычисление комбинаторных оценок вероятности переобучения методом случайных блужданий
- Валентин Полежаев, 5-й курс ММП
  Технология автоматического извлечения графа цитирования из коллекции научных документов
- Евгений Рябенко, асп. ММП
  Матричные разложения больших данных: применения в задачах анализа данных ДНК-микрочипов и тематического моделирования

# Другие проекты

- Комбинаторная теория переобучения
  - повышение точности оценок
  - улучшение с их помощью методов обучения
- Проекты Форексис
  - АнтиПлагиат
  - прогнозирование объёмов продаж
  - клиентская аналитика
  - рекомендательный сервис
  - анализ автотранспортных потоков
- Полигон алгоритмов классификации
  - расширение методики тестирования
  - развитие сервисов

# Спасибо за внимание!

Воронцов Константин Вячеславович
voron@forecsys.ru

Страницы на www.MachineLearning.ru:

- Участник:Vokov
- Математические методы прогнозирования (кафедра ВМиК МГУ)
- Научно-исследовательская работа (рекомендации)
- Машинное обучение (курс лекций, К.В.Воронцов)
- Теория надёжности обучения по прецедентам (курс лекций, К. В. Воронцов)
- Тематическое моделирование
- Полигон алгоритмов