# Вероятностные тематические модели коллекций текстовых документов

К. В. Воронцов
vokov@forecsys.ru

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

ноябрь 2011

# Содержание

# Определения и обозначения

**Дано:**

$W$ — словарь, множество слов (терминов);

$D$ — множество текстовых документов;

каждый $d \in D$ — это последовательность слов из $W$.

**Найти:**

$T$ — множество скрытых (латентных) тем;

$p(w|t)$ — распределение на $W$, задающее тему $t \in T$;

$p(t|d)$ — *тематический профиль* документа, для всех $d \in D$.

**Дополнительно:**

$p(w|t, y)$ — изменение темы по годам $y$;

$p(t|y)$, $p(y|t)$ — распределения тем по годам;

$p(t|a)$, $p(t|a, y)$ — *тематический профиль* автора $a$;

$p(t|x)$, $p(t|x, y)$ — *тематический профиль* объекта $x$, связанного с документами (журнала, конференции, организации, страны);

# Цели тематического моделирования (topic modeling)

- Тематический поиск (запрос — тема или документ)
- Рубрикация и классификации текстов
- Поиск экспертов
- Прослеживание фронта исследований
- Аннотация документов
- Суммаризация множества документов

**Типичные приложения:**

- Анализ коллекций научных статей
- Анализ новостных потоков
- Рубрикация коллекций изображений, видео, музыки
- Аннотация генома и другие задачи биоинформатики

# Стандартные гипотезы тематического моделирования

1. Порядок документов в коллекции не важен
2. Порядок слов в документе не важен (bag of words)
3. Слова, встречающиеся в большинстве документов, не важны
4. Слово в разных формах — это одно и то же слово

**Предварительная обработка текстов:**

- Приведение всех слов к нормальной форме (стемминг или лемматизация)
- Выделение терминов (term extraction) и выделение словосочетаний (key phrase extraction); (сводятся к задачам классификации или ранжирования)
- Удаление стоп-слов $w \in W$: $\#\{d : w \in d\} \geqslant \alpha|D|$, $\alpha \sim 0.05 \ldots 0.5$

**Основная вероятностная гипотеза:**
коллекция документов — i.i.d. выборка $\big\{(d, w) : d \in D,\ w \in d\big\}$.

# Униграммная модель порождения текста

**Дополнительная вероятностная гипотеза:**
появления слов $w$ в документе $d$ — независимые события.

**Вероятностная модель документа:**

$$p(d) = \prod_{w \in d} p(w|d)^{n_{dw}},$$

$p(w|d)$ — неизвестное мультиномиальное распределение на $W$;
$n_{dw}$ — число вхождений слова $w$ в документ $d$.

**Принцип максимума правдоподобия:**

$$\ln \prod_{d \in D} \prod_{w \in d} p(w|d)^{n_{dw}} \to \max_{\{p(w|d)\}} \quad \text{при} \quad \sum_{w \in W} p(w|d) = 1.$$

Лагранжиан:

$$\mathscr{L} = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) - \sum_{d \in D} \lambda_d \left( \sum_{w \in W} p(w|d) - 1 \right).$$

# Униграммная модель порождения текста

Лагранжиан $\mathscr{L}$ распадается на независимые слагаемые $\mathscr{L}(d)$:

$$\mathscr{L}(d) = \sum_{w \in d} n_{dw} \ln p(w|d) - \lambda_d \left( \sum_{w \in W} p(w|d) - 1 \right).$$

$$\frac{\partial \mathscr{L}}{\partial p(w|d)} = \frac{\partial \mathscr{L}(d)}{\partial p(w|d)} = \frac{n_{dw}}{p(w|d)} - \lambda_d = 0,$$

где $n_{dw}$ — число вхождений слова $w$ в документ $d$.

Умножим обе части равенства на $p(w|d)$ и просуммируем по $w$:

$$\sum_{w \in W} \lambda_d p(w|d) = \sum_{w \in W} n_{dw} \;\Rightarrow\; \lambda_d = n_d,$$

где $n_d$ — длина документа $d$.

Получим (тривиальную) оценку максимума правдоподобия:

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

# Униграммная модель порождения текста

**Недостатки униграммной модели:**

- тематика не выявляется
- число $|W| \cdot |D|$ оцениваемых параметров $p(w|d)$ линейно зависит от $|D|$ — числа документов в коллекции
- зависимости между документами не учитываются

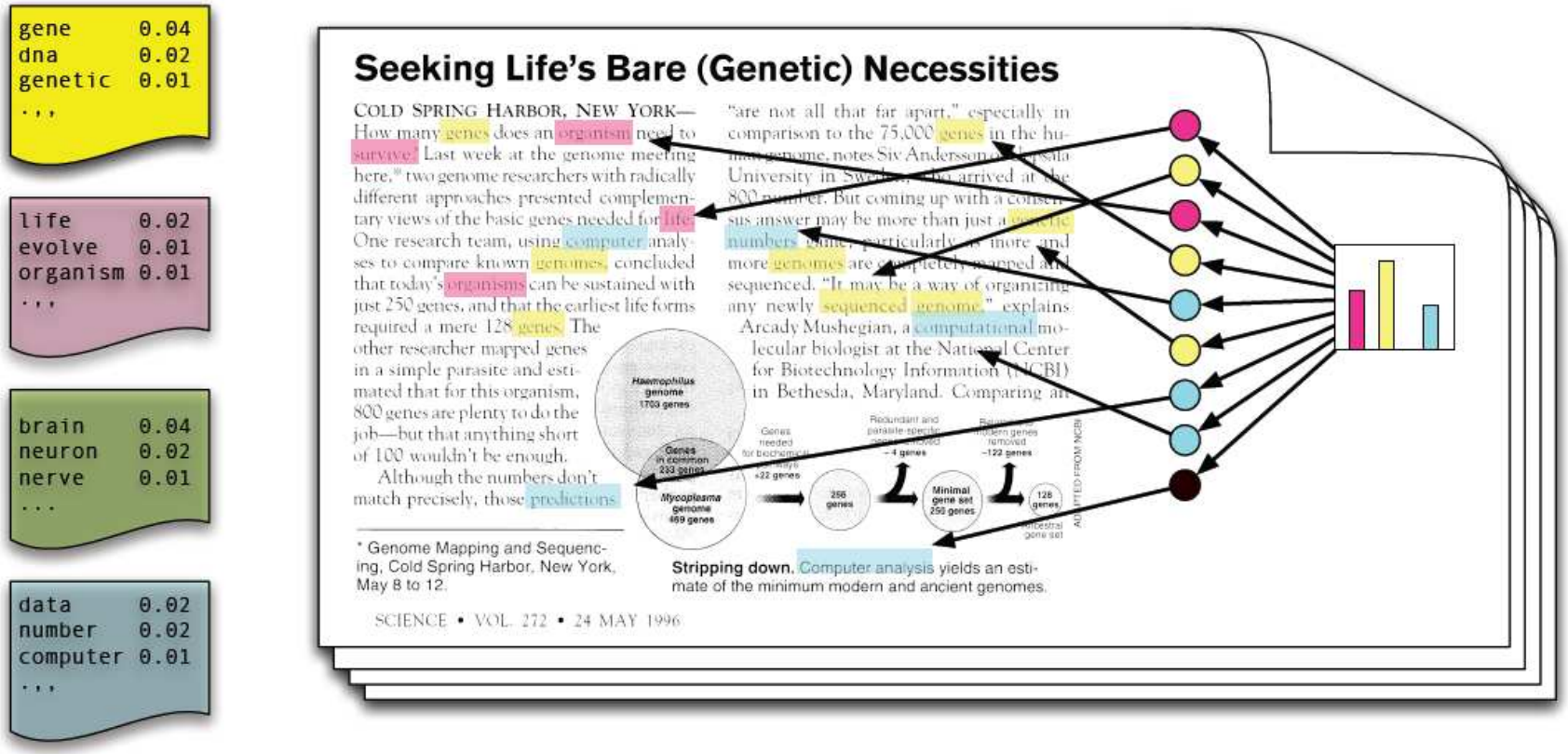
Эти недостатки устраняются в модели смеси униграмм:

*Nigam, McCallum, Thrun, Mitchell.*
Text classification from labeled and unlabeled documents using EM.
Journal of Machine Learning, 2000, 39(2–3): 103–134

# Вероятностные тематические модели порождения текста

темы $t \in T$ документы $d \in D$ слова темы

$p(w|t)$ из $p(w|t)$ из $p(t|d)$



# Модель смеси униграмм [Nigam и др., 2000]

1. Будем описывать униграммной моделью не весь документ $d$, а только ту его часть, которая относится к теме $t \in T$:
$$p(d|t) = \prod_{w \in d} p(w|t,d)^{n_{dw}}.$$
2. **Гипотеза условной независимости:** $p(w|t,d) = p(w|t)$ (распределения слов связаны с темами, а не с документами)
3. Документ — это смесь униграмм:
$$p(d) = \sum_{t \in T} p(t)p(d|t).$$

**Тематическая модель смеси униграмм:**

$$p(d) = \sum_{t \in T} p(t) \prod_{w \in d} p(w|t)^{n_{dw}}.$$

# Преимущества и недостатки модели смеси униграмм

**Преимущества:**

- Модель позволяет выявлять тематику
- *Тематический профиль* каждого документа выражается через параметры модели $p(t)$, $p(w|t)$ по формуле Байеса:

$$p(t|d) = \frac{p(t)p(d|t)}{p(d)} = \frac{p(t)\prod\limits_{w\in d} p(w|t)^{n_{dw}}}{\sum\limits_{s\in T} p(s)\prod\limits_{w\in d} p(w|s)^{n_{dw}}}$$

- Число параметров $|T| + |W|\cdot|T|$ не зависит от $|D|$

**Недостатки:**

- Наивное байесовское предположение о независимости слов:

$$p(w_1,\ldots,w_n|t) = p(w_1|t),\ldots,p(w_n|t)$$

# Обучение модели смеси униграмм

Принцип максимума правдоподобия:

$$\ln \prod_{d\in D} p(d) = \ln \prod_{d\in D} \sum_{t\in T} p(t) \prod_{w\in d} p(w|t)^{n_{dw}} \to \max_{\{p(t),p(w|t)\}};$$

$$\text{при} \quad \sum_{w\in W} p(w|t) = 1,\ t \in T; \quad \sum_{t\in T} p(t) = 1.$$

Лагранжиан:

$$\mathscr{L} = \sum_{d\in D} \ln \underbrace{\sum_{t\in T} p(t) \overbrace{\prod_{w\in d} p(w|t)^{n_{dw}}}^{p(d|t)}}_{p(d)} -$$

$$- \sum_{t\in T} \lambda_t \left( \sum_{w\in W} p(w|t) - 1 \right) - \mu \left( \sum_{t\in T} p(t) - 1 \right).$$

# Оценка МП для $p(t)$

$$\mathscr{L} = \sum_d \ln \underbrace{\sum_t p(t) \overbrace{\prod_{w \in d} p(w|t)^{n_{dw}}}^{p(d|t)}}_{p(d)} - \sum_t \lambda_t \left( \sum_w p(w|t) - 1 \right) - \mu \left( \sum_t p(t) - 1 \right).$$

$$\frac{\partial \mathscr{L}}{\partial p(t)} = \sum_{d \in D} \frac{1}{p(d)} p(d|t) - \mu = 0.$$

Умножим обе части равенства на $p(t)$ и просуммируем по $t$:

$$\mu \sum_{t \in T} p(t) = \sum_{d \in D} \sum_{t \in T} \frac{p(d|t)p(t)}{p(d)} \quad \Rightarrow \quad \mu = |D|.$$

Если умножить, но не суммировать:

$$\mu p(t) = \sum_{d \in D} \frac{p(d|t)p(t)}{p(d)} = \sum_{d \in D} p(t|d) \quad \Rightarrow \quad p(t) = \frac{\sum\limits_{d \in D} p(t|d)}{|D|}.$$

# Оценка МП для $p(w|t)$

$$\mathscr{L} = \sum_d \ln \underbrace{\sum_t p(t) \overbrace{\prod_{w \in d} p(w|t)^{n_{dw}}}^{p(d|t)}}_{p(d)} - \sum_t \lambda_t \left( \sum_w p(w|t) - 1 \right) - \mu \left( \sum_t p(t) - 1 \right).$$

$$\frac{\partial \mathscr{L}}{\partial p(w|t)} = \sum_{d \in D} \frac{1}{p(d)} n_{dw} \frac{p(d|t)}{p(w|t)} - \lambda_t = 0.$$

Умножим обе части равенства на $p(t)p(w|t)$, просуммируем по $w$:

$$\lambda_t p(t) \sum_w p(w|t) = \sum_{d,w} n_{dw} \frac{p(d|t)p(t)}{p(d)} \quad \Rightarrow \quad \lambda_t p(t) = \sum_{d,w} n_{dw} p(t|d).$$

Если умножить, но не суммировать:

$$\lambda_t p(t) p(w|t) = \sum_{d \in D} n_{dw} p(t|d) \quad \Rightarrow \quad p(w|t) = \frac{\sum\limits_{d \in D} n_{dw} p(t|d)}{\sum\limits_{d \in D} \sum\limits_{u \in W} n_{du} p(t|d)}.$$

# EM-алгоритм (оценки максимума правдоподобия)

- **Инициализировать** $p(t|d)$;
- **M-шаг:** оценить параметры модели

$$p(t) := \frac{\sum\limits_{d \in D} p(t|d)}{|D|} \quad \text{для всех } t \in T;$$

$$p(w|t) := \frac{\sum\limits_{d \in D} p(t|d) n_{dw}}{\sum\limits_{d \in D} p(t|d) \sum\limits_{u \in d} n_{du}} \quad \text{для всех } w \in W,\ d \in D;$$

- **E-шаг:** вычислить скрытые профили документов

$$p(t|d) := \frac{p(t) \prod\limits_{w \in d} p(w|t)^{n_{dw}}}{\sum\limits_{s \in T} p(s) \prod\limits_{w \in d} p(w|s)^{n_{dw}}} \quad \text{для всех } d \in D,\ t \in T;$$

# EM-алгоритм (байесовские оценки)

- **Инициализировать** $p(t|d)$;
- **M-шаг:** оценить параметры модели

$$p(t) := \frac{1 + \sum_{d \in D} p(t|d)}{|T| + |D|} \quad \text{для всех } t \in T;$$

$$p(w|t) := \frac{1 + \sum_{d \in D} p(t|d) n_{dw}}{|W| + \sum_{d \in D} p(t|d) \sum_{u \in d} n_{du}} \quad \text{для всех } w \in W,\ d \in D;$$

- **E-шаг:** вычислить скрытые профили документов

$$p(t|d) := \frac{p(t) \prod_{w \in d} p(w|t)^{n_{dw}}}{\sum_{s \in T} p(s) \prod_{w \in d} p(w|s)^{n_{dw}}} \quad \text{для всех } d \in D,\ t \in T;$$

# Инициализация $p(t|d)$ и частичное обучение

1. Использование априорной классификации документов
$y_{dt} = \big[$ документ $d$ относится к теме $t\big]$:

$$p(t|d) := \frac{y_{dt}}{\sum\limits_{s \in T} y_{ds}}.$$

2. Если размечена только часть коллекции $D^\ell \subset D$, то
— на первом M-шаге суммируем только по $d \in D^\ell$,
— на первом E-шаге вычисляем $p(t|d)$ для всех $d \in D$;
— далее EM-алгоритм выполняется как обычно,
на E-шаге вычисляем $p(t|d)$ для всех $d \in D$ либо $D \setminus D^\ell$.

3. Если $|D^\ell| \ll |D|$, то вес неразмеченных документов $\lambda < 1$:

$$\Lambda_d = \begin{cases} 1, & d \in D^\ell; \\ \lambda, & d \notin D^\ell. \end{cases}$$

# EM-алгоритм (с частичным обучением)

- **Инициализировать** $p(t|d)$ для всех $d \in D^\ell$, $t \in T$;
- **M-шаг:** оценить параметры модели

$$p(t) := \frac{1 + \sum_{d \in D} \Lambda_d p(t|d)}{|T| + |D|} \quad \text{для всех } t \in T;$$

$$p(w|t) := \frac{1 + \sum_{d \in D} \Lambda_d p(t|d) n_{dw}}{|W| + \sum_{d \in D} \Lambda_d p(t|d) \sum_{u \in d} n_{du}} \quad \text{для всех } w \in W, \ d \in D;$$

- **E-шаг:** вычислить скрытые профили документов

$$p(t|d) := \frac{p(t) \prod_{w \in d} p(w|t)^{n_{dw}}}{\sum_{s \in T} p(s) \prod_{w \in d} p(w|s)^{n_{dw}}} \quad \text{для всех } d \in D, \ t \in T;$$

# Замечание

В обзорах часто называется недостаток модели смеси униграмм: «каждый документ относится только к одной теме».

Это действительно так при использовании другой техники вывода формул M-шага (через неравенство Йенсена), однако мы получили те же формулы без этого предположения!

*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.
(имеется русский перевод)

# Вероятностный латентный семантический анализ
# Probabilistic Latent Semantic Analysis, PLSA [Hofmann, 1999]

- **Вероятностная модель документа:**

$$p(d) = \prod_{w \in d} p(d,w)^{n_{dw}}$$

- **Гипотеза условной независимости:** $p(w|t,d) = p(w|t)$ (распределения слов связаны с темами, а не с документами)
- **Модель смеси распределений для пар** $(d,w)$**:**

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

**Задача** максимизации правдоподобия по $p(t)$, $p(d|t)$, $p(w|t)$:

$$\mathscr{L} = \sum_{d \in D} \sum_{w \in W} n_{dw} \ln p(d,w) \to \max$$

$$\sum_{t \in T} p(t) = 1; \quad \sum_{d \in D} p(d|t) = 1; \quad \sum_{w \in W} p(w|t) = 1.$$

# Особенности, преимущества, недостатки модели PLSA

- Симметричность модели относительно $d \rightleftarrows w$:

$$\begin{aligned} p(d,w) &= \sum_{t\in T} p(t)\,p(d|t)\,p(w|t); \\ &= \sum_{t\in T} p(d)\,p(t|d)\,p(w|t); \\ &= \sum_{t\in T} p(w)\,p(t|w)\,p(d|t); \end{aligned}$$

- *Тематические профили* вычисляются по формуле Байеса:

$$p(t|d) = \frac{p(d|t)\,p(t)}{\sum_{s\in T} p(d|s)\,p(s)}; \qquad p(t|w) = \frac{p(w|t)\,p(t)}{\sum_{s\in T} p(w|s)\,p(s)}.$$

- Нет наивного байесовского предположения $p(w_1,\ldots,w_n|t) = p(w_1|t),\ldots,p(w_n|t)$
- Число параметров $|D||T| + |W||T|$, возможно переобучение!

# Максимизация правдоподобия: EM-алгоритм

Сформировать начальные приближения $p(t)$, $p(d|t)$, $p(w|t)$;
Повторять итерации до сходимости:

- **E-шаг:** скрытые переменные $H$ по формуле Байеса:

$$H(t|d,w) = \frac{p(t)p(d|t)p(w|t)}{p(d,w)};$$

- **M-шаг:** аналитическое решение задачи $\mathscr{L} \to \max$:

$$p(t) = \frac{S(t)}{S}; \quad S(t) = \sum_{d,w} n_{dw} H(t|d,w); \quad S = \sum_{d,w} n_{dw};$$

$$p(d|t) = \frac{1}{S(t)} \sum_{w} n_{dw} H(t|d,w);$$

$$p(w|t) = \frac{1}{S(t)} \sum_{d} n_{dw} H(t|d,w).$$

# Вывод формул M-шага

Распишем Лагранжиан:

$$\mathscr{L} = \sum_{d,w} n_{dw} \ln p(d,w) - \nu\left(\sum_{t\in T} p(t) - 1\right) -$$

$$- \sum_{t\in T} \lambda_t \left(\sum_{d\in D} p(d|t) - 1\right) - \sum_{t\in T} \mu_t \left(\sum_{w\in W} p(w|t) - 1\right).$$

$$\frac{\partial \mathscr{L}}{\partial p(t)} = \sum_{d,w} n_{dw} \frac{p(d|t)p(w|t)}{p(d,w)} - \nu = 0;$$

$$\sum_{d,w} n_{dw} \frac{p(d|t)p(w|t)p(t)}{p(d,w)} = \nu p(t) \;\Rightarrow\; \nu = \sum_{d,w} n_{dw} = S;$$

$$p(t) = \frac{1}{S} \sum_{d,w} n_{dw} H(t|d,w) = \frac{S(t)}{S};$$

# Вывод формул M-шага (продолжение)

$$\frac{\partial \mathscr{L}}{\partial p(d|t)} = \sum_w n_{dw} \frac{p(t)p(w|t)}{p(d,w)} - \lambda_t = 0;$$

$$\sum_w n_{dw} \frac{p(t)p(w|t)p(d|t)}{p(d,w)} = \lambda_t p(d|t) \;\Rightarrow\; \lambda_t = \sum_{d,w} n_{dw} H(t|d,w);$$

$$p(d|t) = \sum_w n_{dw} H(t|d,w) \Big/ \sum_{d,w} n_{dw} H(t|d,w).$$

$$\frac{\partial \mathscr{L}}{\partial p(w|t)} = \sum_d n_{dw} \frac{p(t)p(d|t)}{p(d,w)} - \mu_t = 0;$$

$$\sum_d n_{dw} \frac{p(t)p(d|t)p(w|t)}{p(d,w)} = \mu_t p(w|t) \;\Rightarrow\; \mu_t = \sum_{d,w} n_{dw} H(t|d,w);$$

$$p(w|t) = \sum_d n_{dw} H(t|d,w) \Big/ \sum_{d,w} n_{dw} H(t|d,w).$$

# Замечания о методе PLSA

1. Как быстро оценить профиль нового документа (folding-in):

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

где $p(w|d) = n_{dw}/n_d$ — оценка униграммной модели.

2. Меры по уменьшению переобучения:
   - обнуление незначимых компонент профилей $p(t|d)$, $p(t|w)$ (сокращение числа параметров модели);
   - обнуление незначимых скрытых переменных $H(t|d, w)$;
   - аккуратное формирование начального приближения;
   - ранний останов (оптимизация числа итераций);
   - симметризованный EM-алгоритм, в котором профили $p(t|w)$ и $p(t|d)$ уточняются по очереди.

# Латентное размещение Дирихле
# Latent Dirichlet Allocation [David Blei, 2003]

**Вероятностная тематическая модель:**

$$p(d) = \prod_{w \in d} p(d,w)^{n_{dw}}; \quad p(d,w) = \sum_{t \in T} p(w|t) \underbrace{p(t|d)p(d)}_{\theta_{dt}}.$$

Обозначим $\theta_{dt} = p(t|d)p(d)$.

**Гипотеза:** $\theta_d = (\theta_{dt})_{t \in T} \in \mathbb{R}^{|T|}$ — случайные векторы из априорного распределения Дирихле с параметром $\alpha \in \mathbb{R}^{|T|}$:

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_t \alpha_t\right)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_t^{\alpha_t - 1}.$$

Тогда тематическая модель примет вид:

$$p(d) = \int p(\theta_d|\alpha) \prod_{w \in d} \left( \sum_{t \in T} p(w|t)\theta_{dt} \right)^{n_{dw}} d\theta_d$$

# Оценивание параметров и вывод профилей

- Задача оценивания параметров в модели LDA:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(d,w|\alpha,\beta) \to \max_{\alpha,\beta},$$

где $\alpha$ — параметр Дирихле, $\beta = \big(p(w|t)\big)_{|W|\times|T|}$.

- Тематический профиль документа $d$:

$$p(t|d) = \frac{p(t,d)}{p(d)} = \frac{\int p(\theta_d|\alpha) \prod_{w\in d} \Big(p(w|t)\theta_{dt}\Big)^{n_{dw}} d\theta_d}{\int p(\theta_d|\alpha) \prod_{w\in d} \Big(\sum_{t\in T} p(w|t)\theta_{dt}\Big)^{n_{dw}} d\theta_d}.$$

Методы решения:

- Самплирование Гиббса (Gibbs sampling).
- Вариационный вывод (variational inference).

# Особенности, преимущества, недостатки LDA

- Модель порождения документов $d$ общая для всей коллекции $D$, а не отдельная для каждого $d \in D$.
- Число параметров $|T| + |T| \cdot |W|$ не зависит от $|D|$.
- Профиль нового документа $p(t|d)$ оценивается по той же модели, что и для всех документов обучающей коллекции.
- Оценивание параметров модели и профилей $p(t|d)$ — сложная вычислительная задача.

*David Blei, Andrew Ng, Michael Jordan.*
Latent Dirichlet allocation.
Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.

**Эффективная реализация — в проекте Vowpal Wabbit:**
`http://hunch.net/~vw/`

# Методы оценивания качества тематических моделей

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

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

# Обобщения и модификации тематических моделей

- Иерархические модели, в том числе с адаптивной иерархией
- Темпоральные модели, учитывающие годы публикаций
- Author-topic models — пытаются приписать распределение авторов $p(a|w)$ каждому слову документа
- Entity-topic models — оценивают тематику авторов, журналов, конференций, организаций, стран
- Модели, учитывающие связь слов внутри документа
- Модели, связи 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.
(имеется русский перевод)

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