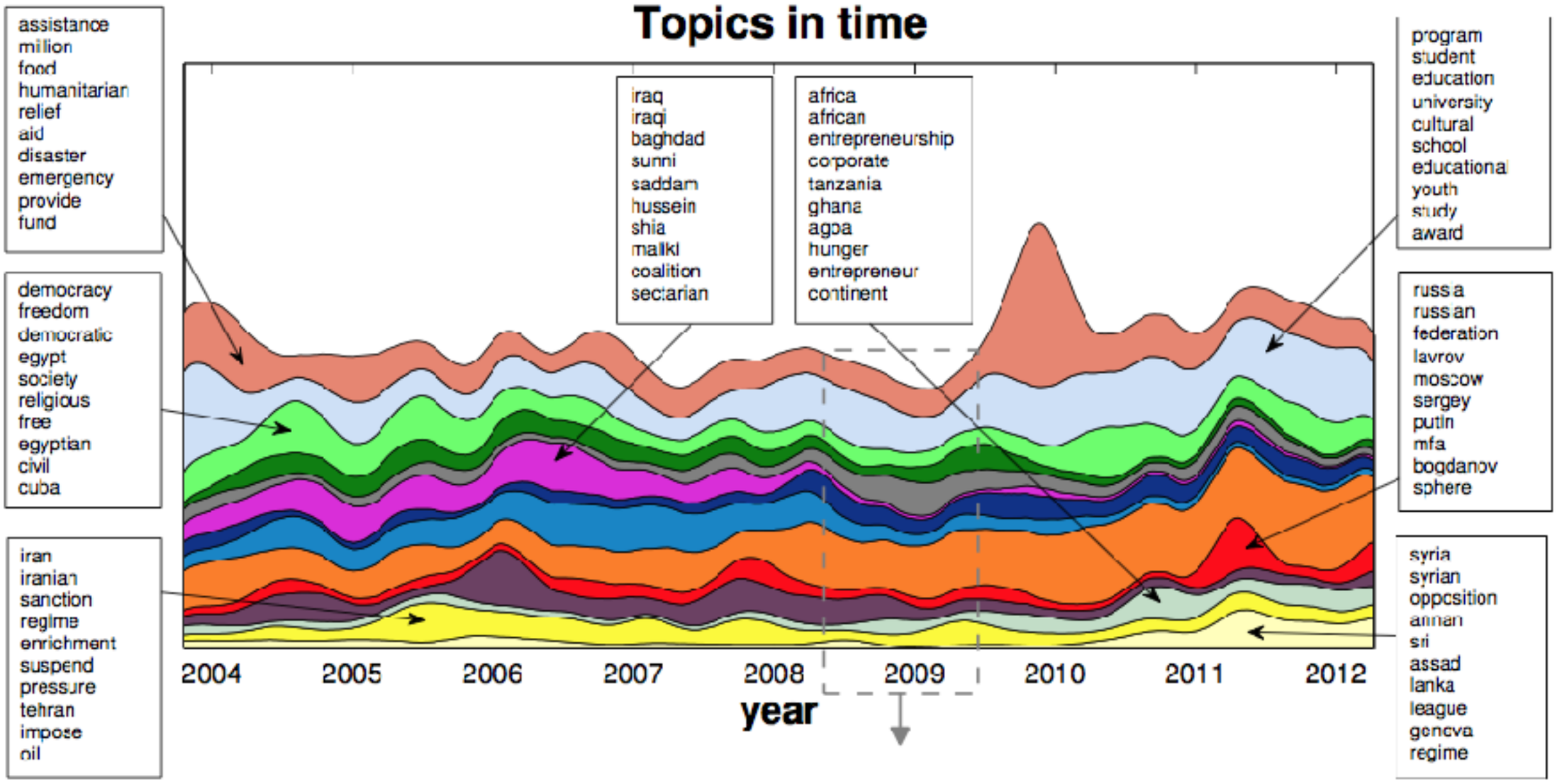
Введение в анализ текстов

Тематическое моделирование

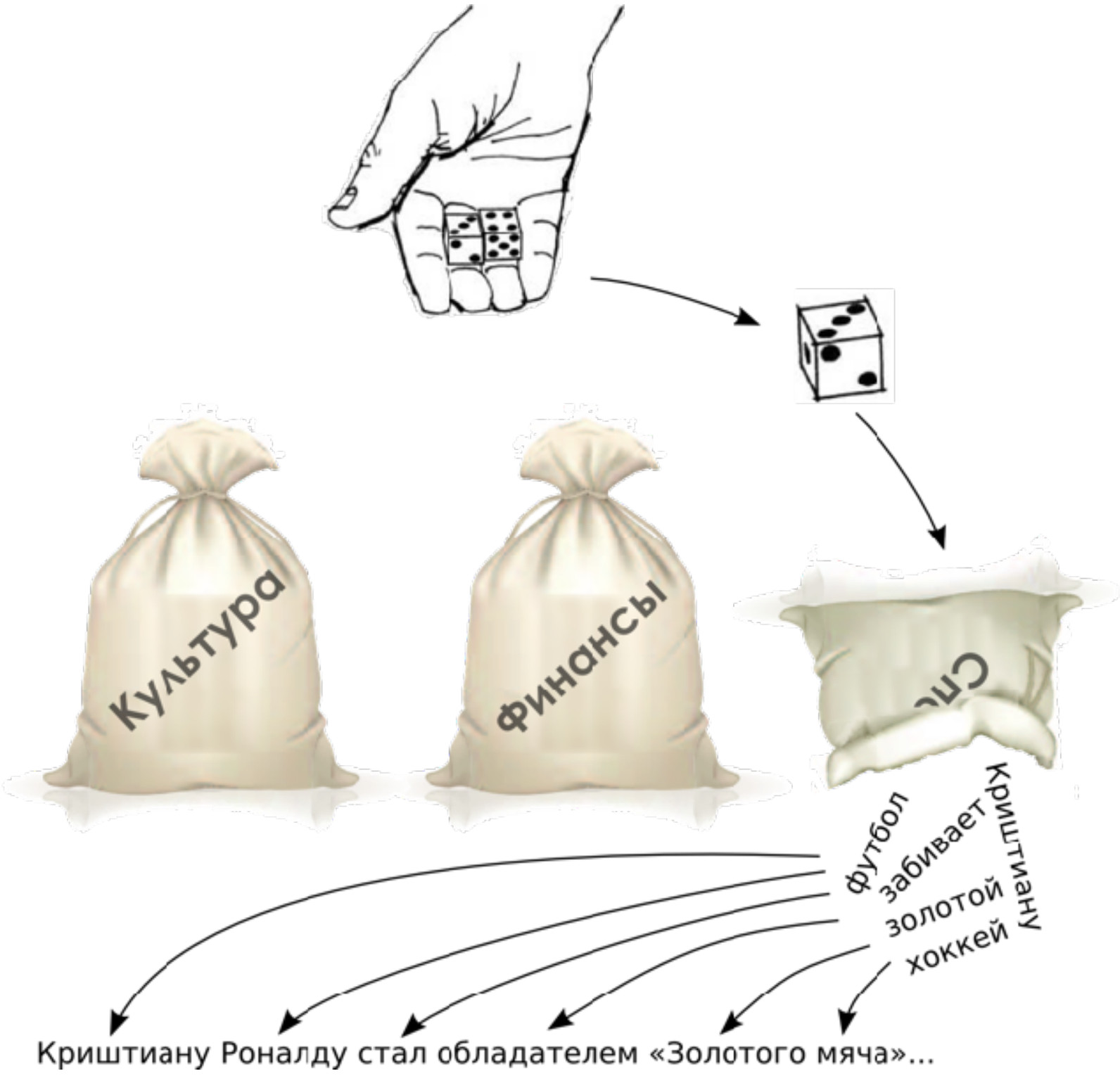
Какие ещё темы?

Коллекция внешнеполитических пресс-релизов ряда стран:
20 тыс. сообщений, 10 лет, 180Мб текста, английский язык.



Отбросим предрассудки

Отбросим предрассудки

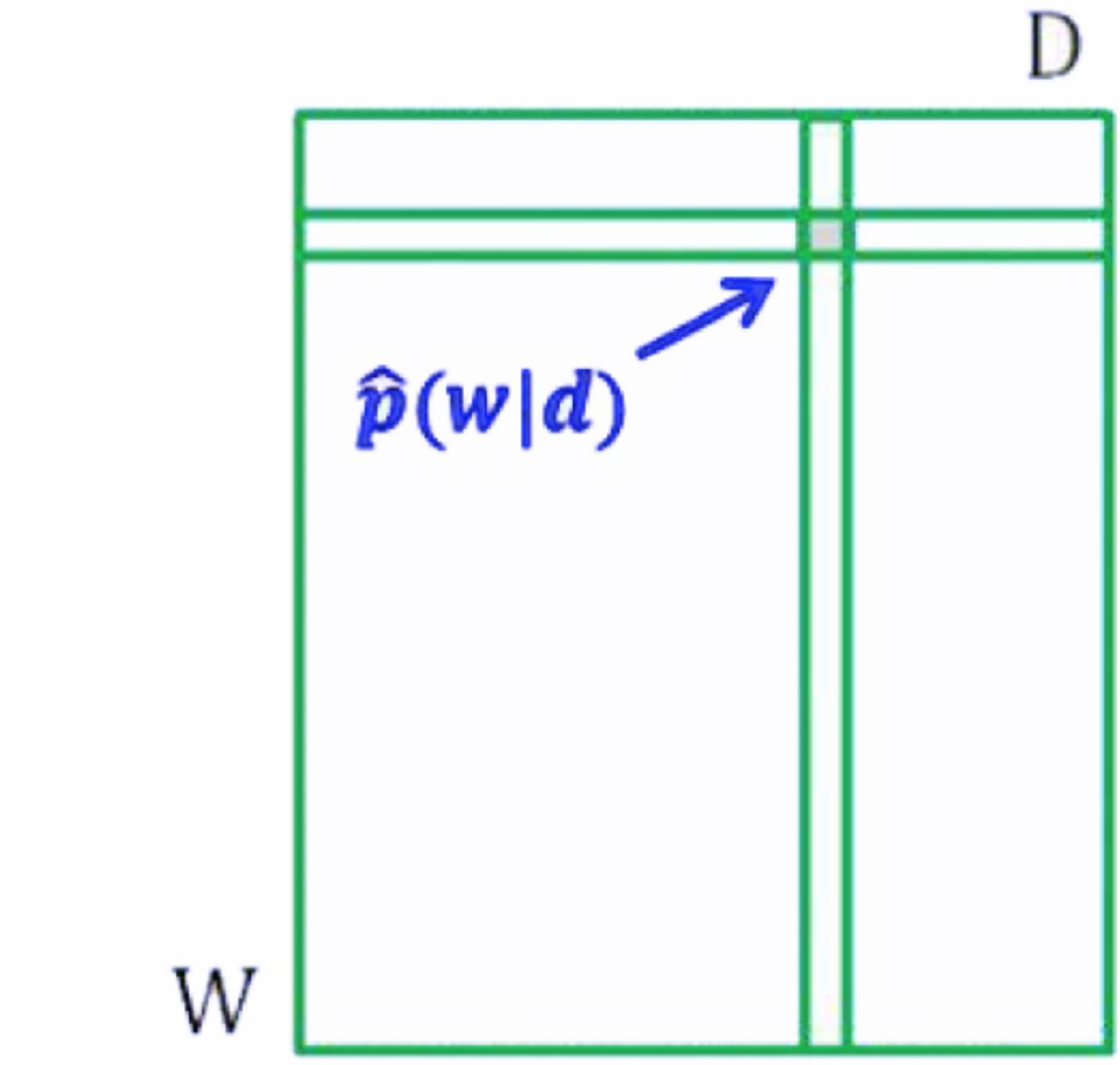


Добавим формальностей

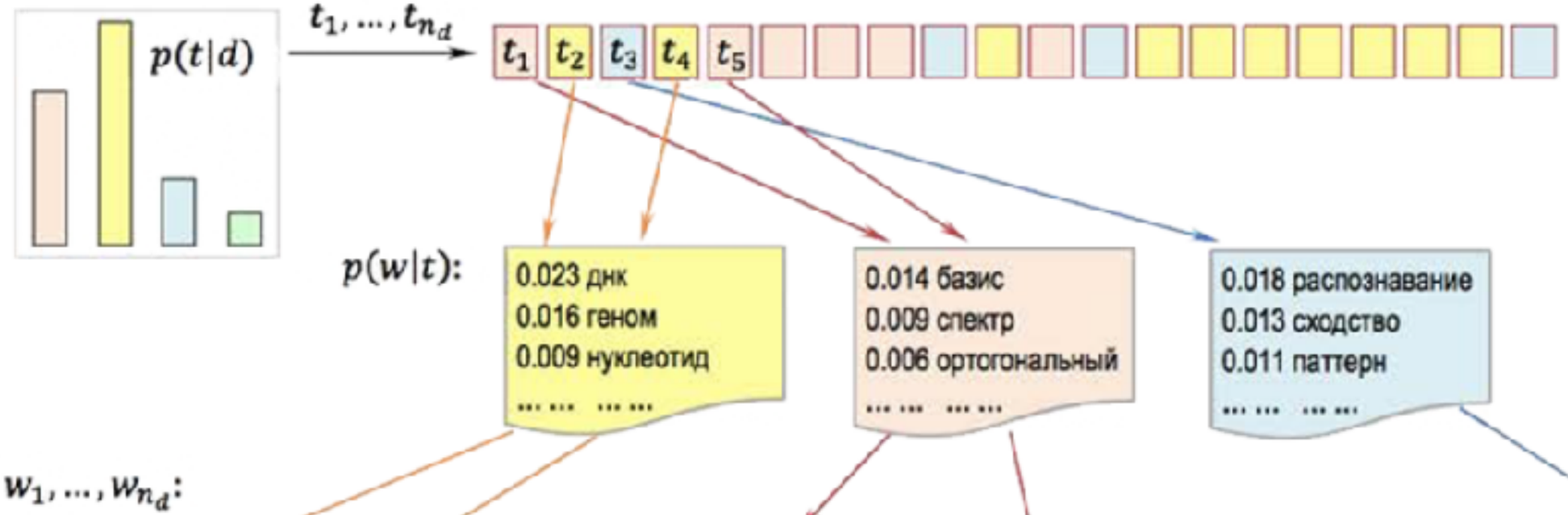
- D - множество документов, которое мы собрали;
- d - отдельный документ;
- T - множество тем, которые задумала природа;
- t - отдельная тема;
- w - отдельное слово, документы и темы состоят из слов.

Каждый документ - это мешок слов. Для каждого слова мы можем посчитать с какой частотой оно входит в документ. Эта частота будет оценкой вероятности встретить в документе конкретное слово, $p(w|d)$.

Матрица термы на документы



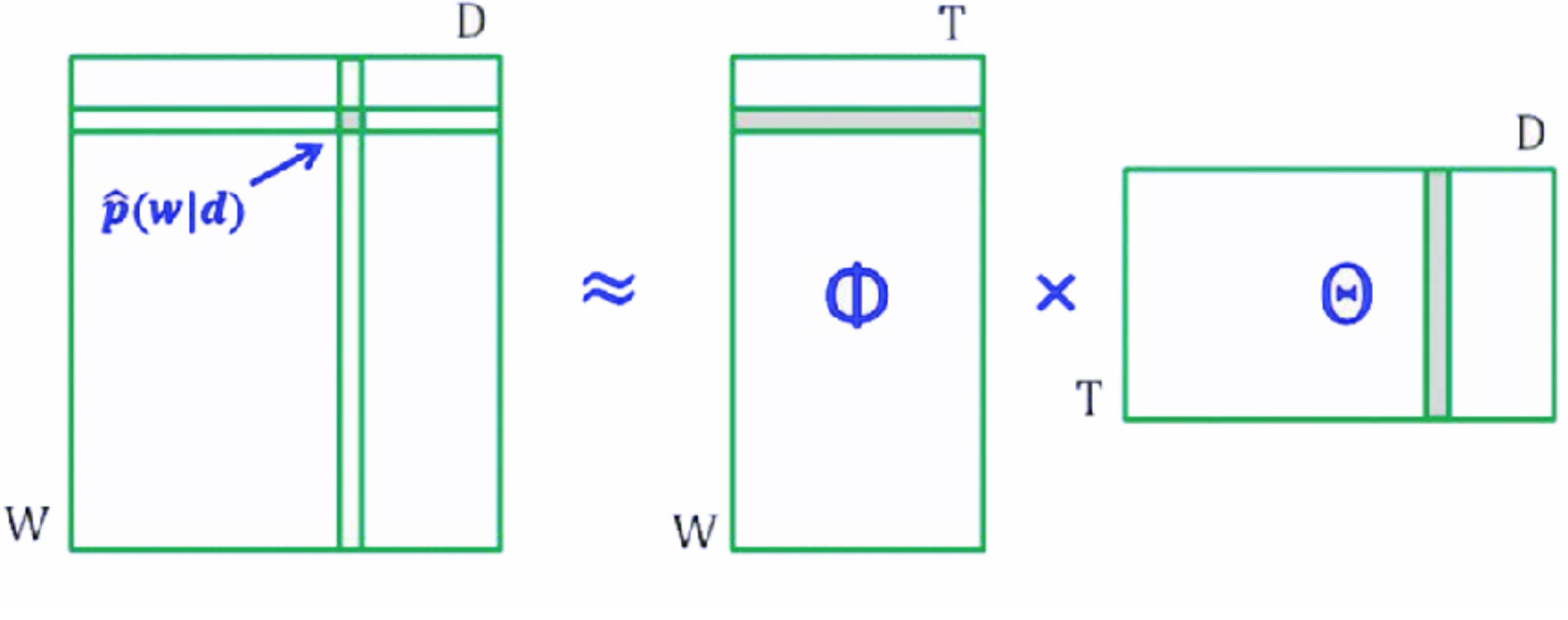
Рождение текста



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

Вероятность термина

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



Метод максимального правдоподобия

Вся наша выборка состоит из текстов $d_1, \ldots, d_{|D|}$. Правдоподобие для неё составит

$$L = p(d_1) \cdot \ldots \cdot p(d_{|D|}) = \prod_{d \in D} p(d).$$

Метод максимального правдоподобия

Каждый текст складывается из конкретных слов. Одно и то же слово может встречаться в тексте несколько раз. Будем считать, что слово w в документе d встретилось n_{wd} раз. Вероятность получить первый текст составит

$$p(d_1) = \prod_{w \in d_1} p(w \mid d_1)^{n_{wd_1}}$$

$$p(d_1) = \prod_{w \in d_1} \left(\sum_{t \in T} p(w \mid t) \cdot p(t \mid d_1) \right)^{n_{wd_1}}$$

Метод максимального правдоподобия

Каждый текст складывается из конкретных слов. Одно и то же слово может встречаться в тексте несколько раз. Будем считать, что слово w в документе d встретилось n_{wd} раз. Вероятность получить первый текст составит

$$p(d_1) = \prod_{w \in d_1} p(w \mid d_1)^{n_{wd_1}}$$

$$p(d_1) = \prod_{w \in d_1} \left(\sum_{t \in T} p(w \mid t) \cdot p(t \mid d_1) \right)^{n_{wd_1}}$$

Метод максимального правдоподобия

$$L = \prod_{d \in D} \prod_{w \in d} \left(\sum_{t \in T} p(w \mid t) \cdot p(t \mid d) \right)^{n_{wd}} = \prod_{(d,w)} \left(\sum_{t \in T} p(w \mid t) \cdot p(t \mid d) \right)^{n_{wd}}$$

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

EM-алгоритм

- Инициализировали матрицы
- E-шаг:

$$H_{dwt} = p(t \mid d, w) = \frac{p(w \mid t)p(t \mid d)}{p(w \mid d)} = \frac{\varphi_{wt}\theta_{td}}{\sum_{s \in T} \varphi_{ws}\theta_{sd}}.$$

- M-шаг:

$$\hat{n}_{dwt} = n_{dw}p(t \mid d, w) = n_{dw}H_{dwt}$$

$$\varphi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}; \qquad \hat{n}_t = \sum_{w \in W} \hat{n}_{wt}; \qquad \hat{n}_{wt} = \sum_{d \in D} n_{dw}H_{dwt}.$$

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

Резюме

Дано: W — словарь терминов (слов или словосочетаний),
D — коллекция текстовых документов $d \subset W$,
n_{dw} — сколько раз термин w встретился в документе d.

Найти: модель $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$ с параметрами $\underset{W \times T}{\Phi}$ и $\underset{T \times D}{\Theta}$:
$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t,
$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d.

Критерий максимума логарифма правдоподобия:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\phi,\theta};$$

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

Проблема: задача стохастического матричного разложения *некорректно поставлена*: $\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$.

ARTM

Максимизация ln правдоподобия с регуляризатором R:

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

Модель PLSA: $R(\Phi,\Theta) = 0$

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

Регуляризаторы

- Разреживающий: хотим в итоговых матрицах много нулей. Чем сильнее разрежено распределение, тем ниже его энтропия. Максимизируем KL-дивиргенцию между нашими распределениями и равномерным.

$$R(\Phi, \Theta) = -\beta \sum_{t \in T} \sum_{w \in W} \ln \varphi_{wt} - \alpha \sum_{d \in D} \sum_{t \in T} \ln \theta_{td} \to \max .$$

Регуляризаторы

- Декоррелирование: хотим, чтобы темы были по своему составу как можно различнее

$$R(\Phi, \Theta) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \operatorname{cov}(\varphi_t, \varphi_s) \to \max,$$