

ВЕРОЯТНОСТНЫЕ ТЕМАТИЧЕСКИЕ МОДЕЛИ

ЦЕЛИ И ЗАДАЧИ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

ЧТО ТАКОЕ “ТЕМА”?

- *Тема* — специальная терминология предметной области
- *Тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах

ЧТО ТАКОЕ “ТЕМА”?

Более формально:

- » *Тема* — условное распределение на множестве терминов, $p(w|t)$ — вероятность (частота) термина w в теме t
- » *Тематика документа* — условное распределение, $p(t|d)$ — вероятность (частота) темы t в документе d

ЧТО ТАКОЕ “ТЕМА”?

- › Тема — условное распределение на множестве терминов, $p(w|t)$ — вероятность (частота) термина w в теме t
- › Тематика документа — условное распределение, $p(t|d)$ — вероятность (частота) темы t в документе d
- › Тематическая модель автоматически выявляет латентные темы по наблюдаемым частотам терминов в документах $p(w|d)$

ЦЕЛИ И ЗАДАЧИ

- Цель — автоматизация анализа текстов
- Задачи:
 - ▶ классификация и категоризация документов
 - ▶ автоматическое аннотирование документов
 - ▶ автоматическая суммаризация коллекций

ЦЕЛИ И ЗАДАЧИ

- Цель — автоматизация анализа текстов
- Задачи:
 - ▶ автоматическое аннотирование документов
 - ▶ автоматическая суммаризация коллекций
 - ▶ тематическая сегментация документов

ЦЕЛИ И ЗАДАЧИ

- Цель — автоматизация анализа текстов
- Идея решения: использовать признаковые описания документов $p(t|d)$

ЦЕЛИ И ЗАДАЧИ

- Цель — систематизация больших объёмов информации, помочь человеку в поиске и понимании текстовой информации, устранение барьеров между Человеком и Знанием
- Семантический (разведочный) поиск информации
- Визуализация тематической структуры коллекции

ЦЕЛИ И ЗАДАЧИ

- Семантический (разведочный) поиск информации
- Визуализация тематической структуры коллекции
- Анализ динамики развития тем
- Тематический мониторинг новых поступлений

ЦЕЛИ И ЗАДАЧИ

- Визуализация тематической структуры коллекции
- Анализ динамики развития тем
- Тематический мониторинг новых поступлений
- Рекомендации документов пользователям

ПРИЛОЖЕНИЯ

- › Поиск научной информации, трендов, фронта исследований
- › Подбор экспертов, рецензентов, исполнителей проектов
- › Агрегирование новостных потоков
- › Создание рекомендательных сервисов

ПРИЛОЖЕНИЯ

- › Агрегирование новостных потоков
- › Создание рекомендательных сервисов
- › Аннотирование и поиск изображений
- › Анализ видеопоследовательностей
- › Аннотация генома и другие задачи биоинформатики

ПРИЛОЖЕНИЯ

- Анализ видеопоследовательностей
- Аннотация генома и другие задачи биоинформатики
- Анализ дискретизированных биомедицинских сигналов
- Мониторинг состояния технических систем

МУЛЬТИЯЗЫЧНАЯ МОДЕЛЬ ВИКИПЕДИИ

- › 216175 русско-английских пар статей
Википедии, 400 тем. Первые 10 слов с их
вероятностями $p(w|t)$ в %:

тема 68		тема 79	
research	4.56	институт	6.03
technology	3.14	университет	3.35
engineering	2.63	программа	3.17
institute	2.37	учебный	2.75
science	1.97	технический	2.70
program	1.60	технология	2.30
education	1.44	научный	1.76
campus	1.43	исследование	1.67
management	1.38	наука	1.64
programs	1.36	образование	1.47
goals	4.48	матч	6.02
league	3.99	игрок	5.56
club	3.76	сборная	4.51
season	3.49	фк	3.25
scored	2.72	против	3.20
cup	2.57	клуб	3.14
goal	2.48	футболист	2.67
apps	1.74	гол	2.65
debut	1.69	забивать	2.53
match	1.67	команда	2.14

МУЛЬТИЯЗЫЧНАЯ МОДЕЛЬ ВИКИПЕДИИ

тема 68		тема 79	
research	4.56	институт	6.03
technology	3.14	университет	3.35
engineering	2.63	программа	3.17
institute	2.37	учебный	2.75
science	1.97	технический	2.70
program	1.60	технология	2.30
education	1.44	научный	1.76
campus	1.43	исследование	1.67
management	1.38	наука	1.64
programs	1.36	образование	1.47
goals	4.48	матч	6.02
league	3.99	игрок	5.56
club	3.76	сборная	4.51
season	3.49	фк	3.25
scored	2.72	против	3.20
cup	2.57	клуб	3.14
goal	2.48	футболист	2.67
apps	1.74	гол	2.65
debut	1.69	забивать	2.53
match	1.67	команда	2.14

- Модель выявляет двуязычные темы без выравнивания, без словарей, даже когда тексты не являются точными переводами

МУЛЬТИЯЗЫЧНАЯ МОДЕЛЬ ВИКИПЕДИИ

тема 88		тема 251	
opera	7.36	опера	7.82
conductor	1.69	оперный	3.13
orchestra	1.14	дирижер	2.82
wagner	0.97	певец	1.65
soprano	0.78	певица	1.51
performance	0.78	театр	1.14
mozart	0.74	партия	1.05
sang	0.70	сопрано	0.97
singing	0.69	вагнер	0.90
operas	0.68	оркестр	0.82
		windows	8.00
		microsoft	4.03
		server	2.93
		software	1.38
		user	1.03
		security	0.92
		mitchell	0.82
		oracle	0.82
		enterprise	0.78
		users	0.78
		windows	6.05
		microsoft	3.76
		версия	1.86
		приложение	1.86
		сервер	1.63
		server	1.54
		программный	1.08
		пользователь	1.04
		обеспечение	1.02
		система	0.96

- ▶ В этом эксперименте независимый ассессор оценил 396 тем из 400 как хорошо интерпретируемые

БИГРАММАЯ МОДЕЛЬ: ТЕРМИНЫ - СЛОВОСОЧЕТАНИЯ

➤ Коллекция 850 статей конференций ММРО,
ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информационность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информационность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

ПОИСК ЭТНО-РЕЛЕВАНТНЫХ ТЕМ В СОЦИАЛЬНЫХ СЕТЯХ

- Цель: поддержка социологических исследований в области межэтнических отношений
- Задача: создание системы *разведочного поиска*, систематизации и мониторинга этно-релевантных тем
- Запрос: словарь этнонимов (более 800 слов)

ПОИСК ЭТНО-РЕЛЕВАНТНЫХ ТЕМ В СОЦИАЛЬНЫХ СЕТЯХ

- **Задача:** создание системы *разведочного поиска*, систематизации и мониторинга этно-релевантных тем
- **Запрос:** словарь этнонимов (более 800 слов)
- **Поисковая выдача:** этно-релевантные темы, списки постов по каждой теме

ПРИМЕРЫ ЭТНО-РЕЛЕВАНТНЫХ ТЕМ

- › (русские): русский, князь, россия, татарин, царить, иван, империя, государь, екатерина
- › (русские): акция, организация, митинг, движение, русский, пикет, москва, оппозиция
- › (сирийцы): асад, боевик, район, террорист, уничтожать, дамаск, оружие, группировка
- › (таджики, узбеки): мигрант, страна, россия, азия, нелегальный, таджикистан, миграция

ПРИМЕРЫ ЭТНО-РЕЛЕВАНТНЫХ ТЕМ

- › (таджики, узбеки): мигрант, страна, россия, азия, нелегальный, таджикистан, миграция
- › (канадцы): команда, игра, канадский, сезон, хоккей, сборная, играть, победа, счет, кубок
- › (норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, опека
- › (китайцы): китайский, производство, страна, продукция, технология, компания, военный

РЕЗЮМЕ

- Тематическое моделирование — современное направление статистического анализа коллекций текстовых документов
- Тематическая модель автоматически:
 - ▶ определяет состав тем в каждом документе
 - ▶ определяет состав терминов в каждой теме

РЕЗЮМЕ

- Тематическая модель автоматически:
 - ▶ определяет состав тем в каждом документе
 - ▶ определяет состав терминов в каждой теме
- Существуют сотни моделей для учёта метаданных, связей пользователей, внешних лингвистических ресурсов и т.п.

ПОРОЖДАЮЩАЯ МОДЕЛЬ И ПОСТАНОВКА ЗАДАЧИ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

ПОДГОТОВКА ДАННЫХ ДЛЯ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

- Предварительная обработка и очистка текстов:
 - ▶ Удаление форматирования и переносов
 - ▶ Удаление обрывочной и нетекстовой информации
 - ▶ Исправление опечаток
 - ▶ Слияние слишком коротких текстов

ПОДГОТОВКА ДАННЫХ ДЛЯ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

➤ Формирование словаря:

- ▶ Приведение слов к нормальной форме
(лемматизация или стемминг)
- ▶ Выделение терминов (term extraction)
- ▶ Удаление стоп-слов и слишком редких
слов

БАЗОВЫЕ ПРЕДПОЛОЖЕНИЯ ПРОСТЫХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

- Базовые предположения:
 - ▶ Порядок документов в коллекции не важен
 - ▶ Порядок терминов в документе не важен
(bag of words)
 - ▶ Каждая пара (d, w) связана с некоторой темой $t \in T$

БАЗОВЫЕ ПРЕДПОЛОЖЕНИЯ ПРОСТЫХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

➤ Базовые предположения:

- ▶ Порядок терминов в документе не важен (bag of words)
- ▶ Каждая пара (d, w) связана с некоторой темой $t \in T$
- ▶ Гипотеза условной независимости:

$$p(w|t, d) = p(w|t)$$

БАЗОВЫЕ ПРЕДПОЛОЖЕНИЯ ПРОСТЫХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

- Дополнительные предположения о разреженности:
 - ▶ Документ, как правило, относится к небольшому числу тем
 - ▶ Тема, как правило, состоит из небольшого числа терминов

ВЕРОЯТНОСТНЫЙ ПРОЦЕСС ПОРОЖДЕНИЯ ТЕКСТОВОЙ КОЛЛЕКЦИИ

- › Документ d — это смесь распределений $p(w|t)$ с весами $p(t|d)$:

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



ПОСТАНОВКА ЗАДАЧИ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

- › **Дано:** W — словарь терминов (слов или словосочетаний)
 D — коллекция текстовых документов $d \in W$
 n_{dw} — сколько раз термин w встретился в документе d
 n_d — длина документа d
- › **Найти:** параметры вероятностной тематической модели

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

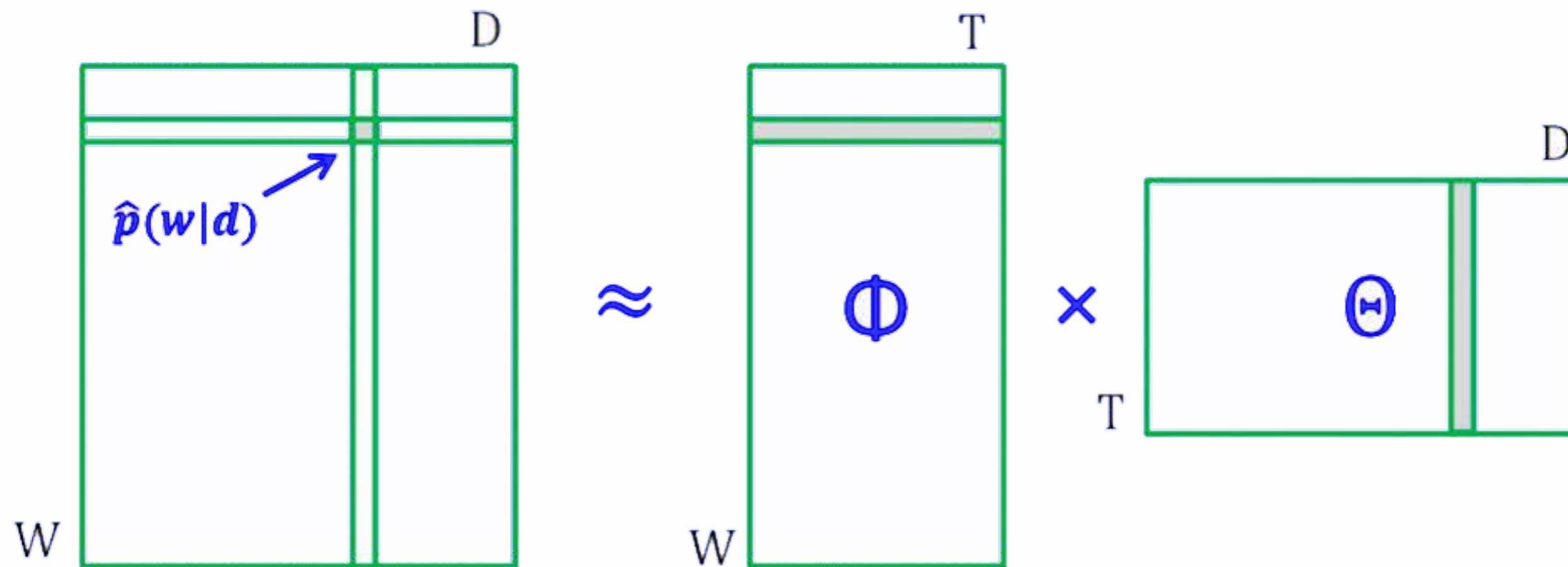
ПОСТАНОВКА ЗАДАЧИ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

- Найти: параметры вероятностной тематической модели

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

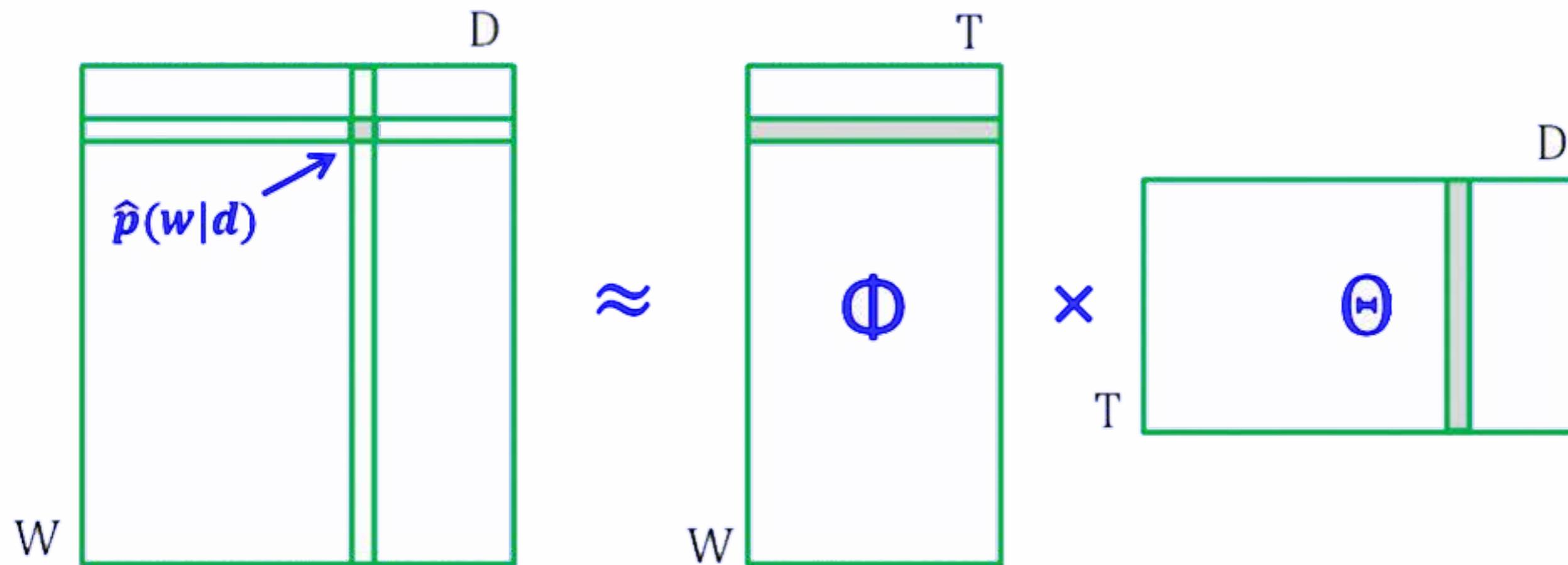
- $\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

ТЕМАТИЧЕСКАЯ МОДЕЛЬ КАК МАТРИЧНОЕ РАЗЛОЖЕНИЕ



- » $\Phi = (\phi_{wt})$ — матрица распределений терминов в темах
- » $\Theta = (\theta_{td})$ — матрица распределений тем в документах

ТЕМАТИЧЕСКАЯ МОДЕЛЬ КАК МАТРИЧНОЕ РАЗЛОЖЕНИЕ



- Наблюдаемые частоты терминов в документах:

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

- Вероятностная тематическая модель:

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

ПРИНЦИП МАКСИМУМА ПРАВДОПОДОБИЯ

- » Для построения тематической модели:

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

- » По наблюдаемым частотам:

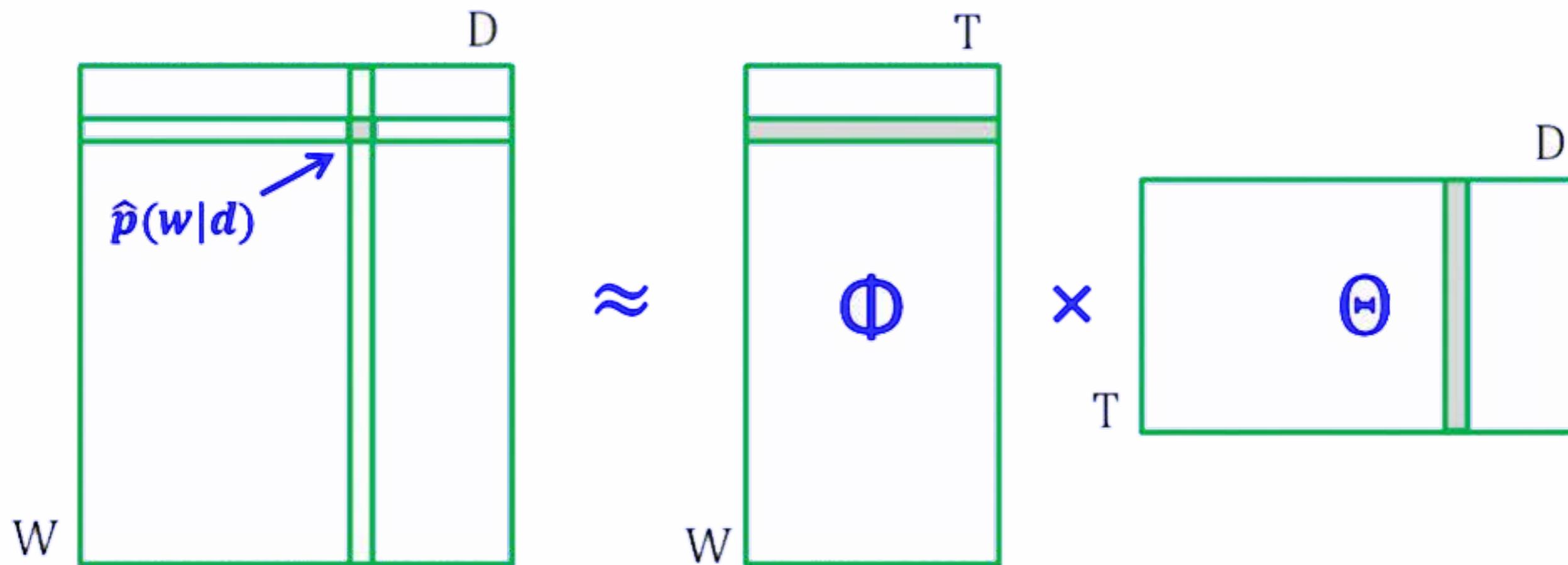
$$\hat{p}(w|d) = \frac{n_{dw}}{n_d}$$

ПРИНЦИП МАКСИМУМА ПРАВДОПОДОБИЯ

- » $p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$
- » $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$
- » Максимизируется логарифм правдоподобия:

$$\left\{ \begin{array}{l} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta} \\ \sum_{w \in W} \phi_{wt} = 1 \quad \phi_{wt} \geq 0 \\ \sum_{t \in T} \theta_{td} = 1 \quad \theta_{td} \geq 0 \end{array} \right.$$

ПРИНЦИП МАКСИМУМА ПРАВДОПОДОБИЯ



- Задача матричного разложения некорректно поставлена, поскольку её решение в общем случае не единственno:

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

ПРИНЦИП МАКСИМУМА РЕГУЛЯРИЗОВАННОГО ПРАВДОПОДОБИЯ

› Чтобы из множества решений выбрать наиболее подходящее, вводится критерий регуляризации $R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$

$$\left\{ \begin{array}{l} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta} \\ \sum_{w \in W} \phi_{wt} = 1 \quad \phi_{wt} \geq 0 \\ \sum_{t \in T} \theta_{td} = 1 \quad \theta_{td} \geq 0 \end{array} \right.$$

РЕЗЮМЕ

- *Вероятностная тематическая модель* описывает:
 - ▶ Каждый документ вероятностной смесью тем
 - ▶ Каждую тему — распределением терминов в темах
- Два основных вероятностных допущения:

РЕЗЮМЕ

- *Вероятностная тематическая модель* описывает:
 - ▶ Каждый документ вероятностной смесью тем
 - ▶ Каждую тему — распределением терминов в темах
- Два основных вероятностных допущения:
 - ▶ Гипотеза “мешка слов” (или терминов)
 - ▶ Гипотеза условной независимости

РЕЗЮМЕ

- Два основных вероятностных допущения:
 - ▶ Гипотеза “мешка слов” (или терминов)
 - ▶ Гипотеза условной независимости
- Трактовки постановки задачи:
 - ▶ “Мягкая” би-кластеризация документов и терминов
 - ▶ Стохастическое матричное разложение

РЕЗЮМЕ

➤ Трактовки постановки задачи:

- ▶ “Мягкая” би-кластеризация документов и терминов
- ▶ Стохастическое матричное разложение
- ▶ Регуляризация некорректно поставленной задачи

БАЗОВЫЕ ТЕМАТИЧЕСКИЕ МОДЕЛИ И ЕМ-АЛГОРИТМЫ

ПРИНЦИП МАКСИМУМА РЕГУЛЯРИЗОВАННОГО ПРАВДОПОДОБИЯ

› Чтобы из множества решений выбрать наиболее подходящее, вводится критерий регуляризации $R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$

$$\left\{ \begin{array}{l} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta} \\ \sum_{w \in W} \phi_{wt} = 1 \quad \phi_{wt} \geq 0 \\ \sum_{t \in T} \theta_{td} = 1 \quad \theta_{td} \geq 0 \end{array} \right.$$

РЕГУЛЯРИЗОВАННЫЙ ЕМ-АЛГОРИТМ

- ЕМ-алгоритм: метод простой итерации для системы уравнений относительно переменных ϕ_{wt} , θ_{td} и $p_{tdw} = p(t|d, w)$

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

РЕГУЛЯРИЗОВАННЫЙ ЕМ-АЛГОРИТМ

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

› Операция нормировки вектора:

$$\underset{t \in T}{\text{norm}}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$$

ДВА САМЫХ ИЗВЕСТНЫХ ЧАСТНЫХ СЛУЧАЯ

- › PLSA, вероятностный латентный семантический анализ:

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

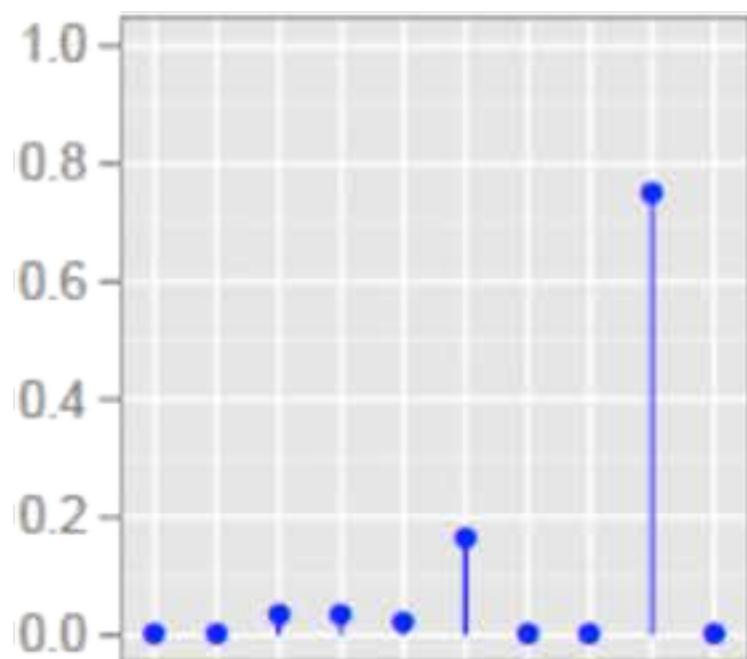
- › LDA, латентное размещение Дирихле:

$$R(\Phi, \Theta) = \sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td}$$

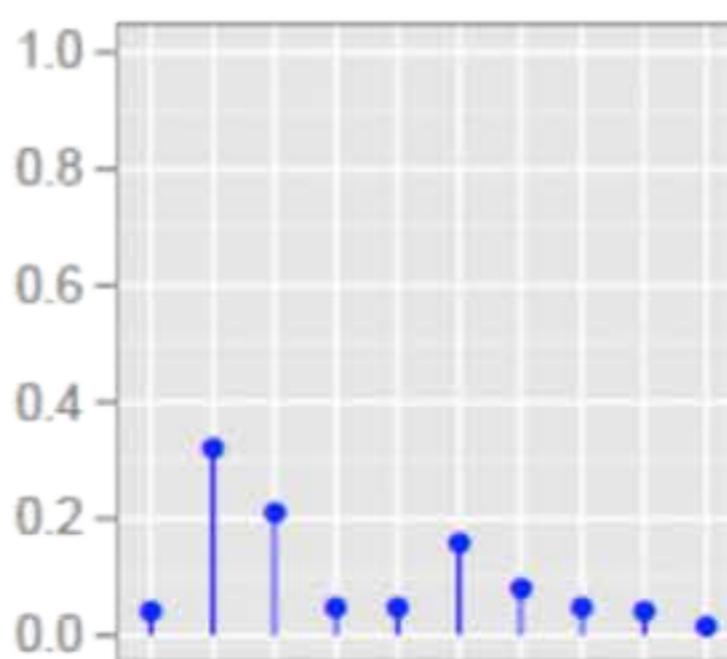
где $\beta_w > 0$, $\alpha_t > 0$ — параметры регуляризатора

БАЙЕСОВСКАЯ ИНТЕРПРЕТАЦИЯ МОДЕЛИ LDA

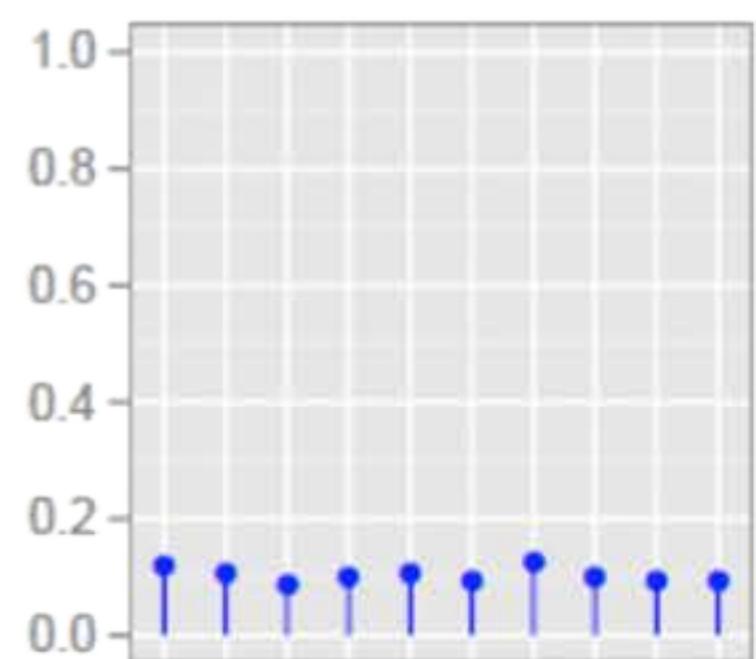
- › Столбцы ϕ_t матрицы Φ порождаются $|W|$ -мерным распределением Дирихле с вектором параметров $\beta = (\beta_w)$
- › Пример: распределение $\phi \sim \text{Dir}(\beta)$ при $|W| = 10$, $\phi, \beta \in \mathbb{R}^{10}$



$\beta_w = 0.1$



$\beta_w = 1$
(равномерное)

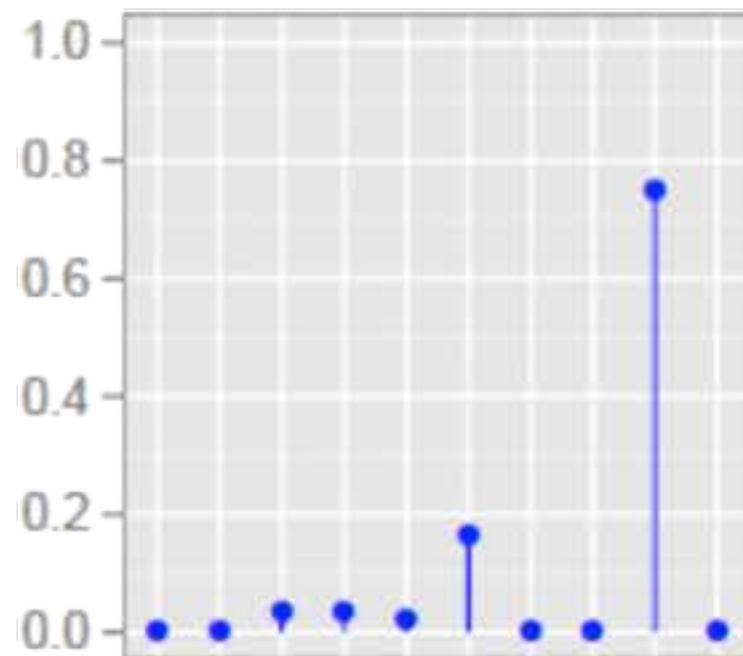


$\beta_w = 100$

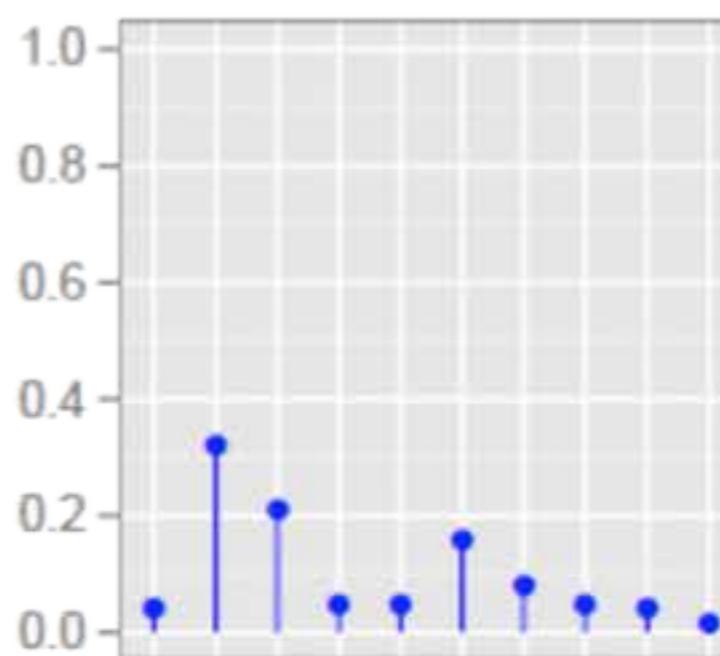
БАЙЕСОВСКАЯ ИНТЕРПРЕТАЦИЯ МОДЕЛИ LDA

› Регуляризатор максимизирует апостериорную вероятность

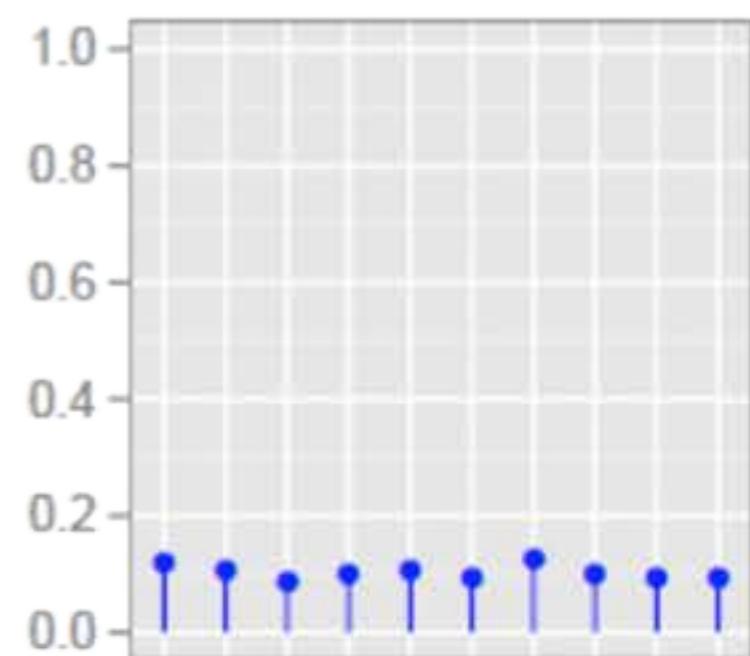
› Пример: распределение $\phi \sim \text{Dir}(\beta)$ при $|W| = 10$, $\phi, \beta \in \mathbb{R}^{10}$



$$\beta_w = 0.1$$



$\beta_w = 1$
(равномерное)



$$\beta_w = 100$$

ДИВЕРГЕНЦИЯ КУЛЬБАКА-ЛЕЙБЛЕРА

➤ Расстояние между распределениями

$$P = (p_i)_{i=1}^n \text{ и } Q = (q_i)_{i=1}^n:$$

$$\text{KL}(P\|Q) = \sum_i p_i \log \frac{p_i}{q_i}$$

➤ Неотрицательность: $\text{KL}(P\|Q) \geq 0$

$$\text{KL}(P\|Q) = 0 \Leftrightarrow P = Q$$

➤ Несимметричность: $\text{KL}(P\|Q) \neq \text{KL}(Q\|P)$

ДИВЕРГЕНЦИЯ КУЛЬБАКА-ЛЕЙБЛЕРА

➤ Расстояние между распределениями

$$P = (p_i)_{i=1}^n \text{ и } Q = (q_i)_{i=1}^n:$$

$$\text{KL}(P||Q) = \sum_i p_i \log \frac{p_i}{q_i}$$

➤ Связь с принципом максимума правдоподобия:

$$\sum_{i=1}^n p_i \ln \frac{p_i}{q_i(\alpha)} \rightarrow \min_{\alpha} \Leftrightarrow \sum_{i=1}^n p_i \ln q_i(\alpha) \rightarrow \max_{\alpha}$$

НЕ-БАЙЕСОВСКАЯ ИНТЕРПРЕТАЦИЯ МОДЕЛИ LDA

- $\beta_w > 1$: значения ϕ_{wt} сглаживаются, приближаясь к β_w^+ :

$$\text{KL}(\beta^+ \parallel \phi_t) \rightarrow \min \quad \beta_w^+ = \underset{w \in W}{\text{norm}}(\beta_w - 1)$$

- $\beta_w < 1$: значения ϕ_{wt} разреживаются, удаляясь от β_w^- к нулю:

$$\text{KL}(\beta^- \parallel \phi_t) \rightarrow \max \quad \beta_w^- = \underset{w \in W}{\text{norm}}(1 - \beta_w)$$

- Априорные распределения Дирихле больше не используются, можно снять ограничения:

$$\beta_w > 0 \quad \alpha_t > 0$$

ОНЛАЙНОВЫЙ ЕМ-АЛГОРИТМ

Вход: коллекция D , число тем $|T|$, параметры $i_{\max}, j_{\max}, \gamma$;

Выход: матрицы терминов тем Θ и тем документов Φ ;

инициализировать $n_{wt} := 0$;

для всех $i = 1, \dots, i_{\max}$ (итерации по коллекции)

для всех документов $d \in D$

для всех $j = 1, \dots, j_{\max}$ (итерации по документу)

$$p_{tdw} := \underset{t \in T}{\text{norm}}(\phi_{wt} \theta_{td});$$

$$\theta_{td} := \underset{t \in T}{\text{norm}} \left(\sum_{w \in d} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right);$$

$$n_{wt} := \gamma n_{wt} + n_{dw} p_{tdw};$$

если пора обновить матрицу Φ **то**

$$\phi_{wt} := \underset{w \in W}{\text{norm}} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right);$$

РЕЗЮМЕ

- *EM-алгоритм* — основной в тематическом моделировании
- Онлайновый *EM-алгоритм* позволяет обрабатывать большие коллекции за один проход
- Кроме PLSA и LDA придуманы сотни моделей, многие из них отличаются лишь регуляризаторами

РЕЗЮМЕ

- Кроме PLSA и LDA придуманы сотни моделей, многие из них отличаются лишь регуляризаторами
- В литературе по тематическому моделированию наиболее распространены методы байесовского вывода. Регуляризационный подход проще и универсальнее

РЕГУЛЯРИЗАЦИЯ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

АДДИТИВНАЯ РЕГУЛЯРИЗАЦИЯ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

- Максимизация правдоподобия с k регуляризаторами R_i :

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + \sum_{i=1}^k \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

где τ_i — коэффициенты регуляризации

АДДИТИВНАЯ РЕГУЛЯРИЗАЦИЯ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

- Максимизация правдоподобия с k регуляризаторами R_i :

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + \sum_{i=1}^k \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

- Типы регуляризаторов:
 - ▶ для учёта дополнительных данных
 - ▶ для получения решения Φ, Θ с заданными свойствами

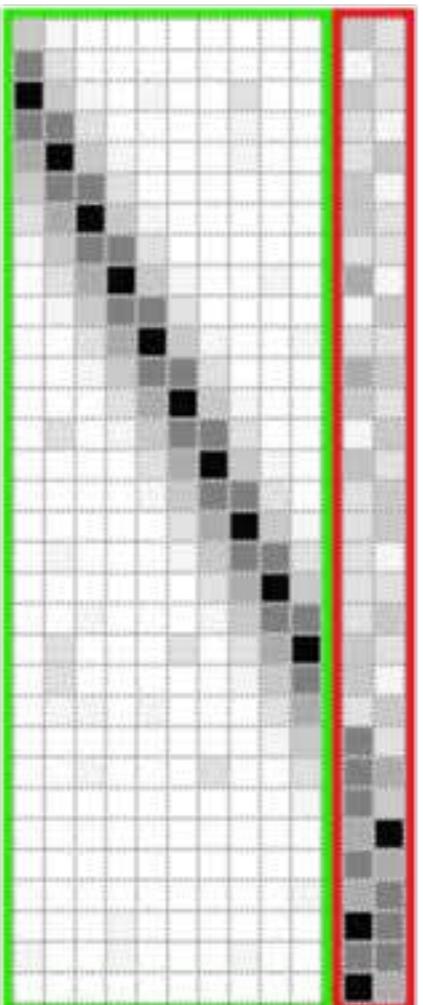
РАЗДЕЛЕНИЕ ТЕМ НА ПРЕДМЕТНЫЕ И ФОНОВЫЕ

- › Предметные темы **S** содержат термины предметной области
- › Фоновые темы **B** содержат слова общей лексики

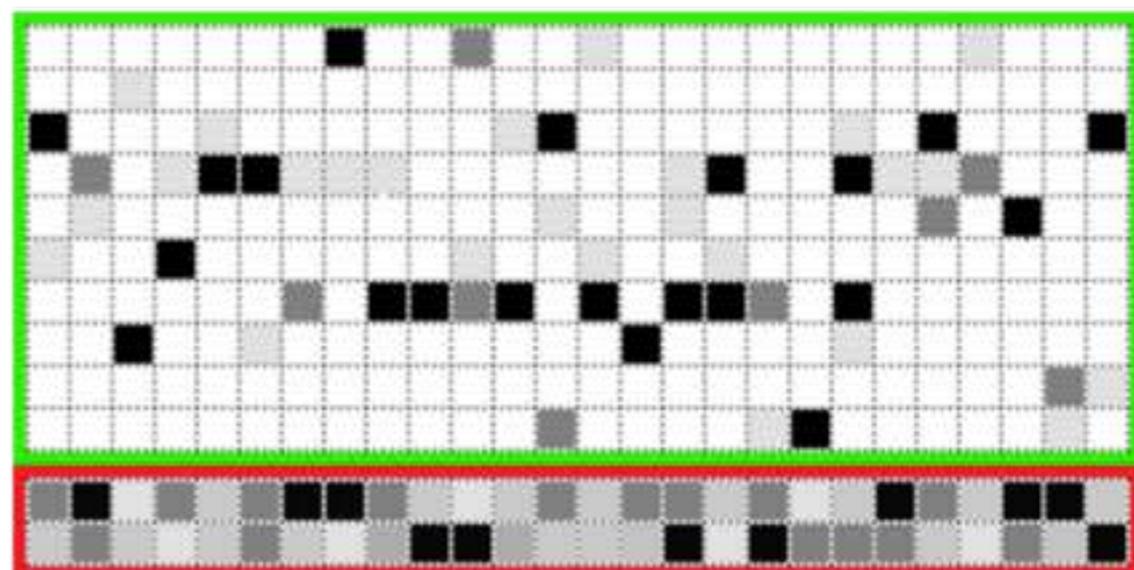
РАЗДЕЛЕНИЕ ТЕМ НА ПРЕДМЕТНЫЕ И ФОНОВЫЕ

- › Предметные темы S : $p(w|t), p(t|d), t \in S$
 - разреженные, существенно различные
- › Фоновые темы B : $p(w|t), p(t|d), t \in B$
 - существенно отличные от нуля

$\Phi_{W \times T}$



$\Theta_{T \times D}$



РЕГУЛЯРИЗАТОР СГЛАЖИВАНИЯ ФОНОВЫХ ТЕМ

- Распределения ϕ_{wt} близки к заданному распределению β_w
- Распределения θ_{td} близки к заданному распределению α_t

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_{w \in W} \beta_w \ln \phi_{wt} + \\ + \alpha_0 \sum_{d \in D} \sum_{t \in B} \alpha_t \ln \theta_{td} \rightarrow \max$$

где β_0, α_0 — коэффициенты регуляризации

РЕГУЛЯРИЗАТОР СГЛАЖИВАНИЯ ФОНОВЫХ ТЕМ

- Распределения ϕ_{wt} далеки от заданного распределения β_w
- Распределения θ_{td} далеки от заданного распределения α_t

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln \phi_{wt} - \\ - \alpha_0 \sum_{d \in D} \sum_{t \in S} \alpha_t \ln \theta_{td} \rightarrow \max$$

где β_0, α_0 — коэффициенты регуляризации

РЕГУЛЯРИЗАТОР ЧАСТИЧНОГО ОБУЧЕНИЯ

- Общий вид регуляризаторов сглаживания и разреживания:

$$\begin{aligned} R(\Phi, \Theta) = & \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \\ & + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \rightarrow \max \end{aligned}$$

РЕГУЛЯРИЗАТОР ЧАСТИЧНОГО ОБУЧЕНИЯ

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \\ + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \rightarrow \max$$

- » $\beta_{wt} = [w \in W_t]$ — белый список W_t терминов темы t
- » $\alpha_{td} = -[d \in D_t]$ — белый список D_t документов темы t
- » $\beta_{wt} = -[w \in W_t]$ — черный список W_t терминов темы t
- » $\alpha_{td} = -[d \in D_t]$ — черный список D_t документов темы t

РЕГУЛЯРИЗАТОР ДЕКОРРЕЛИРОВАНИЯ ТЕМ

- Цель: выделить лексическое ядро каждой темы
- Лексическое ядро темы — множество терминов, отличающее её от других тем.

РЕГУЛЯРИЗАТОР ДЕКОРРЕЛИРОВАНИЯ ТЕМ

- Цель: выделить лексическое ядро каждой темы
- Минимизируем ковариации между вектор-столбцами ϕ_t :

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

где τ — коэффициент регуляризации

РЕГУЛЯРИЗАТОР ДЛЯ ОТБОРА ТЕМ

- Цель: избавиться от мелких, зависимых и расщеплённых тем
- Разреживаем распределение

$$p(t) = \sum_d p(d) \theta_{td}$$

максимизируя **KL**-дивергенцию между $p(t)$ и равномерным распределением:

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

где τ — коэффициент регуляризации

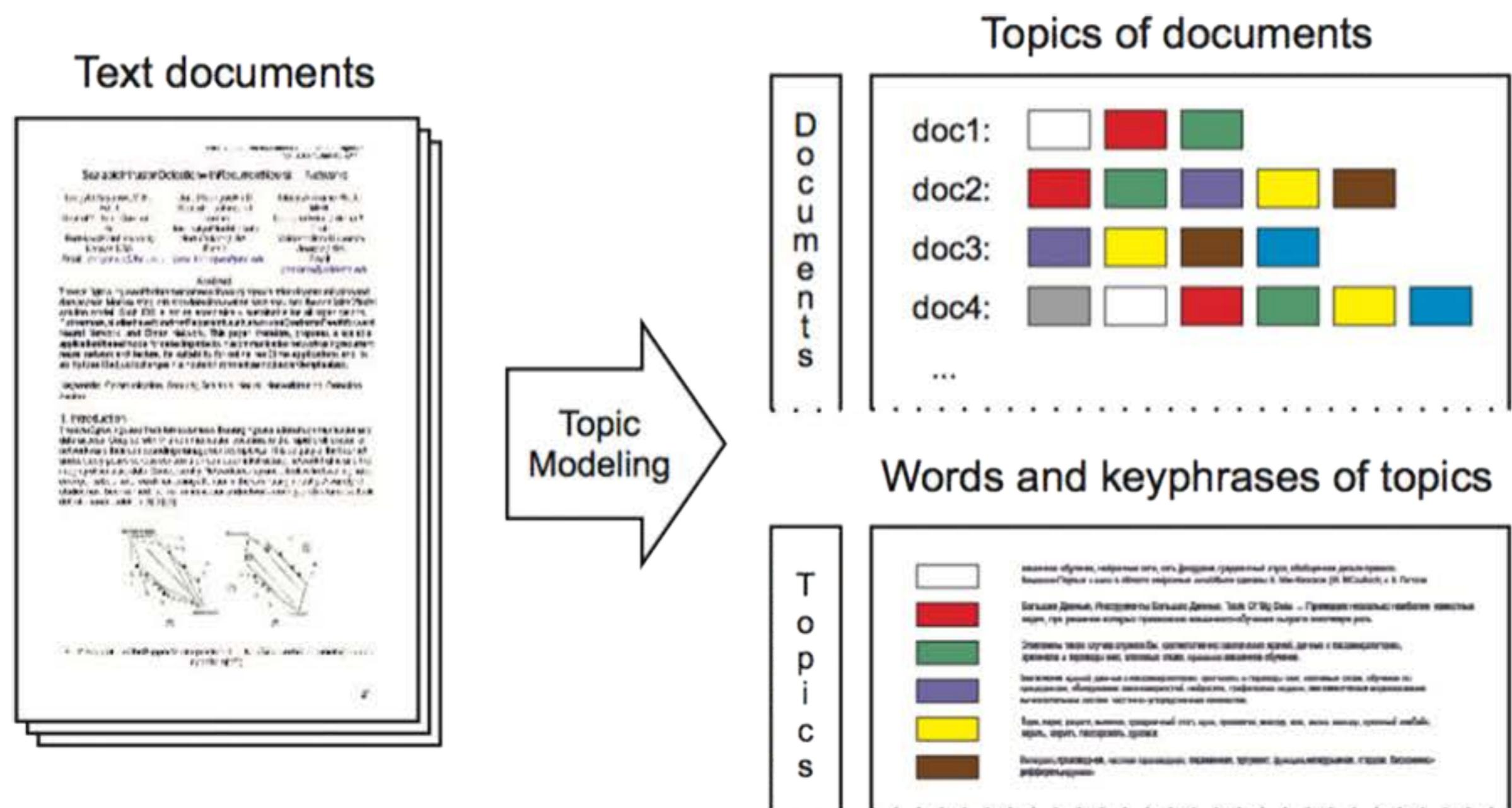
РЕЗЮМЕ

- Аддитивная регуляризация тематических моделей (ARTM) — подход к построению моделей с заданными свойствами
- Разреживание, сглаживание и декоррелирование вместе улучшают интерпретируемость и различность тем, выводят общую лексику из предметных тем в фоновые

МУЛЬТИМОДАЛЬНЫЕ ТЕМАТИЧЕСКИЕ МОДЕЛИ

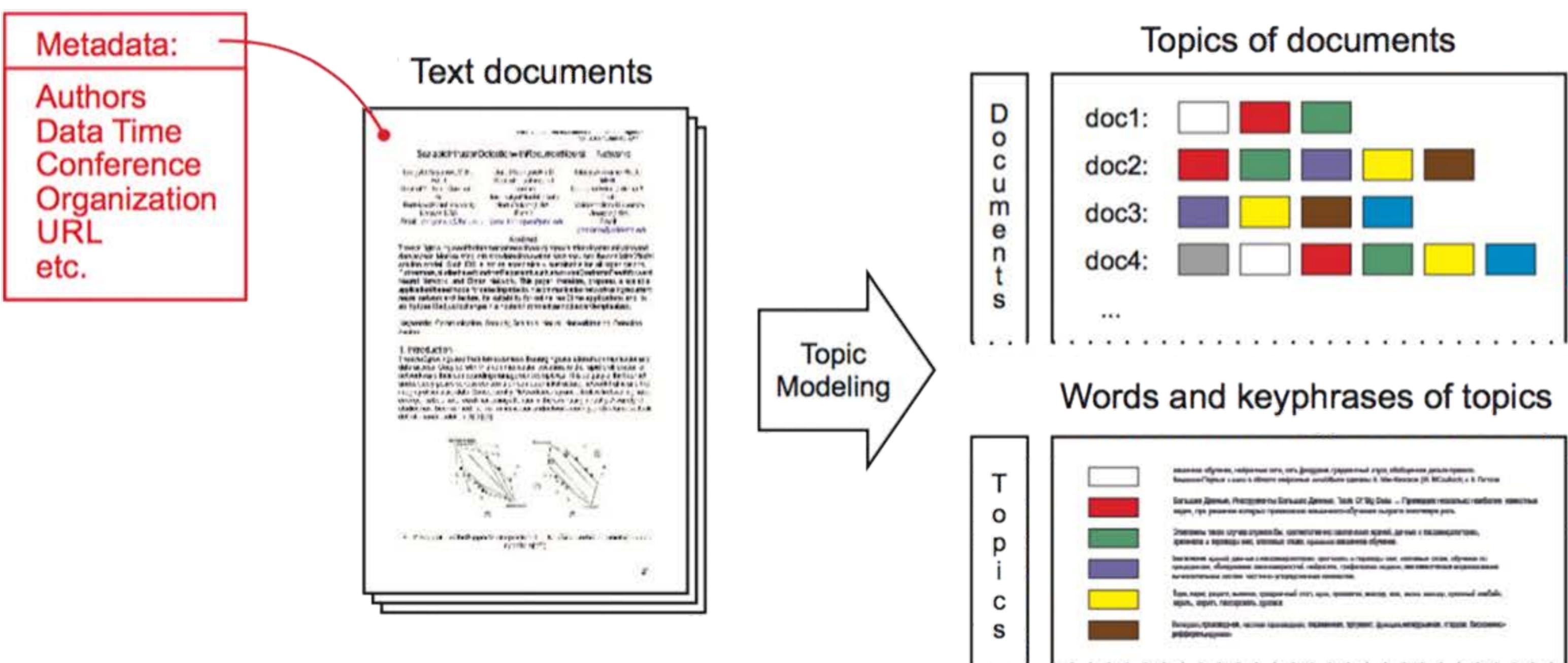
МУЛЬТИМОДАЛЬНАЯ ТЕМАТИЧЕСКАЯ МОДЕЛЬ

- Выявляет тематику документов $p(t|d)$, терминов $p(t|w), \dots$



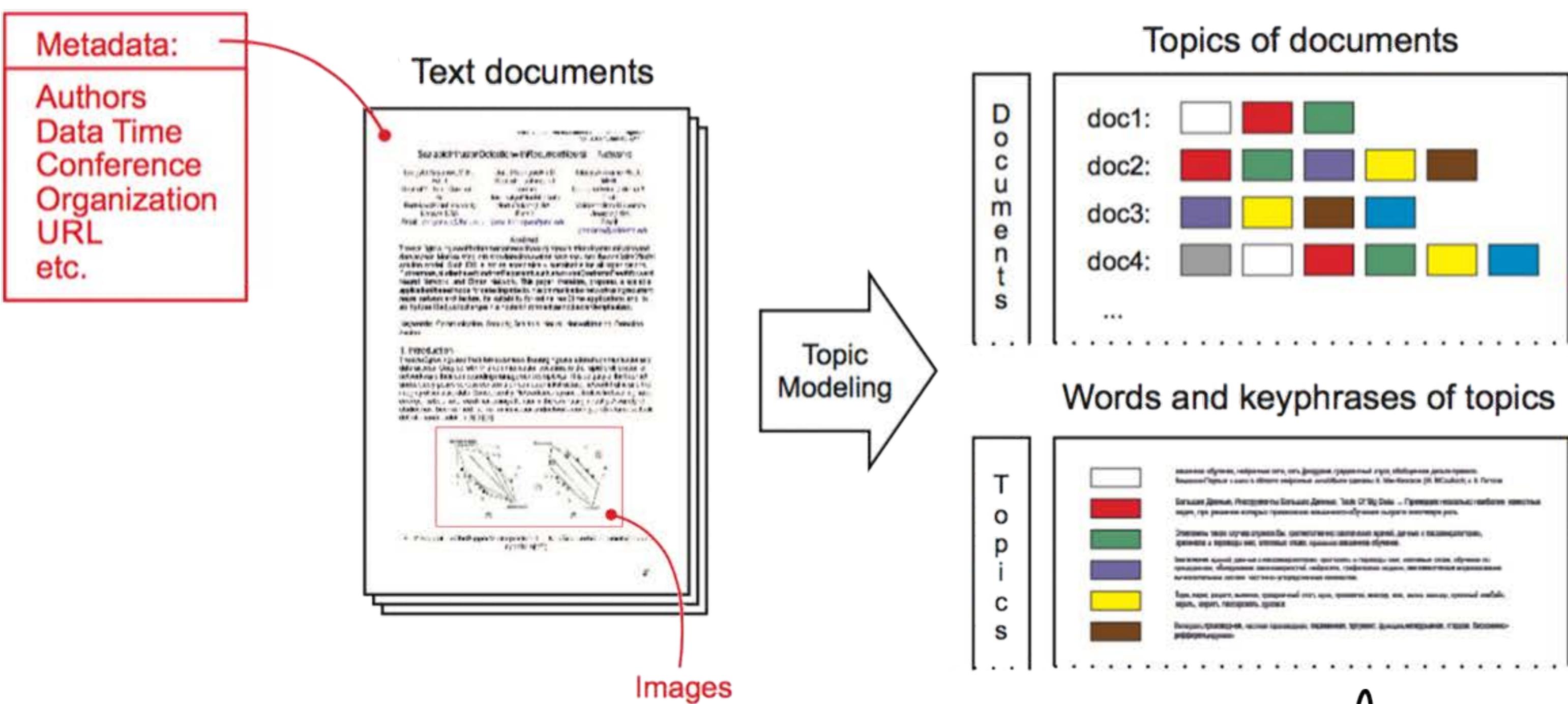
МУЛЬТИМОДАЛЬНАЯ ТЕМАТИЧЕСКАЯ МОДЕЛЬ

- » Выявляет тематику документов $p(t|d)$, терминов $p(t|w)$ и токенов других модальностей: $p(t|\text{автор})$, $p(t|\text{время})$, ...



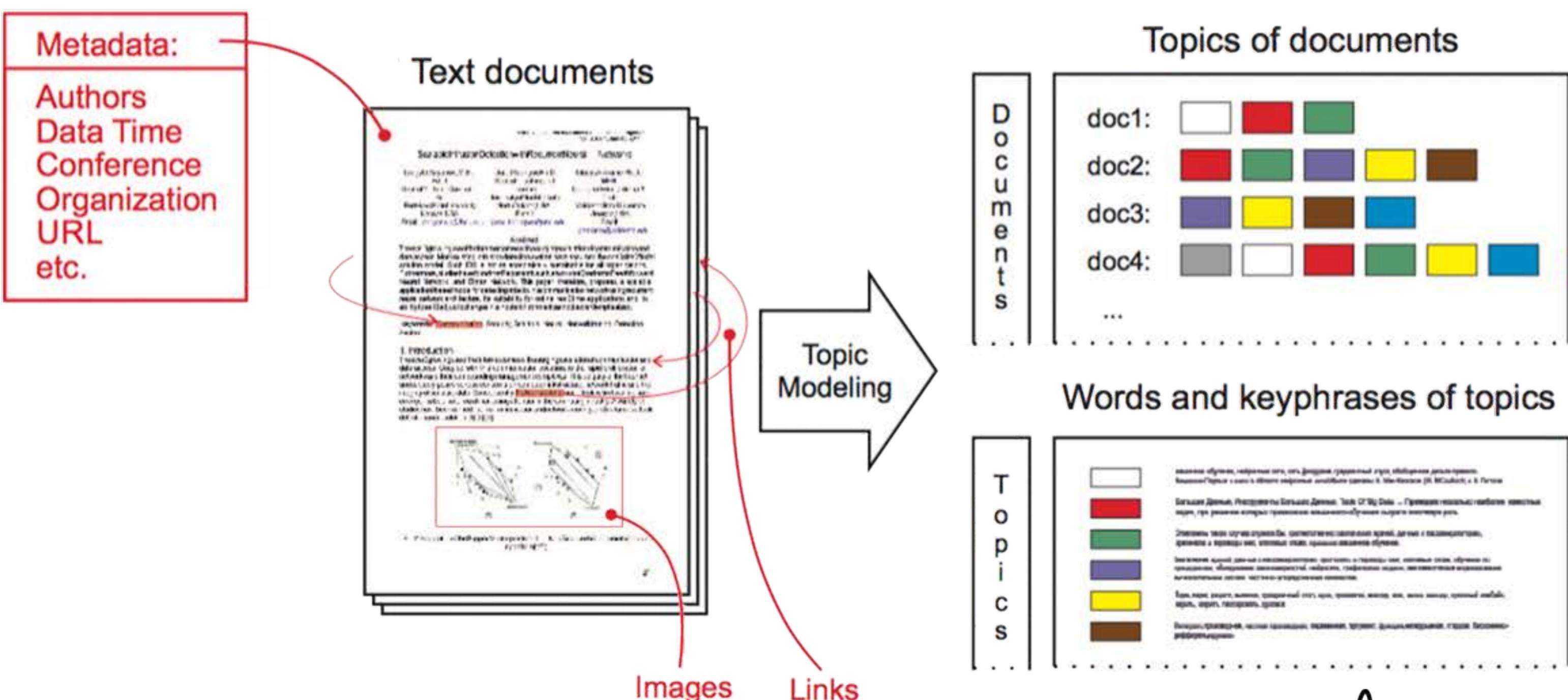
МУЛЬТИМОДАЛЬНАЯ ТЕМАТИЧЕСКАЯ МОДЕЛЬ

- Выявляет тематику токенов других модальностей: $p(t|\text{автор})$, $p(t|\text{время})$, $p(t|\text{элемент})$, ...



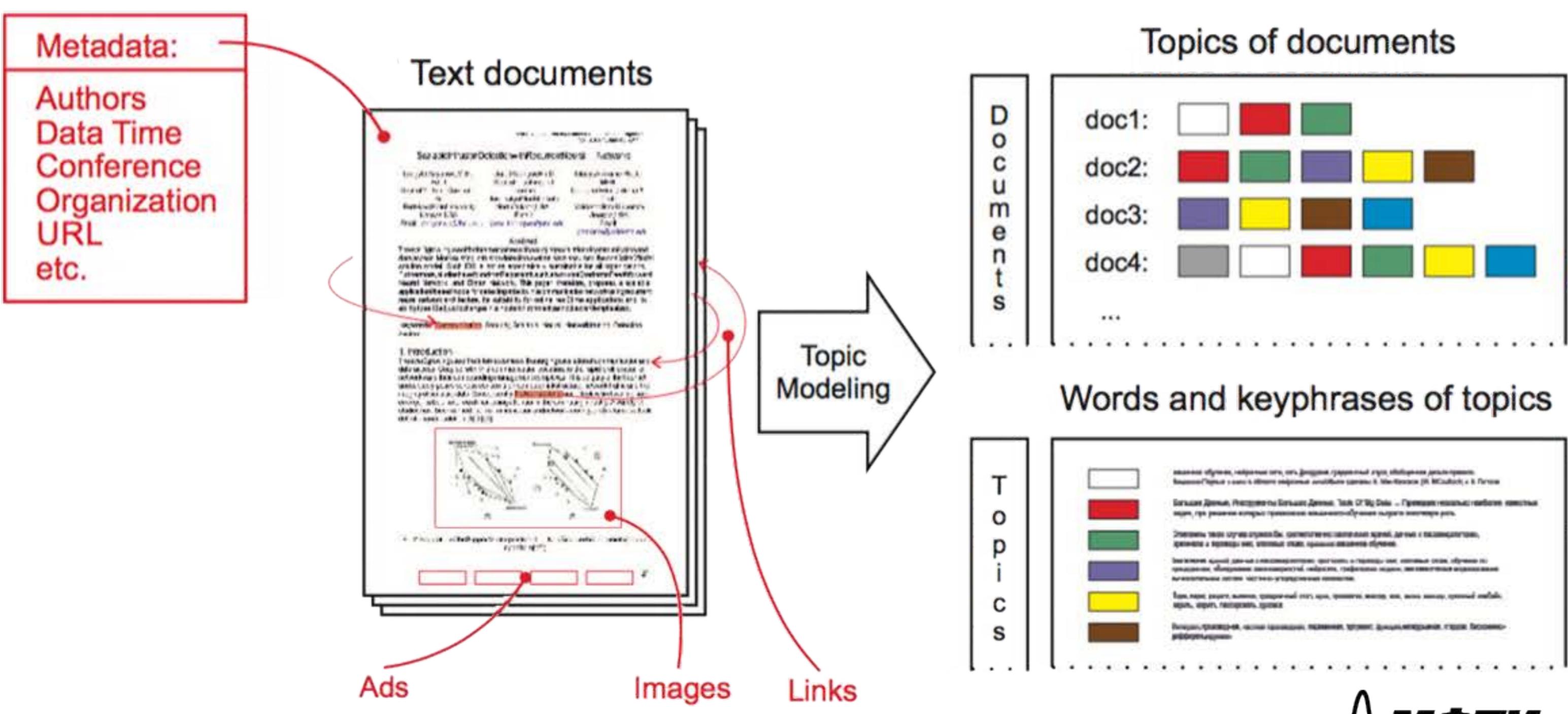
МУЛЬТИМОДАЛЬНАЯ ТЕМАТИЧЕСКАЯ МОДЕЛЬ

- Выявляет тематику токенов других модальностей: $p(t|\text{автор})$, $p(t|\text{время})$, $p(t|\text{элемент})$, $p(t|\text{ссылка})$, ...



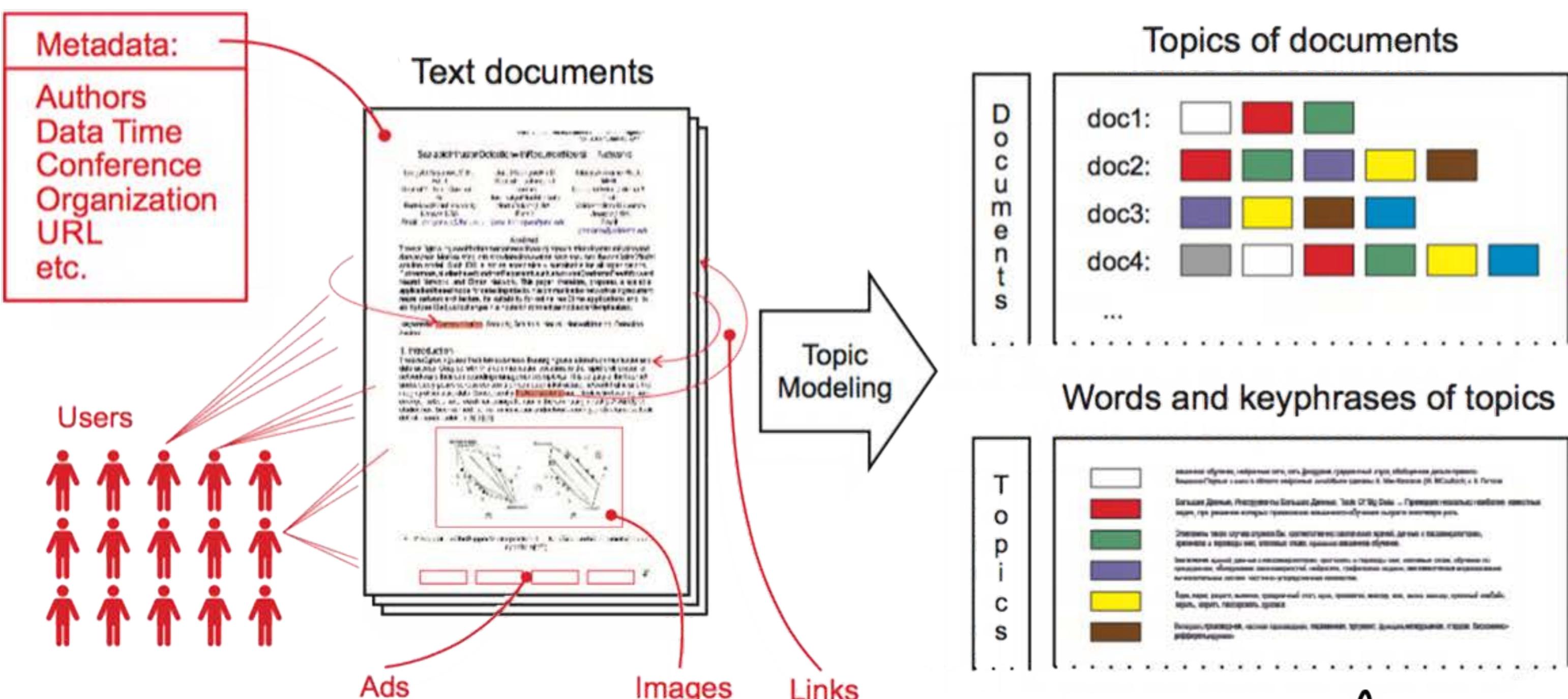
МУЛЬТИМОДАЛЬНАЯ ТЕМАТИЧЕСКАЯ МОДЕЛЬ

- › Выявляет тематику токенов других модальностей: $p(t|\text{автор})$, $p(t|\text{время})$, $p(t|\text{элемент})$, $p(t|\text{ссылка})$, $p(t|\text{баннер})$, ...

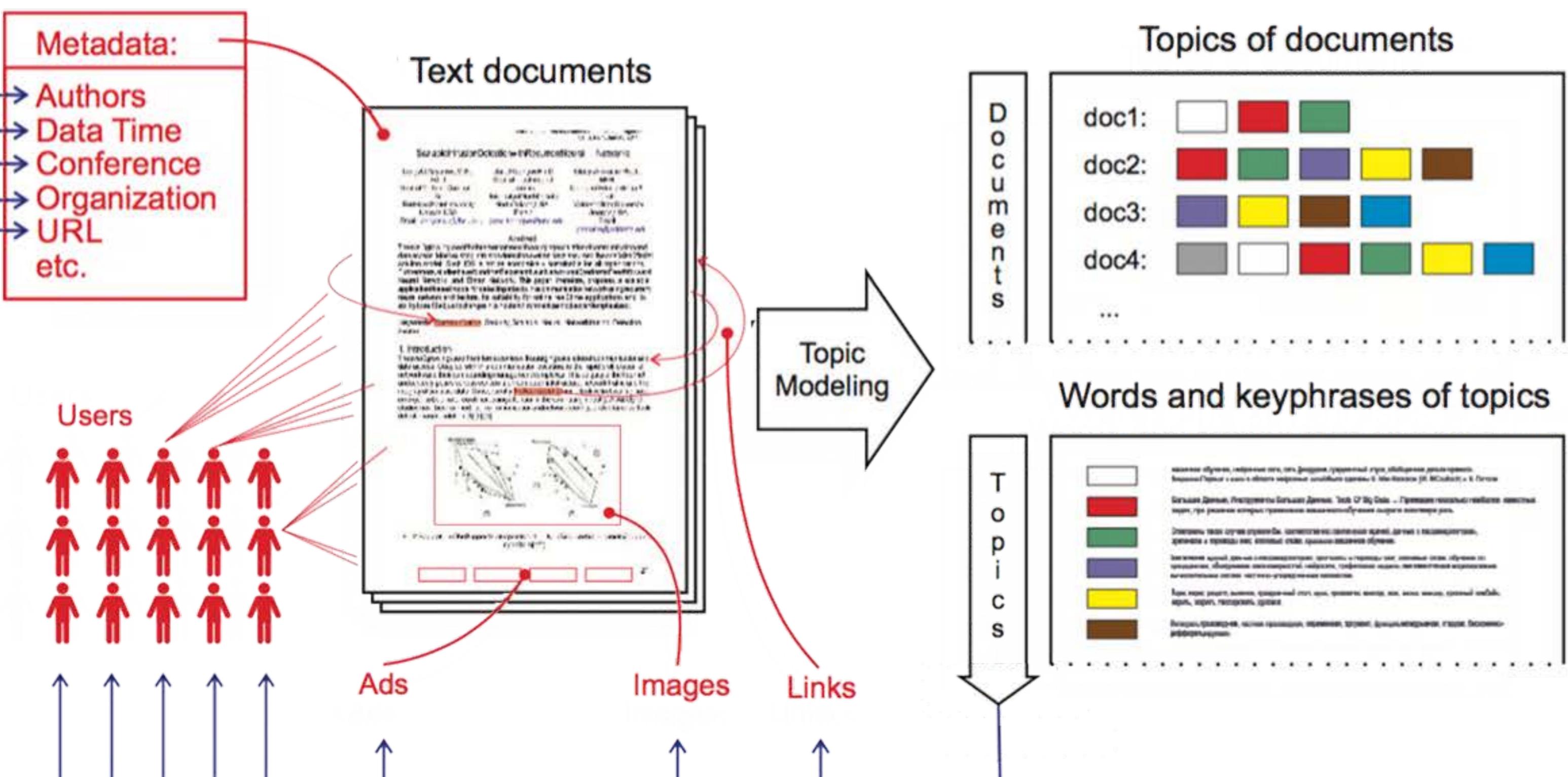


МУЛЬТИМОДАЛЬНАЯ ТЕМАТИЧЕСКАЯ МОДЕЛЬ

➤ ... токенов других модальностей:
 $p(t|\text{автор})$, $p(t|\text{время})$, $p(t|\text{элемент})$,
 $p(t|\text{ссылка})$, $p(t|\text{баннер})$, $p(t|\text{пользователь})$, ...



МУЛЬТИМОДАЛЬНАЯ ТЕМАТИЧЕСКАЯ МОДЕЛЬ



МУЛЬТИМОДАЛЬНАЯ ТЕМАТИЧЕСКАЯ МОДЕЛЬ

- » Выявляет тематику документов $p(t|d)$, терминов $p(t|w)$ и токенов других модальностей: $p(t|\text{автор})$, $p(t|\text{время})$, $p(t|\text{элемент})$, $p(t|\text{ссылка})$, $p(t|\text{баннер})$, $p(t|\text{пользователь})$
- »

описывается
енты могут
одальностей,
ределение:
n

МУЛЬТИМОДАЛЬНАЯ ТЕМАТИЧЕСКАЯ МОДЕЛЬ

- » Каждая модальность $t \in M$ описывается своим словарем W^t , документы могут содержать токены разных модальностей, каждая тема имеет своё распределение:
 $p(w|t), w \in W^t$
- » W^t — словарь токенов t -й модальности,
 $t \in M$
- » $W = W^1 \sqcup \dots \sqcup W^M$ — объединённый словарь всех модальностей

МУЛЬТИМОДАЛЬНАЯ ARTM. ПОСТАНОВКА ЗАДАЧИ

- › W^m — словарь токенов m -й модальности,
 $m \in M$
- › $W = W^1 \sqcup \dots \sqcup W^M$ — объединённый
словарь всех модальностей
- › Максимизация суммы \log -правдоподобий с
регуляризацией:

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

где τ_m — веса модальностей

МУЛЬТИМОДАЛЬНАЯ ARTM. ПОСТАНОВКА ЗАДАЧИ

- EM-алгоритм: метод простой итерации для системы уравнений относительно переменных ϕ_{wt} , θ_{td} и $p_{tdw} = p(t|d, w)$:

$$\left\{ \begin{array}{l} p_{tdw} = \underset{t \in T}{\text{norm}}(\phi_{wt}\theta_{td}) \\ \phi_{wt} = \underset{w \in W^m}{\text{norm}} \left(\sum_{d \in D} \tau_{m(w)} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \underset{t \in T}{\text{norm}} \left(\sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{array} \right.$$

РЕЗЮМЕ

- Модальности очень легко добавляются в ARTM
- Но при этом сильно расширяют спектр решаемых задач
- Каждая модальность, как и слова, имеет свой словарь
- Документы — контейнеры токенов разных модальностей

РЕЗЮМЕ

- Нетрииальные примеры модальностей:
 - ▶ Пользователи документов
 - ▶ Языки в параллельных и сравнимых коллекциях
 - ▶ n -граммы и словосочетания
 - ▶ Время