- **Гипотеза условной независимости:** $p(w|t,d) = p(w|t)$ заключается в том, что вероятность слова документах определяется только темой, а не самим документом. Это предположение позволяет строить легко оцениваемые тематические модели.

Часто используются дополнительные предположения разреженности:

- Предположение, что документ относится к небольшому числу тем.
- Предположение, что тема состоит из небольшого числа терминов, лексического ядра, которое существенно отличает эту тему от остальных.

7.2.3. Вероятностный процесс порождения текстовой коллекции

В вероятностной порождающей модели документ d — это смесь распределений $p(w|t)$ с весами $p(t|d)$:

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

Условное распределение тем в документе $p(t|d)$ — важный параметр модели, который и необходимо оценивать.

Таким образом, процесс порождения текста следующий. Для каждой словопозиции w сначала из распределения тем в документе выбирается тема, к которой это слово будет относиться. После этого из распределения слов в выбранной теме выбирается конкретное слово, которое будет записано в данную словопозицию. Слово за словом так появляется весь текст.

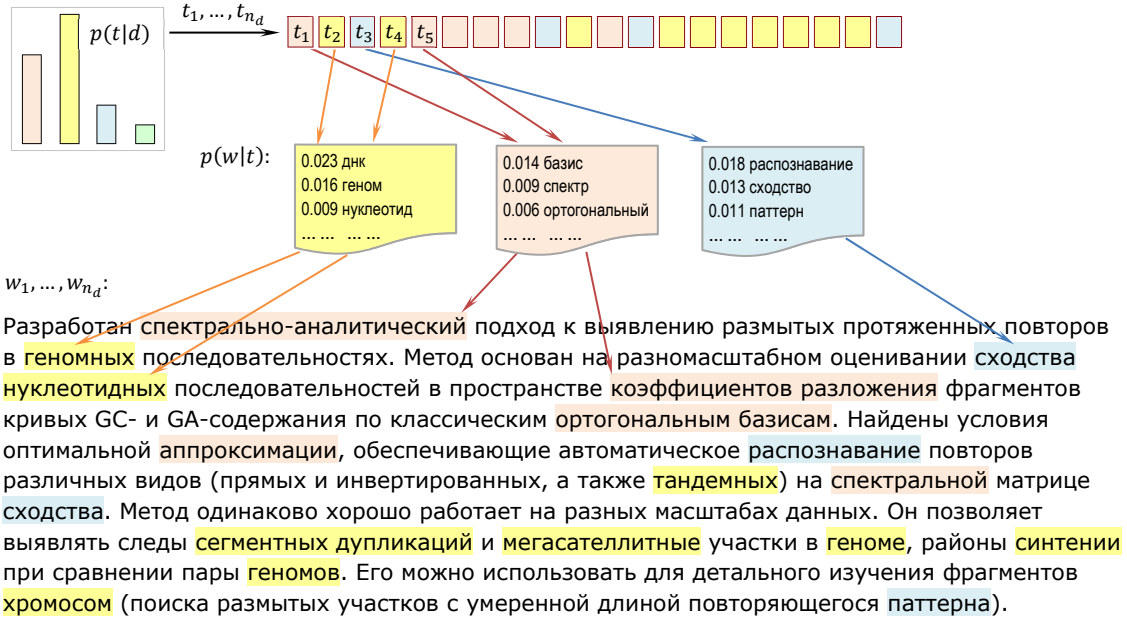
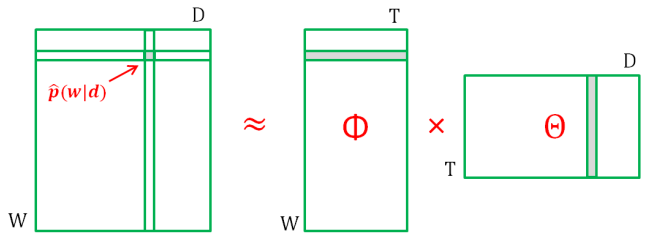


Рис. 7.1: Процесс порождения текстового документа вероятностной тематической моделью

Поскольку выполняется гипотеза «мешка слов» сгенерированный текст вряд ли будет осмысленным. Можно только говорить о том, что с точностью до произвольной перестановки слов, этот текст вполне мог бы нести в себе какую-то тематику. А именно тематику текста и нужно выявить. Другими словами, тематическое моделирование не обеспечивает понимание компьютером смысла текста, а только позволяет выполнить кластеризацию документов по темам.

7.2.4. Постановка задачи тематического моделирования

Формальная постановка задачи тематического моделирования следующая. Пусть зафиксирован словарь терминов W, из элементов которого складываются документы, и дана коллекция D документов $d \subset W$. Для каждого документа d известна его длина n_d и количество n_{dw} использований каждого термина w.



Требуется найти параметры вероятностной порождающей тематической модели, то есть представить вероятность появления $p(w|d)$ слов в документе в виде:

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

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

Порождающая модель описывает процесс построения коллекции по ϕ_{wt} и θ_{td}. Тематическое моделирование представляет собой обратную задачу: по наблюдаемой коллекции необходимо понять, какими распределениями ϕ_{wt} и θ_{td} она могла бы быть получена.

7.2.5. Задача тематического моделирования как задача матричного разложения

Фактически, эту задачу можно трактовать как задачу матричного разложения. Пусть Φ — матрица распределений терминов в темах, а Θ — матрица распределений тем в документах:

$$\Phi = \big(\phi_{wt}\big), \qquad \Theta = \big(\theta_{td}\big).$$

Матрицы называются стохастическими, если каждый их столбец преставляет собой дискретное распределение вероятностей, а ,следовательно, сумма значений по каждому столбцу равна 1 (условие нормировки) и каждое значение является неотрицательным (условие неотрицательности). Следует особо отметить, что стохастические матрицы — это НЕ такие матрицы, элементы которых генерируются случайно. Обе определенные выше матрцы Φ и Θ — стохастические. Согласно вероятностной тематической модели, произведение матриц Φ и Θ должно давать частотные оценки $p(w|d)$ условных вероятностей слов в документах коллекции. Наблюдаемые частоты терминов в документах известны:

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

Задача тематического моделирования, таким образом, стала задачей стохастического матричного разложения матрицы $\big(\hat{p}(w|d)\big)$ на стохастические матрицы Φ и Θ.

Теперь можно воспользоваться приниципом максимума правдоподобия с ограничениями, следующими из условий нормировки и неотрицательности на элементы стохастических матриц. Если максимизировать логарифм правдоподобия, получается:

$$\begin{cases} \sum\limits_{d \in D} \sum\limits_{w \in d} n_{dw} \ln \sum\limits_{t \in T} \phi_{wt} \theta_{td} \to \max\limits_{\Phi, \Theta}; \\ \sum\limits_{w \in W} \phi_{wt} = 1; \qquad \phi_{wt} \geq 0; \\ \sum\limits_{t \in T} \theta_{td} = 1; \qquad \theta_{td} \geq 0. \end{cases}$$

7.2.6. Принцип максимума регуляризованного правдоподобия

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

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

С одной стороны, строящаяся математическая модель получается неустойчивой и невоспроизводимой (результат работы итерационных методов будет зависет от начального приближения), но, с другой стороны, это

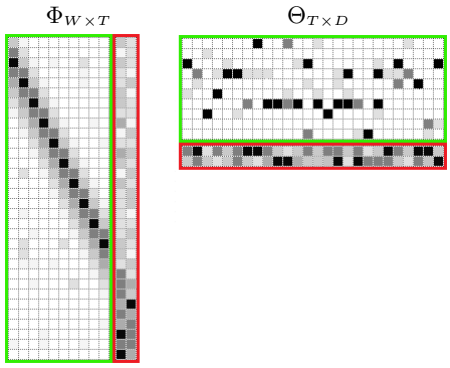


Рис. 7.3: Матрицы Φ и Θ

Такой подход дает возможность наложить сразу несколько условий, но также появляется проблема нахождения коэффициентов регуляризации. На данный момент, в основном, регуляризаторы добавляются по одному и у каждого регуляризатора оптимизируя этот коэффициент в ходе нескольких пробных запусков модели.

7.4.2. Разделение тем на предметные и фоновые

Продемонстрировать, как используя несколько регуляризаторов наделить модель нужными свойствами, можно на следующем примере. Наличие слов общей лексики в теме приводит к плохой интерпретируемости данной темы. Поэтому хотелось бы такие общеупотребительные слова выделить в отдельные темы, так называемые фоновые темы. Все остальные темы называются, соответственно, предметными, так как они описывают предметные области текстовой коллекции.

Предметные темы должны быть достаточно сильно разреженными, чтобы у каждой такой темы существовало свое лексическое ядро, существенно отличающее эту тему от остальных. Другими словами, требуется не только разреженность тем, но и их декоррелированность.

Эти требования можно выразить с помощью регуляризаторов. Пусть S — множество предметных тем, а B — множество фоновых. Поскольку для предметных тем ($t \in S$) матрицы $p(w|t)$ и $p(t|d)$ должны быть разреженными и существенно различными, а для фоновых ($t \in B$) — существенно отличными от нуля (больше половины слов в каждом документе — фоновые), имеет смысл применить регуляризатор, рассмотренный ранее в методе латентного размещения Дирихле. Единственное отличие состоит в том, что тогда он применялся для всего словаря, а в данном случае регуляризатор сглаживания необходимо применить только к фоновым темам:

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

где β_0, α_0 — коэффициенты регуляризации. В этом случае распределения ϕ_{wt} будут близки к заданному распределению β_w, а распределения θ_{td} — к заданному распределению α_t. Распределения β_w и α_t вычисляются заранее. Например, в качестве β_w можно использовать распределение слов в используемом языке.

По аналогии можно построить разреживающий регуляризатор для предметных тем:

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

где β_0, α_0 — коэффициенты регуляризации. В этом случае распределения ϕ_{wt} и θ_{td} будут как можно далеки от заданных распределений β_w и α_t. Определением параметров β_w и α_t занимается специалист, который занимается построением тематической модели. Часто в качестве β_w также используют распределение слов в используемом языке, а в качестве α_t — равномерное распределение.

7.4.3. Регуляризатор частичного обучения (semi-supervised learning)

Интересное обобщение этих двух регуляризаторов — сглаживающего и разреживающего — возникает в том случае, если векторы β_{wt} и α_{td} могут быть свои для каждого столбца:

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

7.5. Мультимодальные тематические модели

7.5.1. Понятие модальности

На практике часто встречаются коллекции документов, которые включают в себя метаинформацию, связывающую каждый документ с элементами (токенами) каких-то конечных множеств (не обязательно слов). Эти конечные множества называются модальностью.

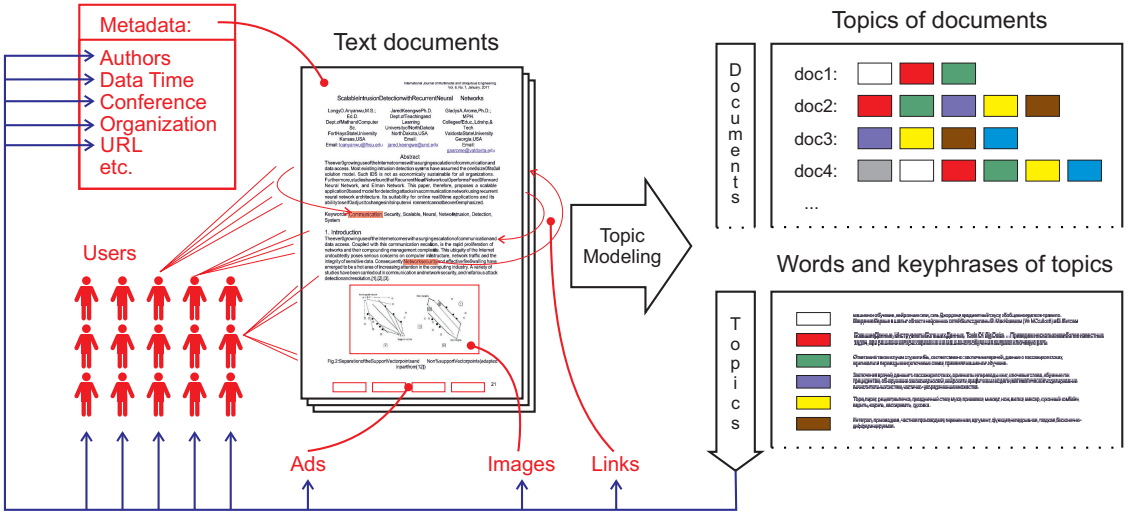


Рис. 7.4: Мультимодальная тематическая модель описывает появление элементов разных модальностей

Примеры модальностей:

- **Авторы, моменты времени и так далее:** в этом случае каждому документу приписывается соответственно метка автора, временная метка и так далее.
- **Элементы изображений**, содержащихся в документе. Изображение в таком случае можно мыслить как мини-документ, состоящий из псевдослов — элементов изображений.
- **Множество ссылок на другие документы**, в том числе гиперссылки в сети Интернет и цитирование других статей в научных трудах.
- **Множество рекламных баннеров, которые появились на данной странице**, а также **множество пользователей, которые кликнули на данные баннеры**, это два примера возможных модальностей.
- **Множество пользователей, сделавших определенное действие с документом (скачал, лайкнул, поставил оценку и так далее)**. После того, как операция выполнена, в системе остается запись о том, что данный пользователь сделал конкретную операцию. И поэтому можно считать, что в документ также включена и эта информация.

Чтобы иметь возможность пользоваться данными типами информации, необходимо строить тематические модели, которые описывают появление элементов разных модальностей в документе по известной тематике. Другими словами, благодаря тому, что документ относится к какой-либо теме, в нем появляются определенные слова из этой темы, на картинках изображены элементы, которые характерны для этой темы, а также его читают пользователи, которым эта тема интересна, и так далее.

7.5.2. Мультимодальная ARTM

Тематическая модель описывает появление элементов всех модальностей исходя из единого тематического профиля всего документа. Каждая модальность $m \in M$ описывается своим словарём токенов W^m, каждая тема имеет своё распределение $p(w|t)$ для каждой модальности $w \in W^m$.