ОТ НАИВНОГО БАЙЕСА ДО ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ II

Сергей Николенко

НИУ ВШЭ — Санкт-Петербург
13 октября 2017 г.

Random facts:

- 13 октября 54 г. жена императора Клавдия отравила его грибами, и на престол взошел Нерон
- 13 октября 1884 г. Гринвич был утвержден как место прохождения нулевого меридиана
- 13 октября 1917 г. в португальском городе Фатима случилось чудо: явление Девы Марии и чудесное превращение Солнца, официально признанные чудом католической церковью

LDA

LDA

- Развитие идей pLSA – LDA (Latent Dirichlet Allocation).
- Это фактически байесовский вариант pLSA, сейчас нарисуем картинки, добавим априорные распределения и посмотрим, как сработают наши методы приближённого вывода.
- Задача та же: смоделировать большую коллекцию текстов (например, для information retrieval или классификации).

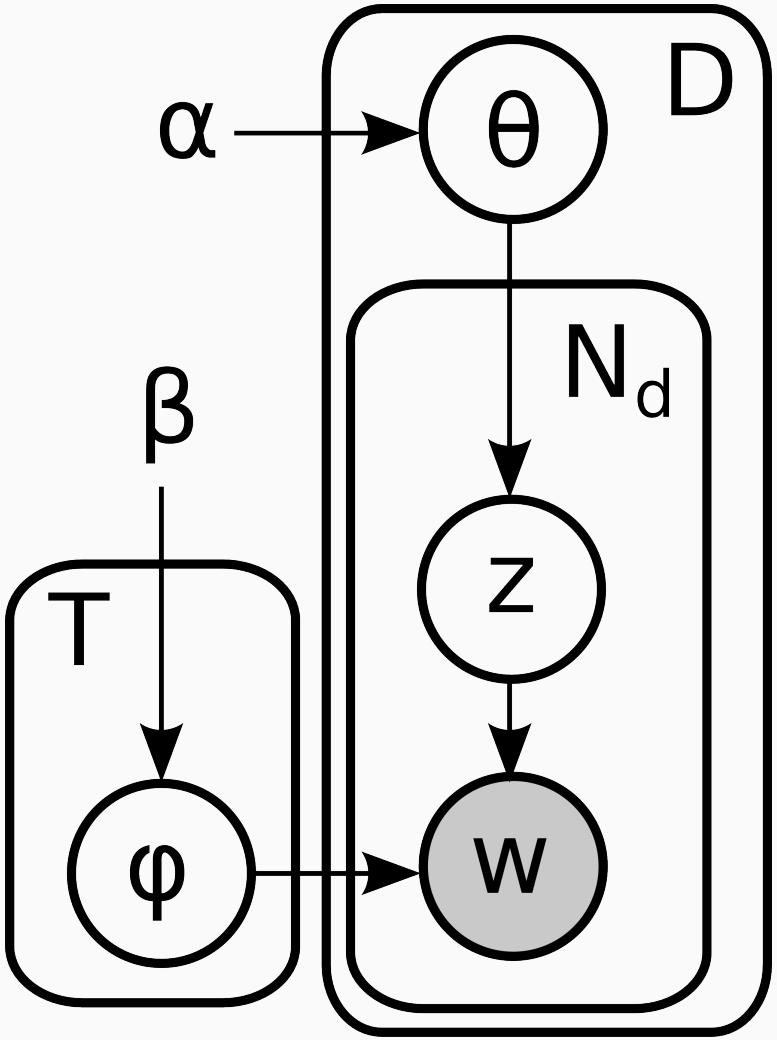
LDA

- У одного документа может быть несколько тем. Давайте построим иерархическую байесовскую модель:
  - на первом уровне – смесь, компоненты которой соответствуют «темам»;
  - на втором уровне – мультиномиальная переменная с априорным распределением Дирихле, которое задаёт «распределение тем» в документе.

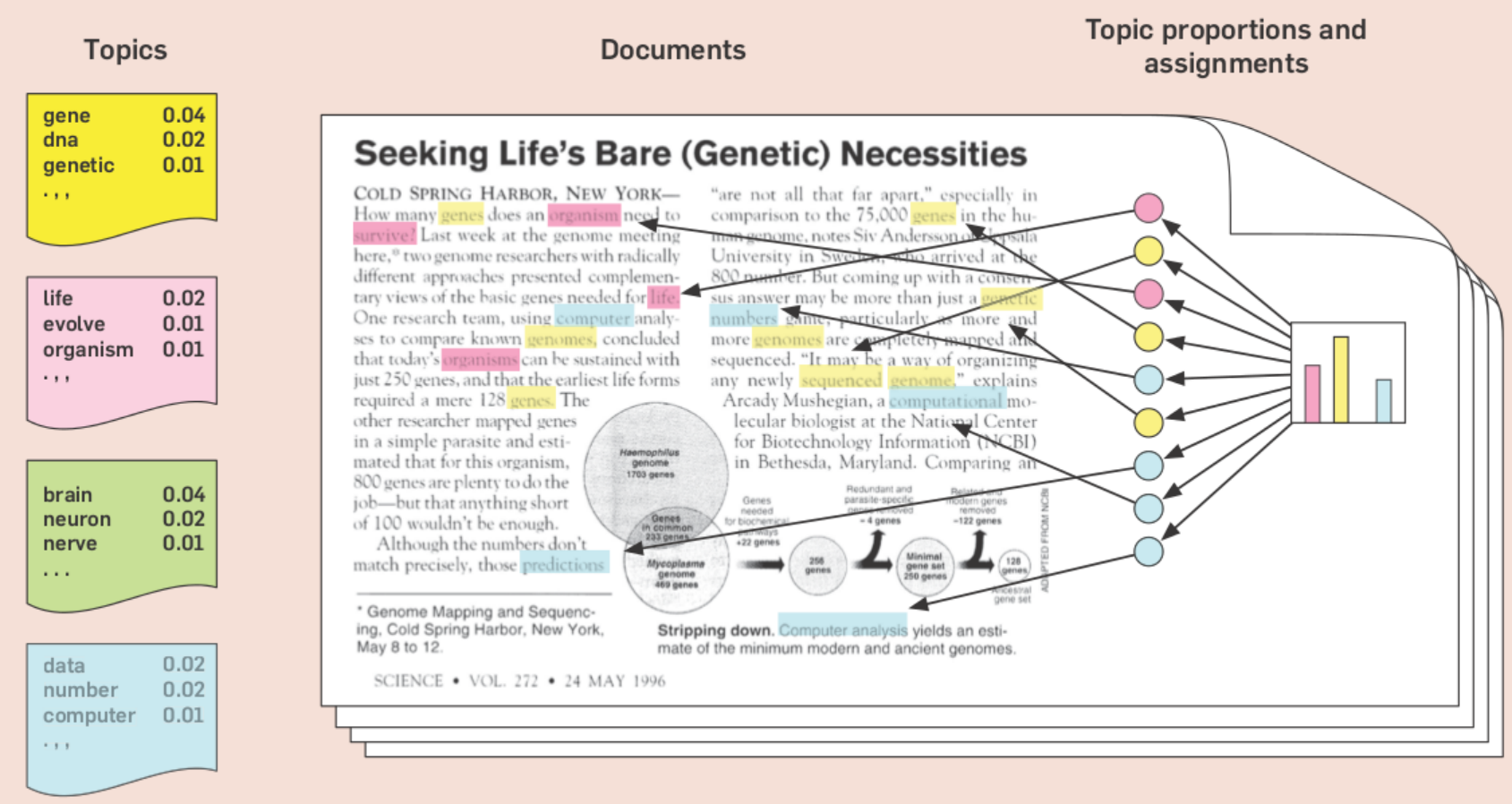
LDA

- Если формально: слова берутся из словаря $\{1, \dots, V\}$; слово – это вектор w, $w_i \in \{0, 1\}$, где ровно одна компонента равна 1.
- Документ – последовательность из N слов $\mathbf{w}$. Нам дан корпус из M документов $\mathcal{D} = \{\mathbf{w}_d \mid d = 1..M\}$.
- Генеративная модель LDA выглядит так:
  - выбрать $\theta \sim \mathbf{Di}(\alpha)$;
  - для каждого из N слов w_n:
    - выбрать тему $z_n \sim \mathbf{Mult}(\theta)$;
    - выбрать слово $w_n \sim p(w_n \mid z_n, \beta)$ по мультиномиальному распределению.

LDA: графическая модель



LDA: что получается [Blei, 2012]



LDA: вывод

- Два основных подхода к выводу в сложных вероятностных моделях, в том числе LDA:
  - *вариационные приближения*: рассмотрим более простое семейство распределений с новыми параметрами и найдём в нём наилучшее приближение к неизвестному распределению;
  - *сэмплирование*: будем набрасывать точки из сложного распределения, не считая его явно, а запуская марковскую цепь под графиком распределения (частный случай – сэмплирование по Гиббсу).
- Сэмплирование по Гиббсу обычно проще расширить на новые модификации LDA, но вариационный подход быстрее и часто стабильнее.

ВЫВОД В LDA

- Рассмотрим задачу байесовского вывода, т.е. оценки апостериорного распределения θ и z после нового документа:

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}.$$

- Правдоподобие набора слов $\mathbf{w}$ оценивается как

$$p(\mathbf{w} \mid \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left[\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}\right] \left[\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j}\right] d\theta,$$

и это трудно посчитать, потому что θ и β путаются друг с другом.

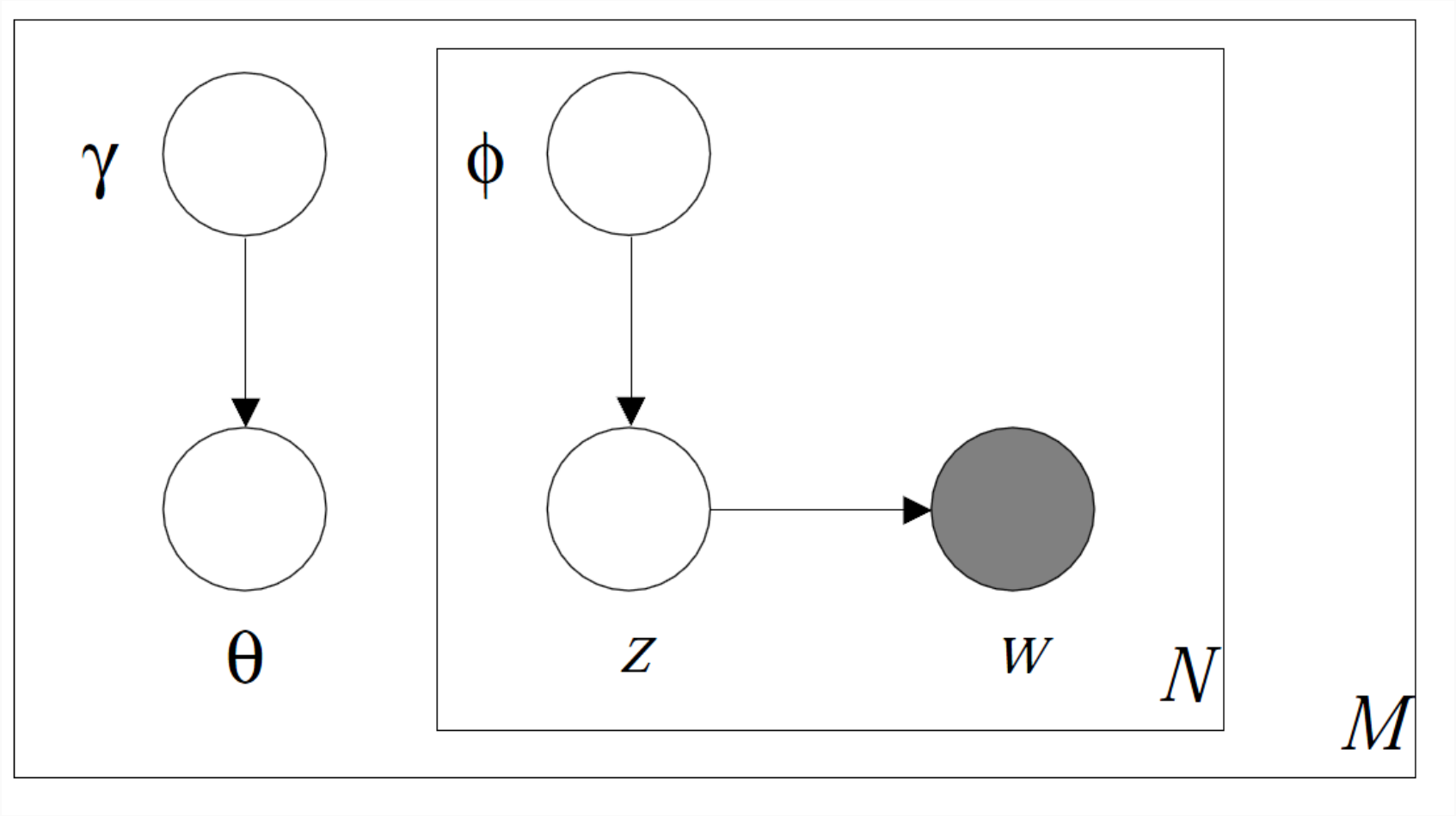
ВЫВОД В LDA

- Вариационное приближение – рассмотрим семейство распределений

$$q(\theta, z \mid \mathbf{w}, \gamma, \phi) = p(\theta \mid \mathbf{w}, \gamma) \prod_{n=1}^{N} p(z_n \mid \mathbf{w}, \phi_n).$$

- Тут всё расщепляется, и мы добавили вариационные параметры γ (Дирихле) и ϕ (мультиномиальный).
- Заметим, что параметры для каждого документа могут быть свои – всё условно по $\mathbf{w}$.

LDA: ВАРИАЦИОННОЕ ПРИБЛИЖЕНИЕ



LDA: вариационный вывод

- Теперь можно искать минимум KL-расстояния:

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} \mathrm{KL}(q(\theta, z \mid \mathbf{w}, \gamma\phi) \| p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)).$$

- Для этого сначала воспользуемся уже известной оценкой из неравенства Йенсена:

$$\log p(\mathbf{w} \mid \alpha, \beta) = \log \int_\theta \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) d\theta =$$

$$= \log \int_\theta \sum_{\mathbf{z}} \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) q(\theta, \mathbf{z})}{q(\theta, \mathbf{z})} d\theta \geq$$

$$\geq E_q \left[\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)\right] - E_q \left[\log q(\theta, \mathbf{z})\right] =: \mathcal{L}(\gamma, \phi; \alpha, \beta).$$

LDA: вариационный вывод

- Распишем произведения:

$$\mathcal{L}(\gamma, \phi; \alpha, \beta) = E_q\left[p(\theta \mid \alpha)\right] + E_q\left[p(\mathbf{z} \mid \theta)\right] + E_q\left[p(\mathbf{w} \mid \mathbf{z}, \beta)\right] - \\ - E_q\left[\log q(\theta)\right] - E_q\left[\log q(\mathbf{z})\right].$$

- Свойство распределения Дирихле: если $X \sim \mathrm{Di}(\alpha)$, то

$$E[\log(X_i)] = \Psi(\alpha_i) - \Psi(\sum_i \alpha_i),$$

  где $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ – дигамма-функция.
- Теперь можно выписать каждый из пяти членов.

LDA: вариационный вывод

- Теперь осталось только брать частные производные этого выражения.
- Сначала максимизируем его по ϕ_{ni} (вероятность того, что n-е слово было порождено темой i); надо добавить λ-множители Лагранжа, т.к. $\sum_{j=1}^{k} \phi_{nj} = 1$.
- В итоге получится:

$$\phi_{ni} \propto \beta_{iv} e^{\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)},$$

где v – номер того самого слова, т.е. единственная компонента $w_n^v = 1$.

LDA: вариационный вывод

- Потом максимизируем по γ_i, i-й компоненте апостериорного Дирихле-параметра.
- Получится

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}.$$

- Соответственно, для вывода нужно просто пересчитывать ϕ_{ni} и γ_i друг через друга, пока оценка не сойдётся.

LDA: ОЦЕНКА ПАРАМЕТРОВ

- Теперь давайте попробуем оценить параметры α и β по корпусу документов $\mathcal{D}$.
- Мы хотим найти α и β, которые максимизируют

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} \log p(\mathbf{w}_d \mid \alpha, \beta).$$

- Подсчитать $p(\mathbf{w}_d \mid \alpha, \beta)$ мы не можем, но у нас есть нижняя оценка $\mathcal{L}(\gamma, \phi; \alpha, \beta)$, т.к.

$$p(\mathbf{w}_d \mid \alpha, \beta) = \\ = \mathcal{L}(\gamma, \phi; \alpha, \beta) + \mathrm{KL}(q(\theta, z \mid \mathbf{w}_d, \gamma\phi) \| p(\theta, \mathbf{z} \mid \mathbf{w}_d, \alpha, \beta)).$$

LDA: оценка параметров

- EM-алгоритм:
  1. найти параметры $\{\gamma_d, \phi_d \mid d \in \mathcal{D}\}$, которые оптимизируют оценку (как выше);
  2. зафиксировать их и оптимизировать оценку по α и β.

LDA: ОЦЕНКА ПАРАМЕТРОВ

- Для β это тоже делается нехитро:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_n^j.$$

- Для α_i получается система уравнений, которую можно решить методом Ньютона.

LDA: сэмплирование по Гиббсу

- В базовой модели LDA сэмплирование по Гиббсу после несложных преобразований сводится к так называемому *сжатому сэмплированию по Гиббсу* (collapsed Gibbs sampling), где переменные z_w итеративно сэмплируются по следующему распределению:

$$p(z_w = t \mid \mathbf{z}_{-w}, \mathbf{w}, \alpha, \beta) \propto q(z_w, t, \mathbf{z}_{-w}, \mathbf{w}, \alpha, \beta) =$$

$$\frac{n^{(d)}_{-w,t} + \alpha}{\sum_{t' \in T} \left(n^{(d)}_{-w,t'} + \alpha\right)} \frac{n^{(w)}_{-w,t} + \beta}{\sum_{w' \in W} \left(n^{(w')}_{-w,t} + \beta\right)},$$

где $n^{(d)}_{-w,t}$ – число слов в документе d, выбранных по теме t, а $n^{(w)}_{-w,t}$ – число раз, которое слово w было порождено из темы t, не считая текущего значения z_w; заметим, что оба этих счётчика зависят от других переменных $\mathbf{z}_{-w}$.

LDA: сэмплирование по Гиббсу

- Из сэмплов затем можно оценить переменные модели

$$\theta_{d,t} = \frac{n^{(d)}_{-w,t} + \alpha}{\sum_{t' \in T} \left(n^{(d)}_{-w,t'} + \alpha \right)},$$

$$\phi_{w,t} = \frac{n^{(w)}_{-w,t} + \beta}{\sum_{w' \in W} \left(n^{(w')}_{-w,t} + \beta \right)},$$

где $\phi_{w,t}$ – вероятность получить слово w в теме t, а $\theta_{d,t}$ – вероятность получить тему t в документе d.

Варианты и расширения модели LDA

- В последние десять лет эта модель стала основой для множества различных расширений.
- Каждое из этих расширений содержит либо вариационный алгоритм вывода, либо алгоритм сэмплирования по Гиббсу для модели, которая, основываясь на LDA, включает в себя ещё и какую-либо дополнительную информацию или дополнительные предполагаемые зависимости.
- Обычно – или дополнительная структура на темах, или дополнительная информация.

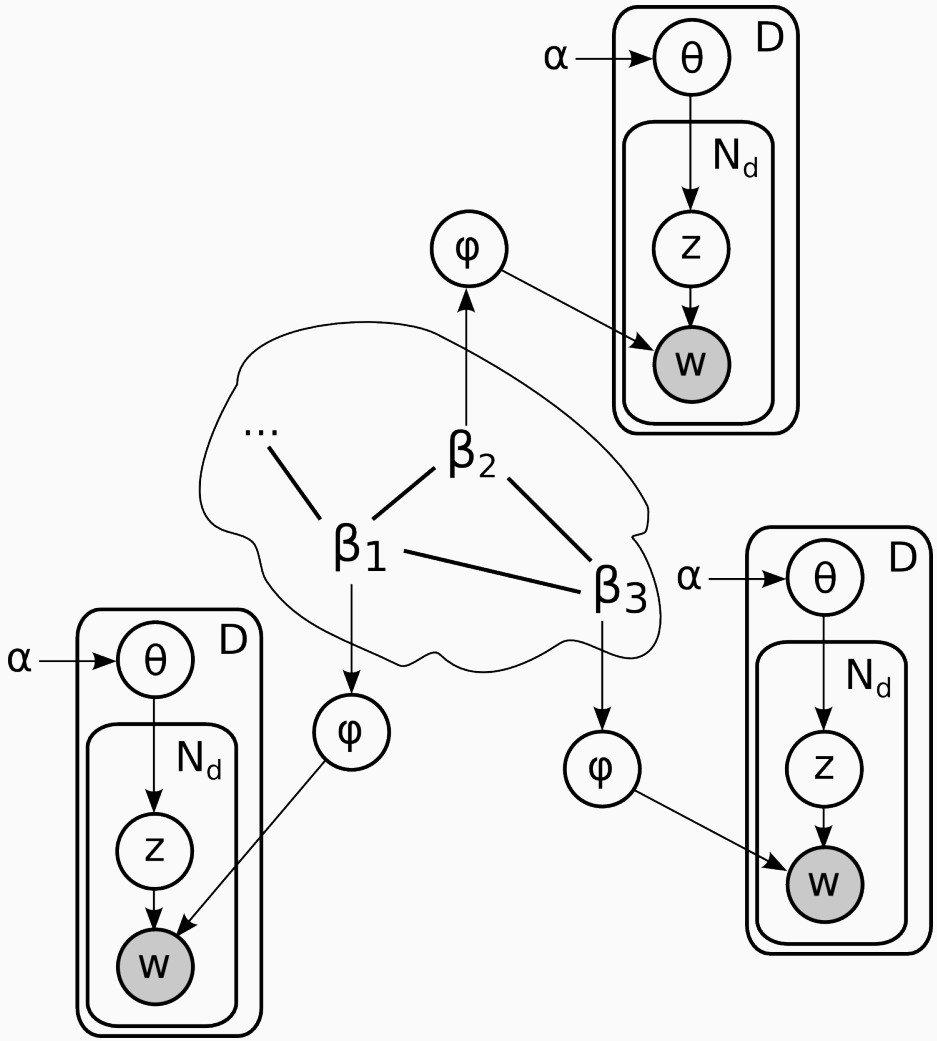
КОРРЕЛИРОВАННЫЕ ТЕМАТИЧЕСКИЕ МОДЕЛИ

- В базовой модели LDA распределения слов по темам независимы и никак не скоррелированы; однако на самом деле, конечно, некоторые темы ближе друг к другу, многие темы делят между собой слова.
- Коррелированные тематические модели (correlated topic models, CTM); отличие от базового LDA здесь в том, что используется логистическое нормальное распределение вместо распределения Дирихле; логистическое нормальное распределение более выразительно, оно может моделировать корреляции между темами.
- Предлагается алгоритм вывода, основанный на вариационном приближении.

МАРКОВСКИЕ ТЕМАТИЧЕСКИЕ МОДЕЛИ

- Марковские тематические модели (Markov topic models, MTM): марковские случайные поля для моделирования взаимоотношений между темами в разных частях датасета (разных корпусах текстов).
- MTM состоит из нескольких копий гиперпараметров β_i в LDA, описывающих параметры разных корпусов с одними и теми же темами. Гиперпараметры β_i связаны между собой в марковском случайном поле (Markov random field, MRF).
- В результате тексты из i-го корпуса порождаются как в обычном LDA, используя соответствующее β_i.
- В свою очередь, β_i подчиняются априорным ограничениям, которые позволяют «делить» темы между корпусами, задавать «фоновые» темы, присутствующие во всех корпусах, накладывать ограничения на взаимоотношения между темами и т.д.

МАРКОВСКИЕ ТЕМАТИЧЕСКИЕ МОДЕЛИ



РЕЛЯЦИОННАЯ ТЕМАТИЧЕСКАЯ МОДЕЛЬ

- Реляционная тематическая модель (relational topic model, RTM) – иерархическая модель, в которой отражён граф структуры сети документов.
- Генеративный процесс в RTM работает так:
  - сгенерировать документы из обычной модели LDA;
  - для каждой пары документов d_1, d_2 выбрать бинарную переменную y_{12}, отражающую наличие связи между d_1 и d_2:

$$y_{12} \mid \mathbf{z}_{d_1}, \mathbf{z}_{d_2} \sim \psi(\cdot \mid \mathbf{z}_{d_1}, \mathbf{z}_{d_2}, \eta).$$

- В качестве ψ берутся разные сигмоидальные функции; разработан алгоритм вывода, основанный на вариационном приближении.

Модели, учитывающие время

- Ряд важных расширений LDA касается учёта трендов, т.е. изменений в распределениях тем, происходящих со временем.
- Цель – учёт времени, анализ «горячих» тем, анализ того, какие темы быстро становятся «горячими» и столь же быстро затухают, а какие проходят «красной нитью» через весь исследуемый временной интервал.

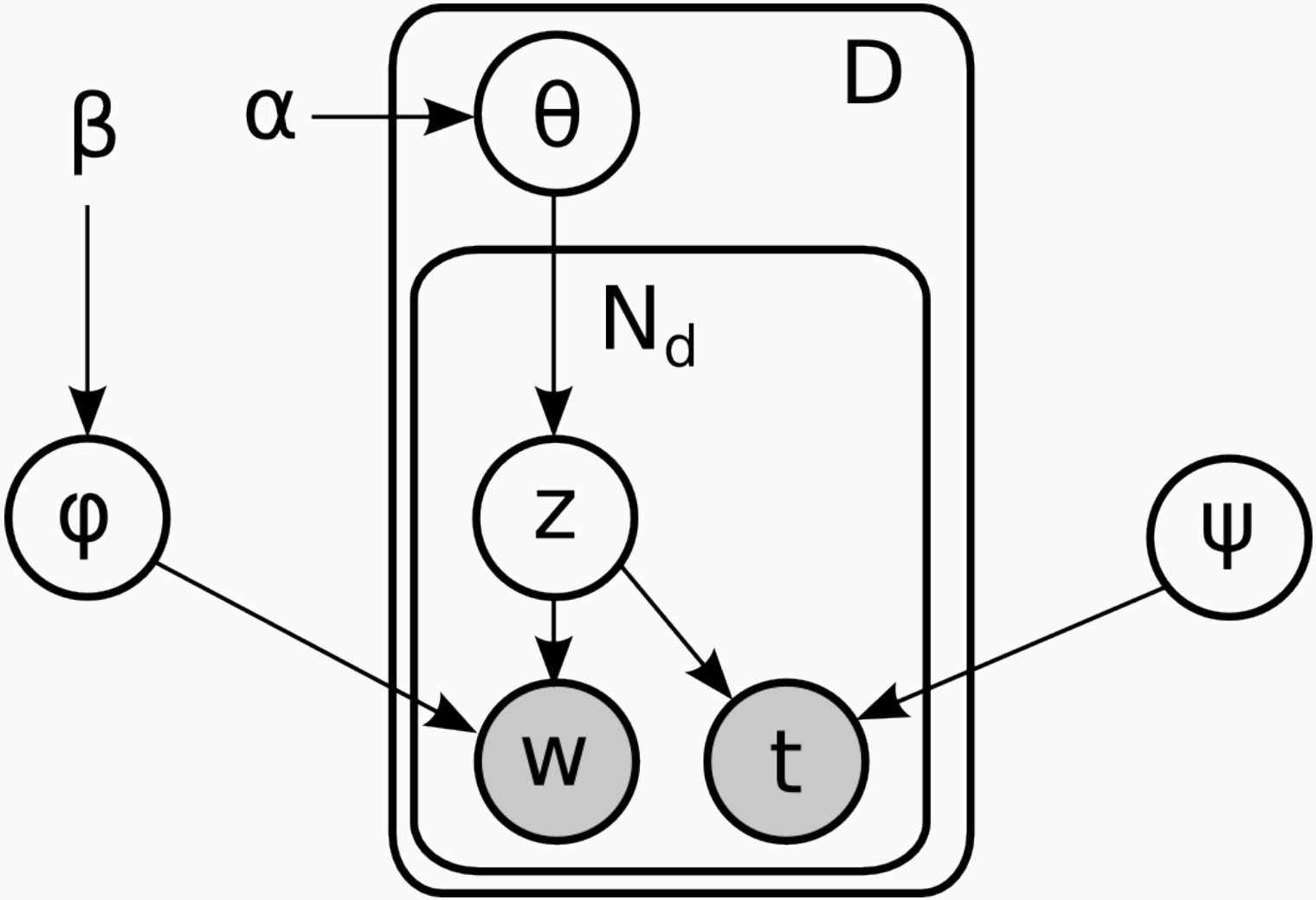
TOPICS OVER TIME

- В модели TOT (Topics over Time) время предполагается непрерывным, и модель дополняется бета-распределениями, порождающими временные метки (timestamps) для каждого слова.
- Генеративная модель модели Topics over Time такова:
  - для каждой темы $z = 1..T$ выбрать мультиномиальное распределение ϕ_z из априорного распределения Дирихле β;
  - для каждого документа d выбрать мультиномиальное распределение θ_d из априорного распределения Дирихле α, затем для каждого слова $w_{di} \in d$:
    - выбрать тему z_{di} из θ_d;
    - выбрать слово w_{di} из распределения $\phi_{z_{di}}$;
    - выбрать время t_{di} из бета-распределения $\psi_{z_{di}}$.

TOPICS OVER TIME

- Основная идея заключается в том, что каждой теме соответствует её бета-распределение ψ_z, т.е. каждая тема локализована во времени (сильнее или слабее, в зависимости от параметров ψ_z).
- Таким образом можно как обучить глобальные темы, которые всегда присутствуют, так и подхватить тему, которая вызвала сильный краткий всплеск, а затем пропала из виду; разница будет в том, что дисперсия ψ_z будет в первом случае меньше, чем во втором.

TOPICS OVER TIME



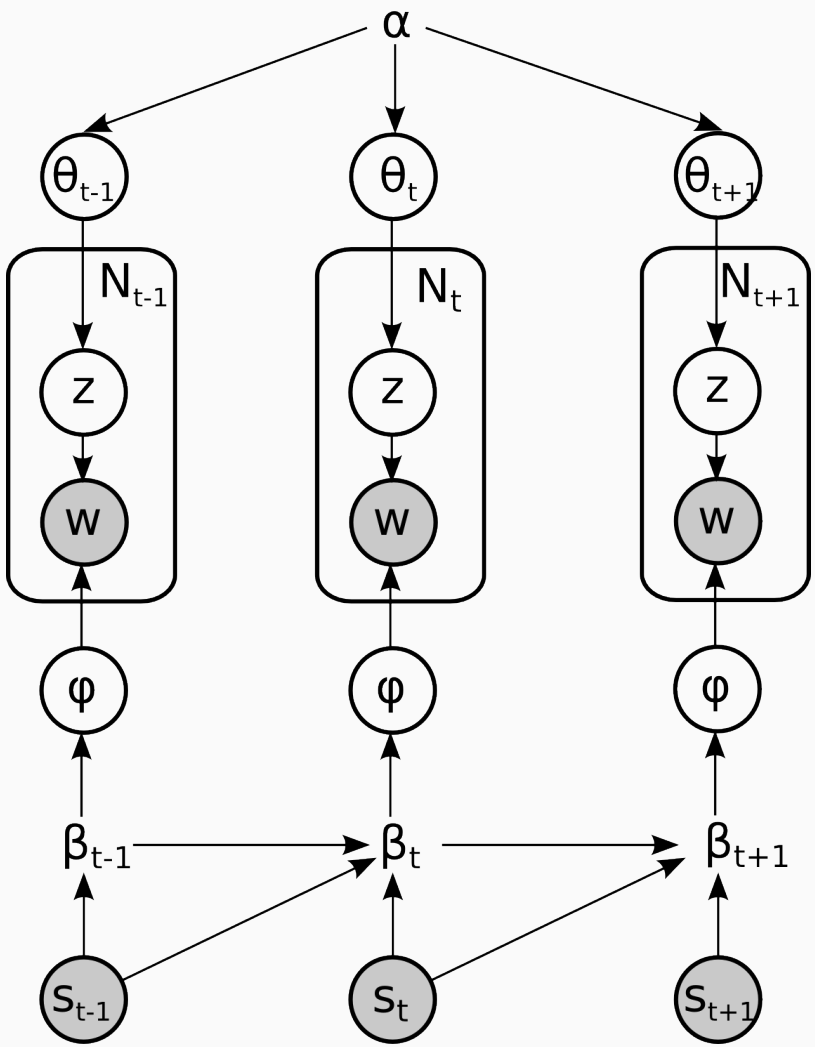
ДИНАМИЧЕСКИЕ ТЕМАТИЧЕСКИЕ МОДЕЛИ

- *Динамические тематические модели* представляют временную эволюцию тем через эволюцию их гиперпараметров α и/или β.
- Бывают дискретные ([d]DTM), в которых время дискретно, и непрерывные, где эволюция гиперпараметра β (α здесь предполагается постоянным) моделируется посредством броуновского движения: для двух документов i и j (j позже i) верно, что

$$\beta_{j,k,w} \mid \beta_{i,k,w}, s_i, s_j \sim \mathcal{N}(\beta_{i,k,w}, v\Delta_{s_i,s_j}),$$

  где s_i и s_j – это отметки времени (timestamps) документов i и j, $\Delta(s_i, s_j)$ – интервал времени, прошедший между ними, v – параметр модели.
- В остальном генеративный процесс остаётся неизменным.

Непрерывная динамическая тематическая модель (cDTM)



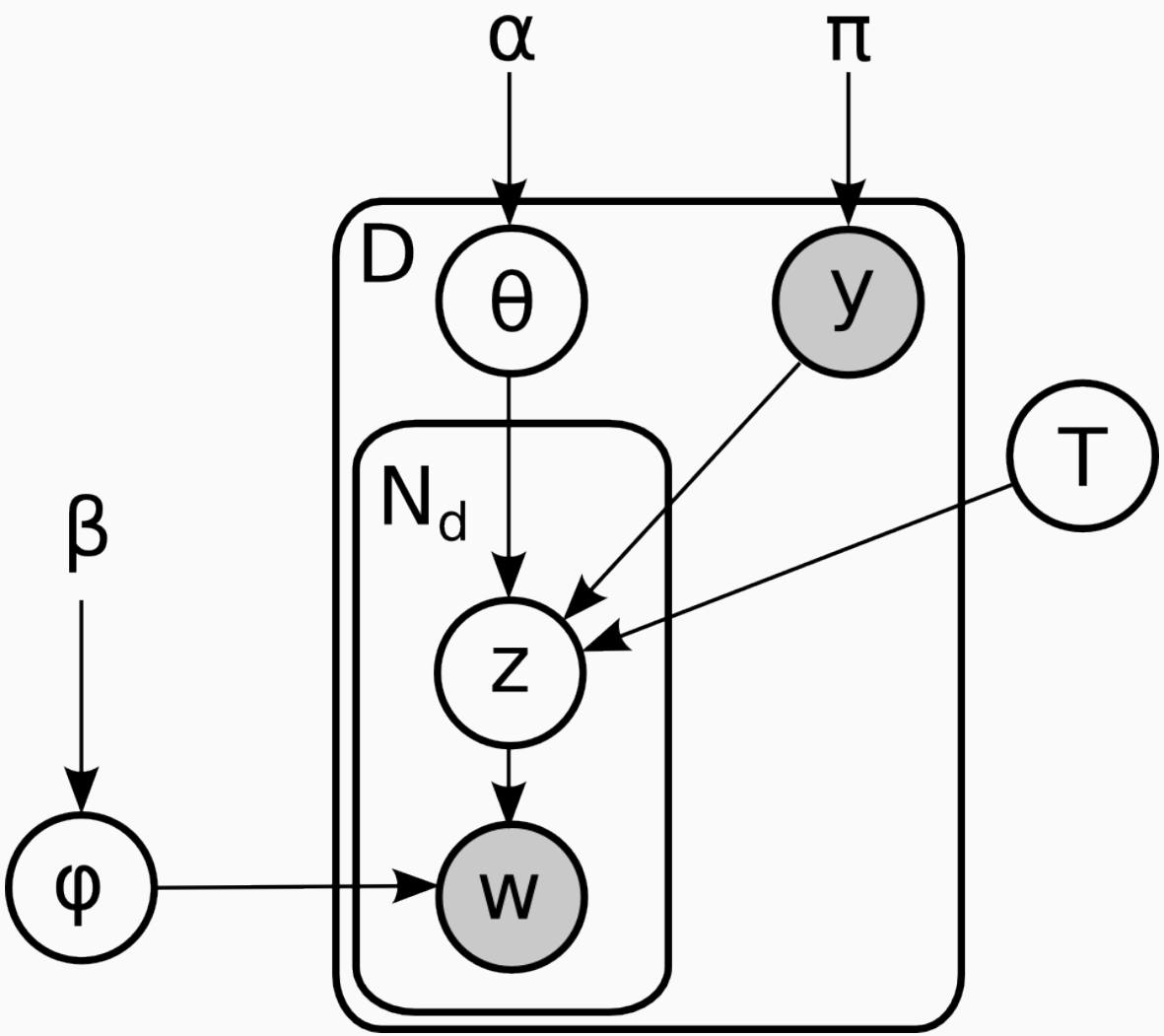
SUPERVISED LDA

- Supervised LDA: документы снабжены дополнительной информацией, дополнительной переменной отклика (обычно известной).
- Распределение отклика моделируется обобщённой линейной моделью (распределением из экспоненциального семейства), параметры которой связаны с полученным в документе распределением тем.
- Т.е. в генеративную модель добавляется ещё один шаг: после того как темы всех слов известны,
  - сгенерировать переменную–отклик $y \sim \mathrm{glm}(\mathbf{z}, \eta, \delta)$, где $\mathbf{z}$ – распределение тем в документе, а η и δ – другие параметры glm.
- К примеру, в контексте рекомендательных систем дополнительный отклик может быть реакцией пользователя.

DISCLDA

- Дискриминативное LDA (DiscLDA), другое расширение модели LDA для документов, снабжённых категориальной переменной y, которая в дальнейшем станет предметом для классификации.
- Для каждой метки класса y в модели DiscLDA вводится линейное преобразование $T^y : \mathbb{R}^K \to \mathbb{R}^L_+$, которое преобразует K-мерное распределение Дирихле θ в смесь L-мерных распределений Дирихле $T^y\theta$.
- В генеративной модели меняется только шаг порождения темы документа z: вместо того чтобы выбирать z по распределению θ, сгенерированному для данного документа,
    - сгенерировать тему z по распределению $T^y\theta$, где T^y – преобразование, соответствующее метке данного документа y.

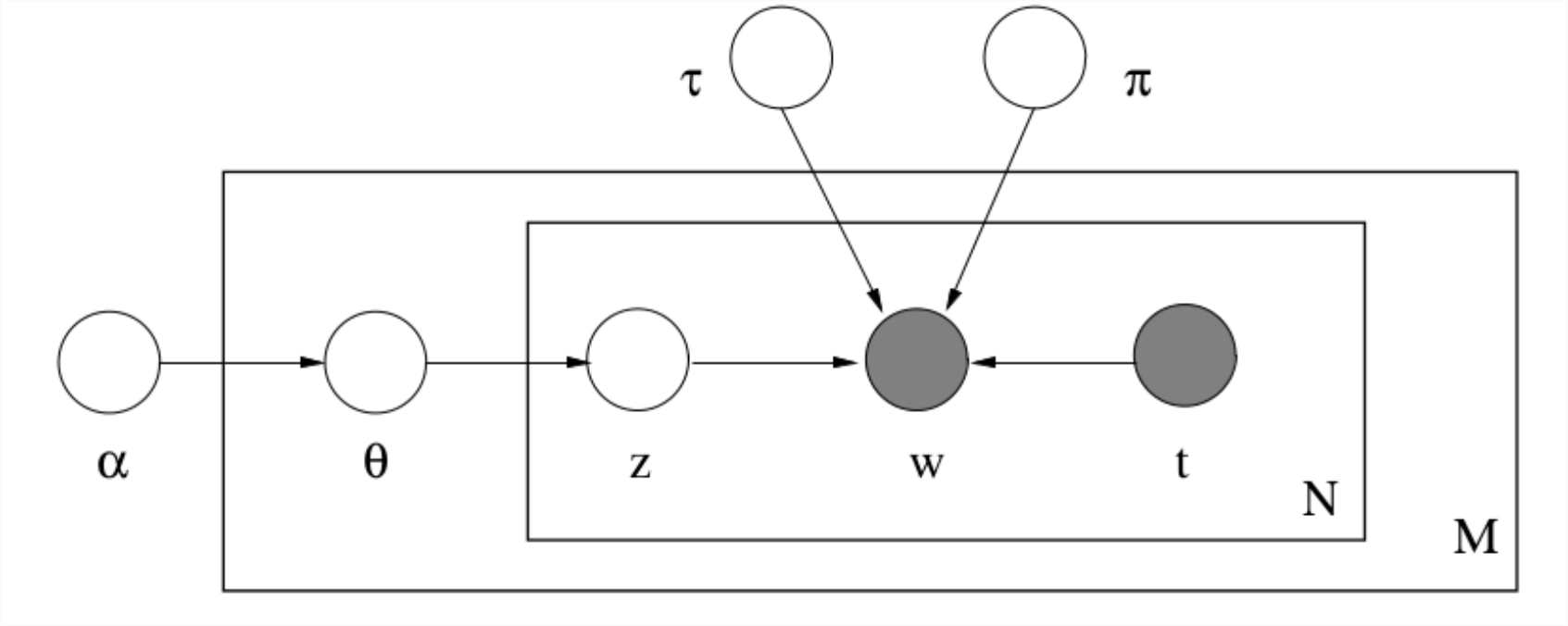
DISCLDA



TAGLDA

- TagLDA: слова имеют теги, т.е. документ не является единым мешком слов, а состоит из нескольких мешков, и в разных мешках слова отличаются друг от друга.
- Например, у страницы может быть название – слова из названия важнее для определения темы, чем просто из текста. Или, например, теги к странице, поставленные человеком – опять же, это слова гораздо более важные, чем слова из текста.
- Математически разница в том, что теперь распределения слов в темах – это не просто мультиномиальные дискретные распределения, они факторизованы на распределение слово-тема и распределение слово-тег.

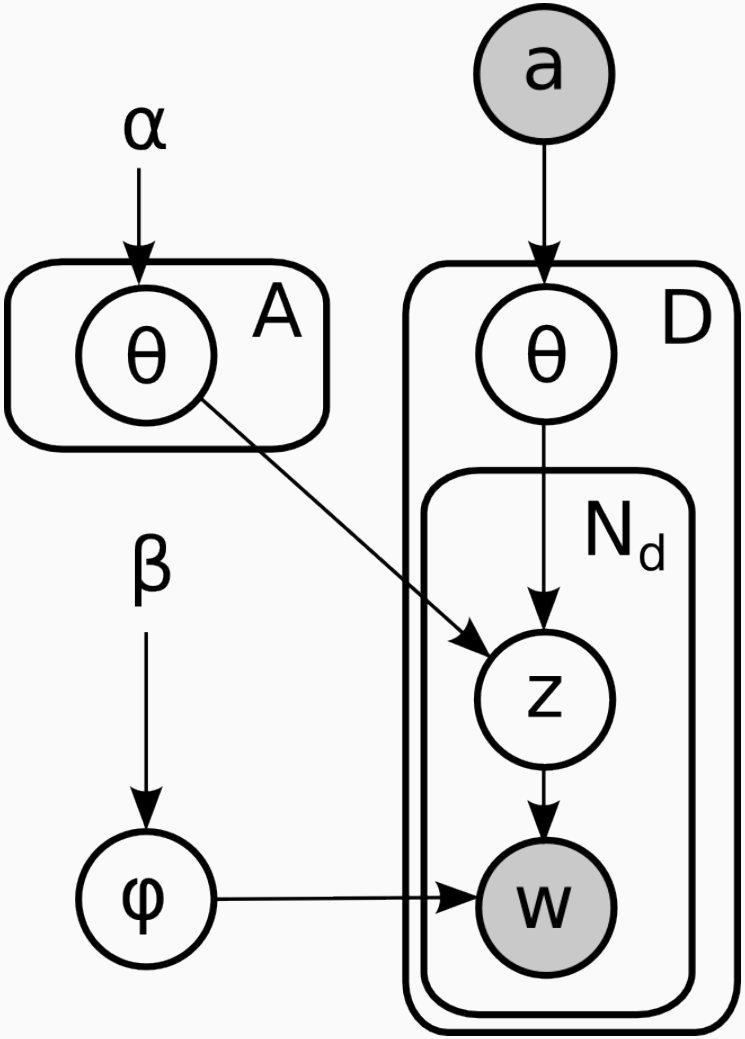
TAGLDA



AUTHOR-TOPIC MODEL

- Author-Topic modeling: кроме собственно текстов, присутствуют их авторы; или автор тоже представляется как распределение на темах, на которые он пишет, или тексты одного автора даже на разные темы будут похожи.
- Базовая генеративная модель Author-Topic model (остальное как в базовом LDA):
  - для каждого слова w:
    - выбираем автора x для этого слова из множества авторов документа a_d;
    - выбираем тему из распределения на темах, соответствующего автору x;
    - выбираем слово из распределения слов, соответствующего этой теме.

AUTHOR-TOPIC MODEL



- Алгоритм сэмплирования, соответствующий такой модели, является вариантом сжатого сэмплирования по Гиббсу:

$$p(z_w = t, x_w = a \mid \mathbf{z}_{-w}, \mathbf{x}_{-w}, \mathbf{w}, \alpha, \beta) \propto$$

$$\propto \frac{n^{(a)}_{-a,t} + \alpha}{\sum_{t' \in T} \left(n^{(a)}_{-w,t'} + \alpha \right)} \frac{n^{(w)}_{-w,t} + \beta}{\sum_{w' \in W} \left(n^{(w')}_{-w,t} + \beta \right)},$$

где $n^{(a)}_{-a,t}$ – то, сколько раз автору a соответствовала тема t, не считая текущего значения x_w, а $n^{(w)}_{-w,t}$ – число раз, которое слово w было порождено из темы t, не считая текущего значения z_w; заметим, что оба этих счётчика зависят от других переменных $\mathbf{z}_{-w}$, $\mathbf{x}_{-w}$.

КРАТКОЕ РЕЗЮМЕ

1. Основы обработки текстов: какие задачи нужно уметь решать.
2. Классический метод категоризации: наивный байесовский классификатор.
3. Обобщаем наивный байес: кластеризация EM-алгоритмом.
4. Тематическое моделирование: pLSA, LDA, расширения LDA.

СПАСИБО!

Спасибо за внимание!