ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ II

Сергей Николенко

СПбГУ — Санкт-Петербург

25 марта 2022 г.

Random facts:

- 25 марта 421 г., по легенде, в день Благовещения Девы Марии римляне, спасшиеся от готов на пустынных островах болотистого побережья, основали Венецию
- 25 марта 1238 г. началась героическая оборона Козельска, длившаяся 7 недель (дольше простоял только Киев); монголы потеряли несколько тысяч человек, а затем сровняли город с землёй, убив всех козлян, включая грудных детей
- 25 марта 1604 г. Борис Годунов послал казачьего голову Гаврилу Писемского из Сургута и стрелецкого голову Василия Тыркова из Тобольска с заданием основать крепость на берегу реки Томи в татарской земле; так появился Томск
- 25 марта 1918 г. Рада Белорусской Народной Республики приняла Третью Уставную грамоту, в которой провозглашалась независимость БНР; впрочем, части Красной армии заняли Минск уже в декабре
- 25 марта 1969 г. Джон Леннон и Йоко Оно начали акцию «В постели за мир» (Bed-In for Peace) против войны во Вьетнаме; семь дней молодожёны приглашали прессу и по 12 часов сидели в постели, призывая к миру

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).

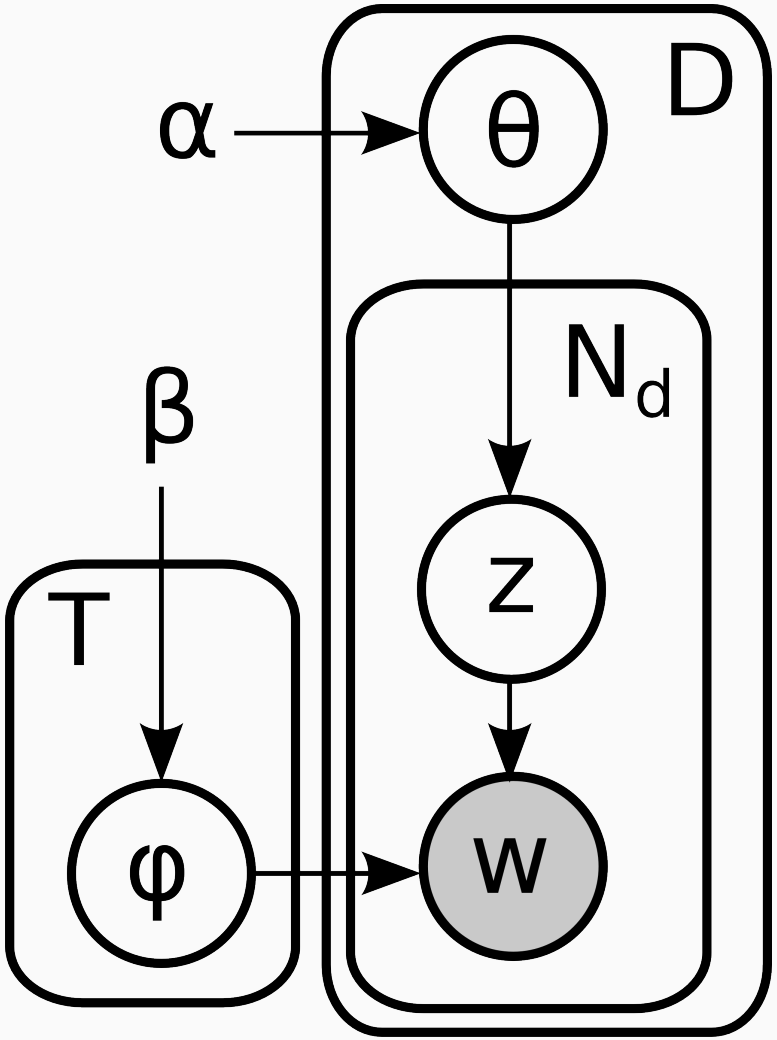
LDA

- 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.

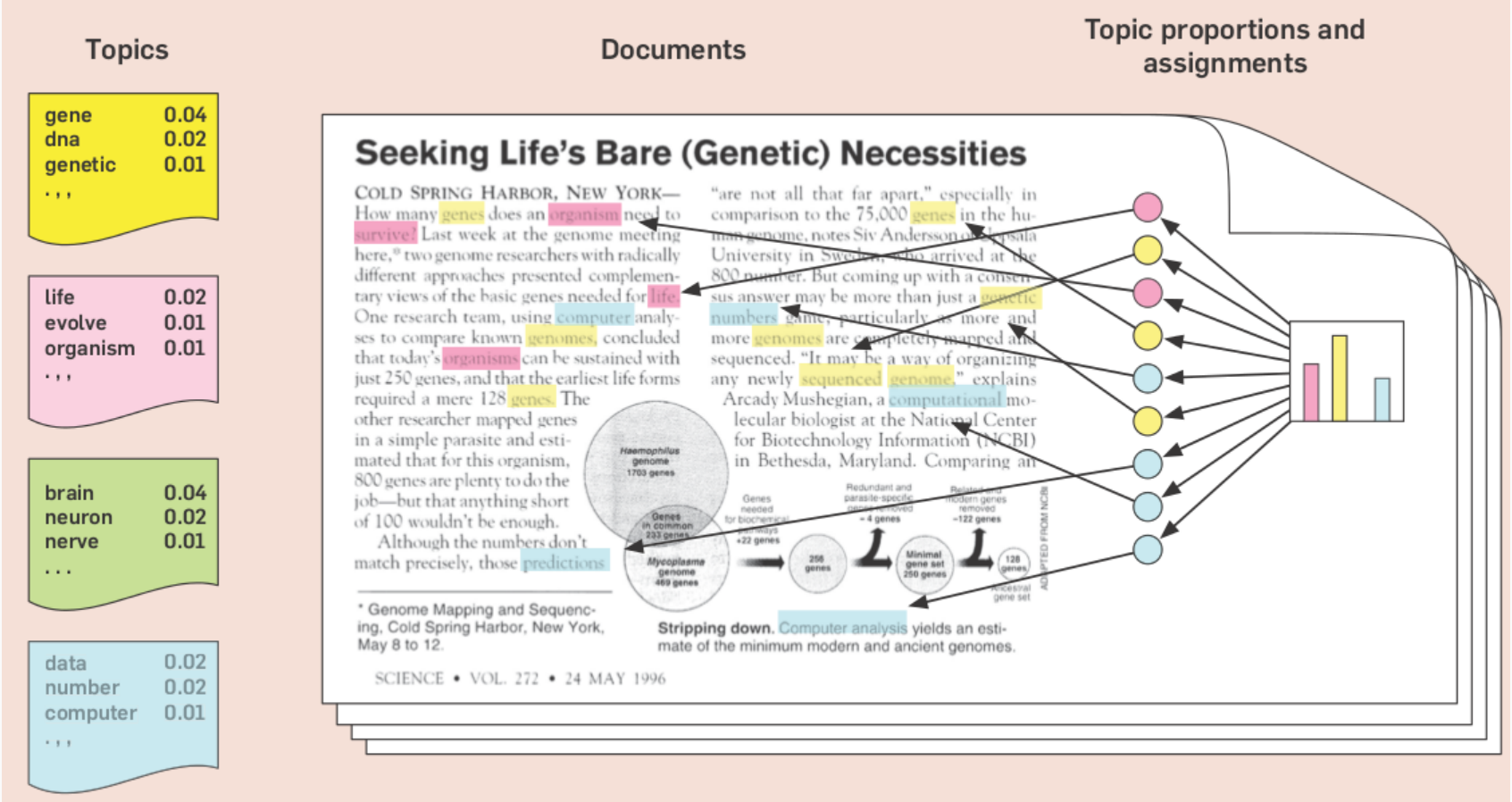
LDA

- 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 $D = \{\mathbf{w}_d \mid d = 1..M\}$.
- Генеративная модель LDA выглядит так:
  - выбрать $\theta \sim \mathrm{Di}(\alpha)$;
  - для каждого из N слов w_n:
    - выбрать тему $z_n \sim \mathrm{Mult}(\theta)$;
    - 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.

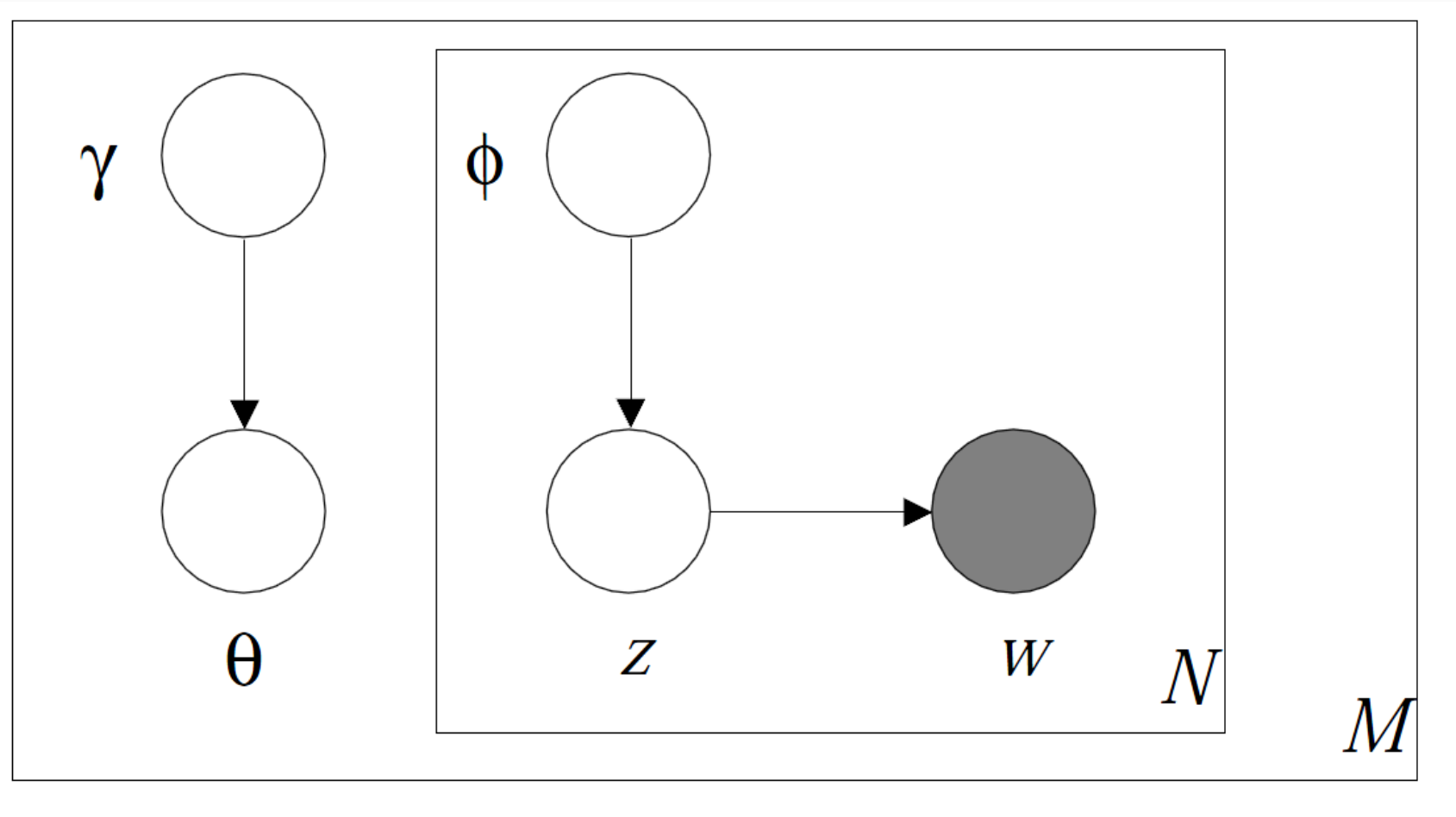
Вывод в LDA

- 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:

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

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

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

LDA: вариационный вывод

- Распишем произведения:

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

- 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

$$
\begin{aligned}
L(\gamma, \phi; \alpha, \beta) = \log\Gamma(\sum_{i=1}^{k}\alpha_i) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) + \sum_{i=1}^{k}(\alpha_i - 1)\left[\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right] + \\
+ \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\left[\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right] + \\
+ \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V} w_n^j \phi_{ni}\log\beta_{ij} - \\
- \log\Gamma(\sum_{i=1}^{k}\gamma_i) + \sum_{i=1}^{k}\log\Gamma(\gamma_i) - \sum_{i=1}^{k}(\gamma_i - 1)\left[\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right] - \\
- \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\log\phi_{ni}.
\end{aligned}
$$

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 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 $L(\gamma, \phi; \alpha, \beta)$, т.к.

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

LDA: оценка параметров

- EM-алгоритм:
  1. найти параметры $\{\gamma_d, \phi_d \mid d \in D\}$, 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.

Спасибо!

Спасибо за внимание!