ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

Сергей Николенко

СПбГУ — Санкт-Петербург

18 марта 2022 г.

Random facts:

- 18 марта 1314 г. состоялся суд над великим магистром Жаком Моле и другими тамплиерами; де Моле и приора Нормандии сожгли, остальных посадили
- 18 марта 1869 г. Н.А. Меншуткин от имени Д.И. Менделеева сообщил Русскому химическому обществу об открытии соотношения между свойствами элементов и их атомными весами, более известном как периодическая система
- 18 марта 1871 г. была создана Парижская коммуна, а 18 марта 1920 г. на II конгрессе Коминтерна В.И. Ленин напомнил о предстоящем через год 50-летии и поставил задачу: «К этому времени Франция должна стать советской республикой!»
- 18 марта 1917 г. в «Вестнике Временного правительства» были опубликованы Акты об отречении Государя Императора Николая II и Великого Князя Михаила Александровича, Николай II был арестован, началось всероссийское совещание Советов рабочих и солдатских депутатов и снова начала выходить запрещённая газета «Правда»
- 18 марта 1965 г. Алексей Леонов впервые в истории человечества вышел в открытый космос, а каждый член группы «Rolling Stones» был оштрафован на 5 фунтов за публичное справление малой нужды

Тематическое моделирование

Как ещё можно обобщить наивный байес

- В прошлый раз мы много говорили о наивном байесе и начали его обобщать.
- Пока обобщили на обучение без учителя (кластеризацию).
- А ещё в наивном байесе у документа только одна тема.
- Но это же не так! На самом деле документ говорит о многих темах (но не слишком многих).
- Давайте попробуем это учесть.

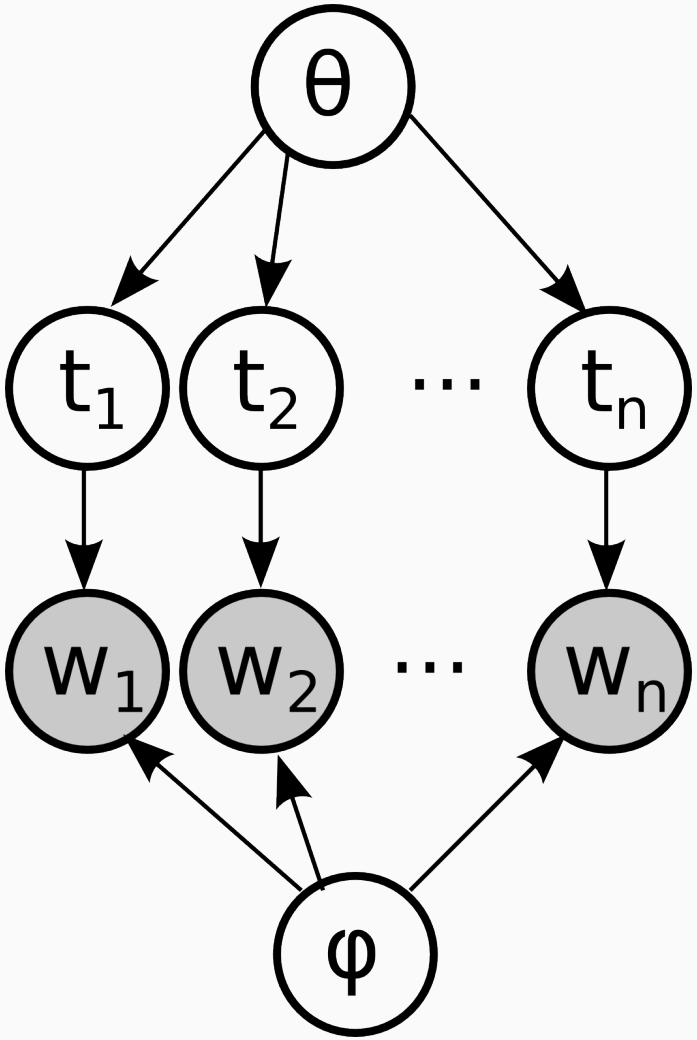
pLSA

- Рассмотрим такую модель:
  - каждое слово в документе d порождается некоторой темой $t \in T$;
  - документ порождается некоторым распределением на темах $p(t \mid d)$;
  - слово порождается именно темой, а не документом: $p(w \mid d, t) = p(w \mid d)$;
  - итого получается такая функция правдоподобия:

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

- Эта модель называется probabilistic latent semantic analysis, pLSA (Hoffmann, 1999).

pLSA: графическая модель документа



Пример

- Получается как-то так:

Алгоритм 2. Рациональный EM-алгоритм для тематической модели (2).

Вход: коллекция D, число тем $|T|$, начальные приближения матриц Φ и Θ;
Выход: параметры модели Φ и Θ;

повторять
  обнулить n_{wt}, n_{td}, n_t для всех $d \in D,\ w \in W,\ t \in T$;
  для всех $d \in D,\ w \in d$
    $n_{tdw} := n_{dw}\varphi_{wt}\theta_{td} \,/\, \sum_\tau \varphi_{w\tau}\theta_{\tau d}$ для всех $t \in T$;
    увеличить n_{wt}, n_{td}, n_t на n_{tdw} для всех $t \in T$;
  $\varphi_{wt} := n_{wt}/n_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := n_{td}/n_d$ для всех $d \in D,\ t \in T$;
пока Φ и Θ не сойдутся;

pLSA

- Как её обучать? Мы можем оценить $p(w \mid d) = \frac{n_{wd}}{n_d}$, а нужно найти:
  - $\phi_{wt} = p(w \mid t)$;
  - $\theta_{td} = p(t \mid d)$.
- Максимизируем правдоподобие

$$p(D) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}} = \prod_{d \in D} \prod_{w \in d} \left[\sum_{t \in T} p(w \mid t) p(t \mid d) \right]^{n_{dw}}.$$

- Как максимизировать такие правдоподобия?

pLSA

- EM-алгоритмом. На E-шаге ищем, сколько слов w в документе d из темы t:

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

- А на M-шаге пересчитываем параметры модели:

$$n_{wt} = \sum_d n_{dwt}, \qquad n_t = \sum_w n_{wt}, \quad \phi_{wt} = \frac{n_{wt}}{n_t},$$
$$n_{td} = \sum_{w \in d} n_{dwt}, \quad \theta_{td} = \frac{n_{td}}{n_d}.$$

- Вот и весь вывод в pLSA.

- Можно даже не хранить всю матрицу n_{dwt}, а двигаться по документам, каждый раз добавляя n_{dwt} сразу к счётчикам n_{wt}, n_{td}.

Алгоритм 2. Рациональный EM-алгоритм для тематической модели (2).

Вход: коллекция D, число тем $|T|$, начальные приближения матриц Φ и Θ;
Выход: параметры модели Φ и Θ;

повторять
  обнулить n_{wt}, n_{td}, n_t для всех $d \in D,\ w \in W,\ t \in T$;
  для всех $d \in D,\ w \in d$
    $n_{tdw} := n_{dw}\varphi_{wt}\theta_{td} \,/\, \sum_\tau \varphi_{w\tau}\theta_{\tau d}$ для всех $t \in T$;
    увеличить n_{wt}, n_{td}, n_t на n_{tdw} для всех $t \in T$;
  $\varphi_{wt} := n_{wt}/n_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := n_{td}/n_d$ для всех $d \in D,\ t \in T$;
пока Φ и Θ не сойдутся;

pLSA

- Чего тут не хватает?
  - Во-первых, разложение такое, конечно, будет сильно не единственным.
  - Во-вторых, параметров очень много, явно будет оверфиттинг, если корпус не на порядки больше числа тем.
  - А совсем хорошо было бы получать не просто устойчивое решение, а обладающее какими-нибудь заданными хорошими свойствами.
- Всё это мы можем решить как?

- Правильно, регуляризацией. Есть целая наука о разных регуляризаторах для pLSA (К.В. Воронцов).
- В общем виде так: добавим регуляризаторы R_i в логарифм правдоподобия:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta).$$

- Тогда в EM-алгоритме на M-шаге появятся частные производные R:

$$n_{wt} = \left[\sum_{d \in D} n_{dwt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right]_+,$$

$$n_{td} = \left[\sum_{w \in d} n_{dwt} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right]_+$$

- Чтобы доказать, EM надо рассмотреть как решение задачи оптимизации через условия Каруша-Куна-Такера.

pLSA

- И теперь мы можем кучу разных регуляризаторов вставить в эту модель:
  - регуляризатор сглаживания (позже, это примерно как LDA);
  - регуляризатор разреживания: максимизируем KL-расстояние между распределениями ϕ_{wt} и θ_{td} и равномерным распределением;
  - регуляризатор контрастирования: минимизируем ковариации между векторами ϕ_t, чтобы в каждой теме выделилось своё лексическое ядро (характерные слова);
  - регуляризатор когерентности: будем награждать за слова, которые в документах стоят ближе друг к другу;
  - и так далее, много всего можно придумать.

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).

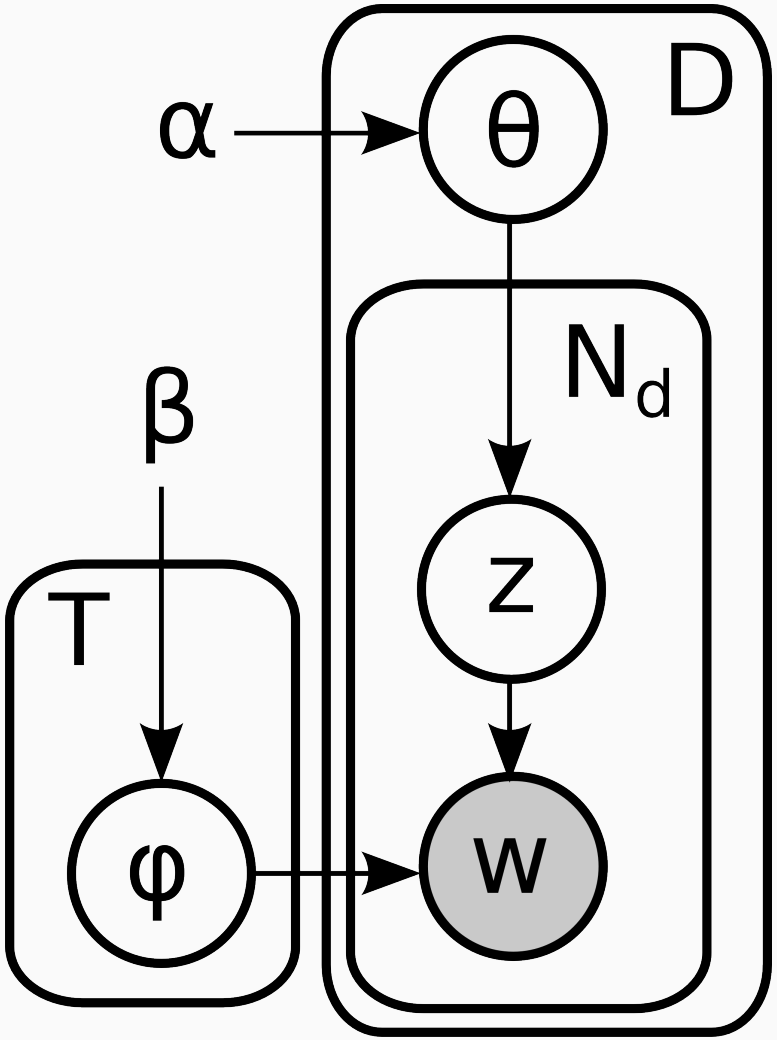
LDA

- 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.

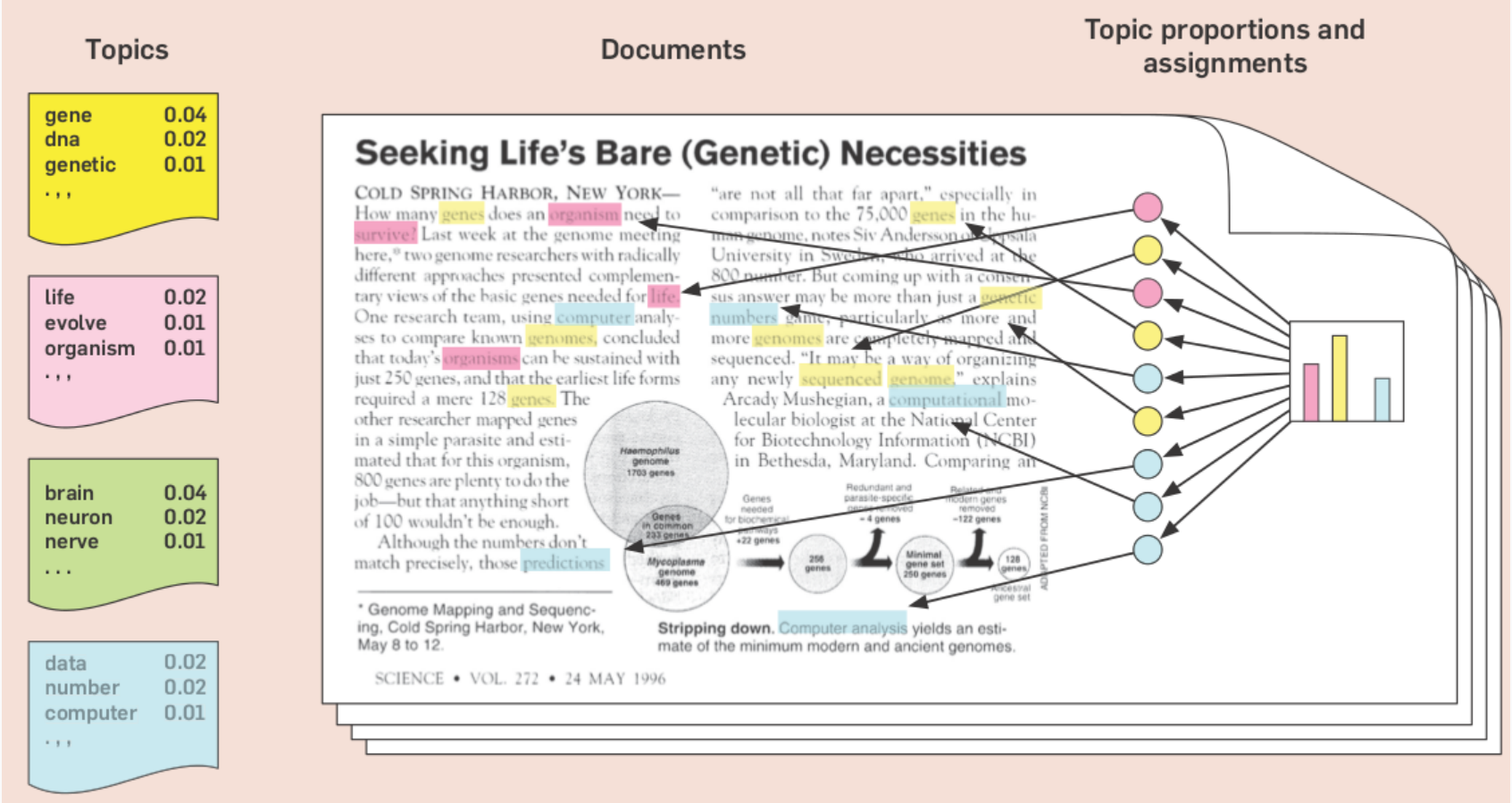
LDA

- 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.

Спасибо!

Спасибо за внимание!