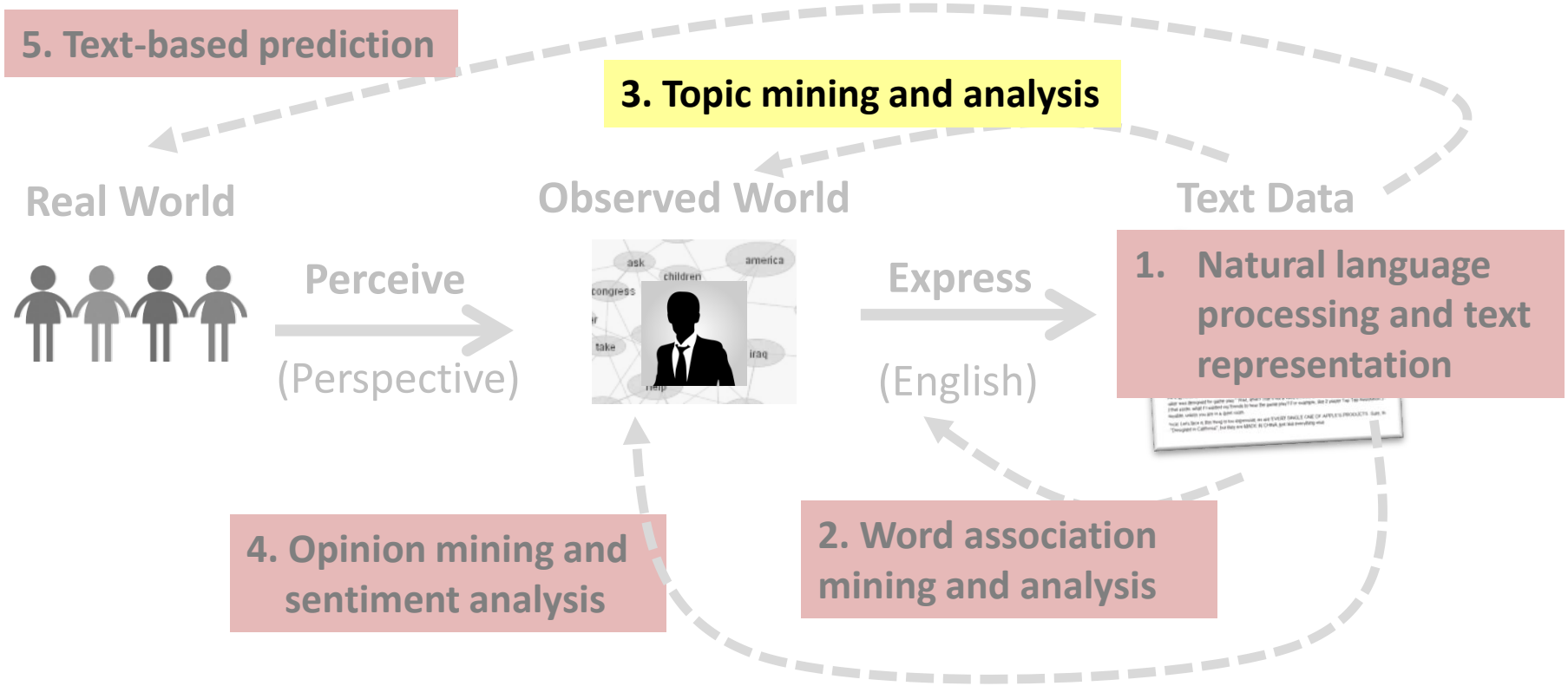
Topic Mining and Analysis: Mining One Topic

ChengXiang “Cheng” Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Probabilistic Topic Models: Mining One Topic



Simplest Case of Topic Model: Mining One Topic

INPUT: C={d}, V

OUTPUT: { θ}

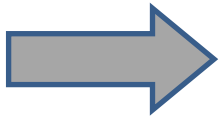
Text Data

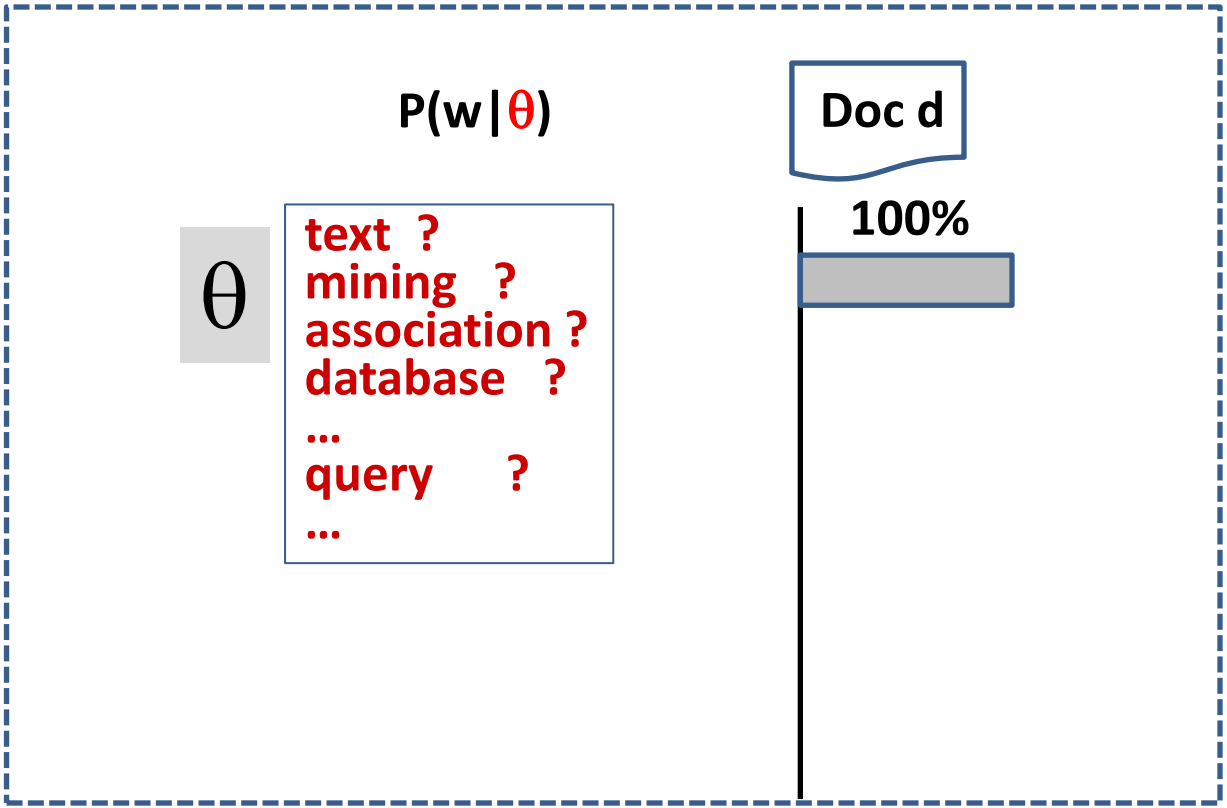

thing. I am still considering taking it back. Here's why:

Speaker quality is ABSOLUTELY HORRENDOUS. The speaker is simply not functional, unless you are in perfec ecording studio. When you turn it up all the way (because you can't hear it) it becomes fuzzy, sounding like a blowe eaker. What is this thing, the size of a flea? And if there is ANY background noise, you cannot hear it at all with the ume up (in case you're wondering, no, the speaker did not fail. I exchanged the iPod, and the second one sounde same: really bad). I can't share videos, share music, share anything, unless my friends and I are in a wine cellar.

heard excuses for the speaker's embarrassing quality, including "it's a small device". My cheap Samsung cell one's speaker is 4x better than the iPod Touch, and the phone is smaller. The other excuse I've heard is "The eaker was designed for game play." Wait, what? That's not a valid excuse, because the IPOD is a MUSIC device d that aside, what if I wanted my friends to hear the game play? For example, like 2 player Tip Tap Revolution 3 i playable, unless you are in a quiet room.

Price: Let's face it, this thing is too expensive, as are EVERY SINGLE ONE OF APPLE'S PRODUCTS. Sure, the s "Designed in California", but they are MADE IN CHINA, just like everything else.





Language Model Setup

- **Data**: Document d= $x_1\ x_2\ \ldots\ x_{|d|}$, $x_i \in V=\{w_1, \ldots, w_M\}$ is a word
- **Model**: Unigram LM θ(=topic) : $\{\theta_i=p(w_i|\theta)\}$, i=1, …, M; $\theta_1+\ldots+\theta_M=1$
- **Likelihood** function:

$$p(d|\theta) = p(x_1|\theta)\times\ldots\times p(x_{|d|}|\theta)$$
$$= p(w_1|\theta)^{c(w_1,d)}\times\ldots\times p(w_M|\theta)^{c(w_M,d)}$$
$$= \prod_{i=1}^{M} p(w_i|\theta)^{c(w_i,d)} = \prod_{i=1}^{M}\theta_i^{c(w_i,d)}$$

- ML **estimate:** $(\hat{\theta}_1,\ldots,\hat{\theta}_M) = \arg\max_{\theta_1,\ldots,\theta_M} p(d|\theta) = \arg\max_{\theta_1,\ldots,\theta_M}\prod_{i=1}^{M}\theta_i^{c(w_i,d)}$

Computation of Maximum Likelihood Estimate

Maximize p(d|θ) $(\hat{\theta}_1,...,\hat{\theta}_M) = \arg\max_{\theta_1,...,\theta_M} p(d|\theta) = \arg\max_{\theta_1,...,\theta_M} \prod_{i=1}^{M} \theta_i^{c(w_i,d)}$

Max. Log-Likelihood $(\hat{\theta}_1,...,\hat{\theta}_M) = \arg\max_{\theta_1,...,\theta_M} \log[p(d|\theta)] = \arg\max_{\theta_1,...,\theta_M} \sum_{i=1}^{M} c(w_i,d)\log\theta_i$

Subject to constraint: $\sum_{i=1}^{M} \theta_i = 1$

Use Lagrange multiplier approach

Lagrange function: $f(q|d) = \sum_{i=1}^{M} c(w_i,d)\log q_i + \lambda(\sum_{i=1}^{M} q_i - 1)$

$$\frac{\partial f(q|d)}{\partial q_i} = \frac{c(w_i,d)}{q_i} + \lambda = 0 \rightarrow q_i = -\frac{c(w_i,d)}{\lambda}$$

$$\sum_{i=1}^{M} -\frac{c(w_i,d)}{\lambda} = 1 \rightarrow \lambda = -\sum_{i=1}^{N} c(w_i,d) \rightarrow \hat{q}_i = p(w_i|\hat{q}) = \frac{c(w_i,d)}{\sum_{i=1}^{M} c(w_i,d)} = \frac{c(w_i,d)}{|d|}$$

Normalized Counts

What Does the Topic Look Like?