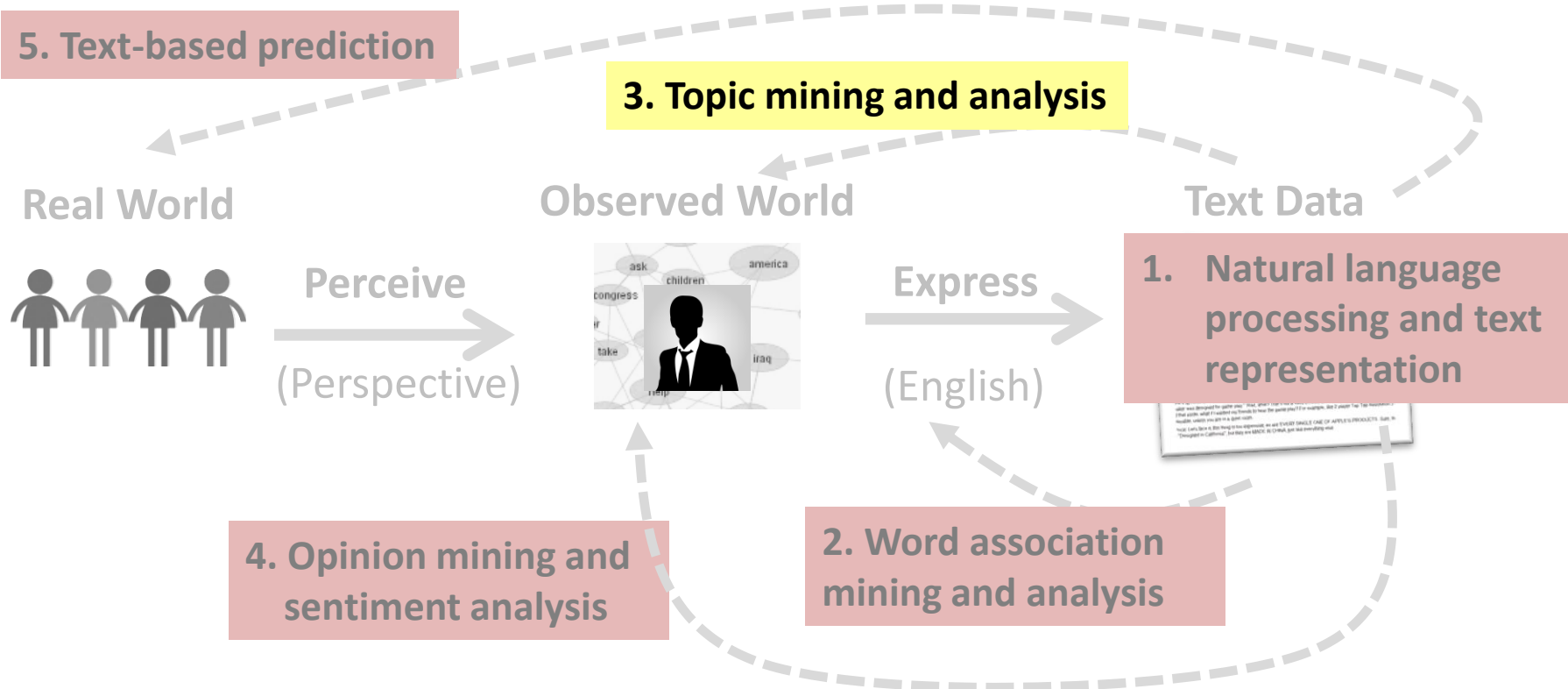




Topic Mining and Analysis: Overview of Statistical Language Models

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

Probabilistic Topic Models: Overview of Statistical Language Models



What Is a Statistical Language Model (LM)?

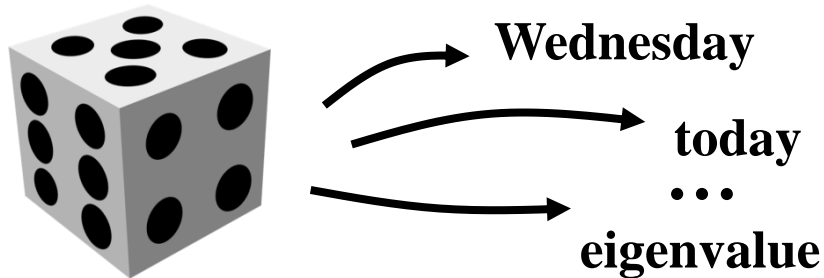
- A probability distribution over word sequences
 - $p(\text{"Today is Wednesday"}) \approx 0.001$
 - $p(\text{"Today Wednesday is"}) \approx 0.00000000000001$ 不符合语法
 - $p(\text{"The eigenvalue is positive"}) \approx 0.00001$
- Context-dependent!
- Can also be regarded as a probabilistic mechanism for "generating" text – thus also called a "generative" model



→ **Today is Wednesday**
→ Today Wednesday is
→ ...
→ **The eigenvalue is positive**

The Simplest Language Model: Unigram LM

- Generate text by generating each word INDEPENDENTLY
- Thus, $p(w_1 w_2 \dots w_n) = p(w_1)p(w_2)\dots p(w_n)$
- Parameters: $\{p(w_i)\}$ $p(w_1) + \dots + p(w_N) = 1$ (N is voc. size)
- Text = sample drawn according to this **word distribution**



$$\begin{aligned} p(\text{"today is Wed"}) \\ &= p(\text{"today"})p(\text{"is"})p(\text{"Wed"}) \\ &= 0.0002 \times 0.001 \times 0.000015 \end{aligned}$$

Text Generation with Unigram LM

Unigram LM $p(w|\theta)$

Sampling



Document d

$p(d|\theta)=?$

Topic 1:
Text mining

...
text 0.2
mining 0.1
association 0.01
clustering 0.02
...
food 0.00001
...



**Text mining
paper**

Topic 2:
Health

...
food 0.25
nutrition 0.1
healthy 0.05
diet 0.02
...



**Food nutrition
paper**

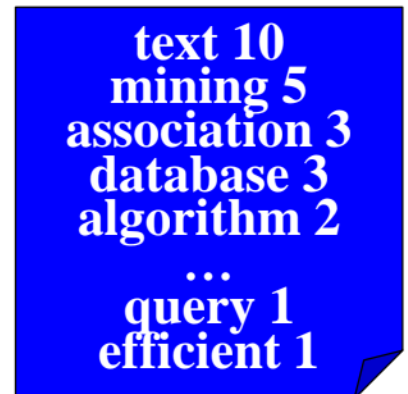
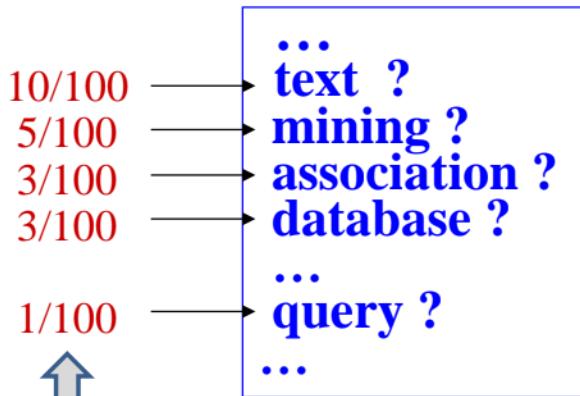
Estimation of Unigram LM

Unigram LM $p(w|\theta)=?$

Estimation

Text Mining Paper d

Total #words=100



从被观察数据中,使概率达到最大值.

Is this our best estimate?
How do we define "best"?

Maximum Likelihood vs. Bayesian

- Maximum likelihood estimation

- “Best” means “data likelihood reaches maximum”

$$\hat{\theta} = \arg \max_{\theta} P(X | \theta)$$

- Problem: Small sample

- Bayesian estimation:

- “Best” means being consistent with our “prior” knowledge and explaining data well

$$\hat{\theta} = \arg \max_{\theta} P(\theta | X) = \arg \max_{\theta} P(X | \theta) P(\theta)$$

- Problem: How to define prior?

Maximum a Posteriori (MAP) estimate

返回已故取得最大值时的参数。

观察到Y之后再对X的检验概率。

后验概率

先验概率。

对特定的X, 观察到Y的概率。

Bayes Rule

$$p(X | Y) = \frac{p(Y | X) p(X)}{p(Y)}$$

最大后验估计

Illustration of Bayesian Estimation

Bayesian inference: $f(\theta)=?$

$$\hat{f}(\theta) = \sum_{\theta} f(\theta)p(\theta | X)$$

Posterior Mean

$$\hat{\theta} = \sum_{\theta} \theta * p(\theta | X)$$

Posterior:

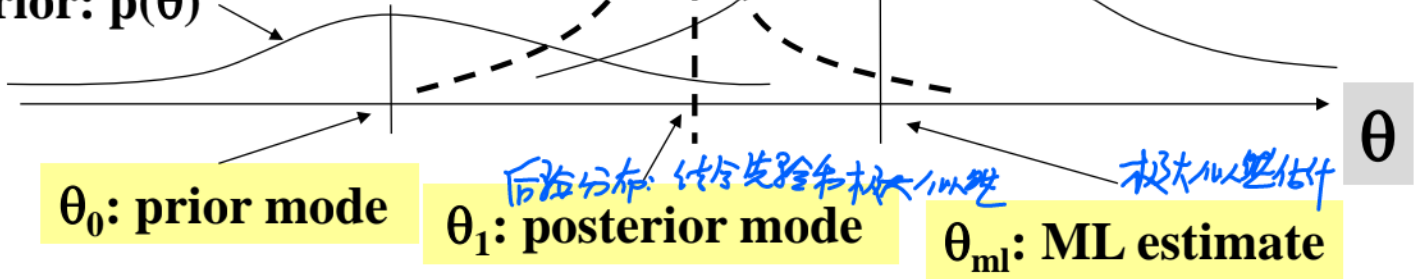
$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

Likelihood:

$$p(X|\theta)$$

$$X=(x_1, \dots, x_N)$$

Prior: $p(\theta)$



Summary

- **Language Model** = probability distribution over text = generative model for text data
- **Unigram Language Model** = **word distribution** 最简单的LH.
- **Likelihood function: $p(X|\theta)$**
 - Given $\theta \rightarrow$ which X has a higher likelihood?
 - Given $X \rightarrow$ which θ maximizes $p(X|\theta)$? [ML estimate] 最大似然估计
- **Bayesian estimation/inference**
 - Must define a **prior: $p(\theta)$**
 - **Posterior distribution: $p(\theta|X) \propto p(X|\theta)p(\theta)$**
 - ➔ Allows for inferring any “derived value” from θ !