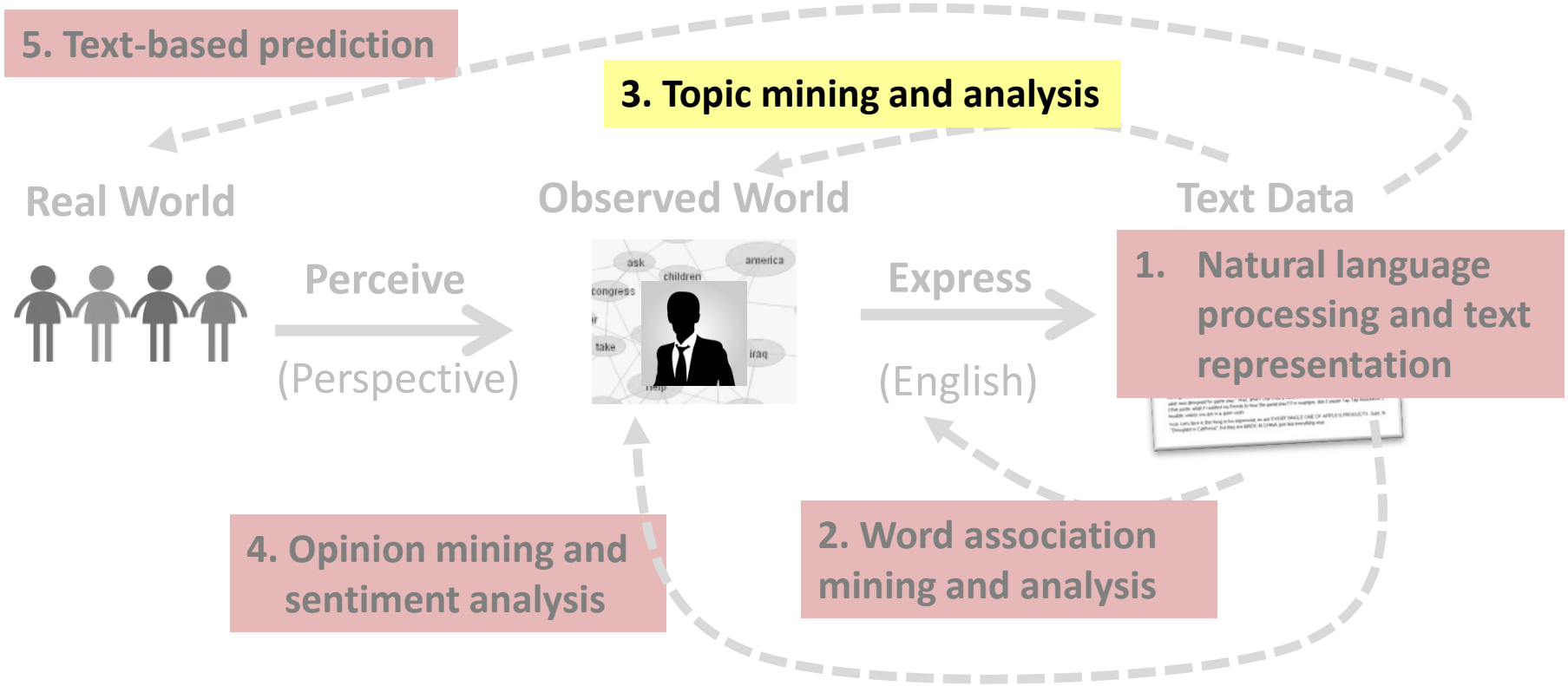
Topic Mining and Analysis: Overview of Statistical Language Models

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("*Today is Wednesday*") ≈ 0.001
  - p("*Today Wednesday is*") ≈ 0.0000000000001 不符合语法
  - p("*The eigenvalue is positive*") ≈ 0.00001
- Context-dependent!
- Can also be regarded as a probabilistic mechanism for "generating" text – thus also called a "generative" model

Today is Wednesday

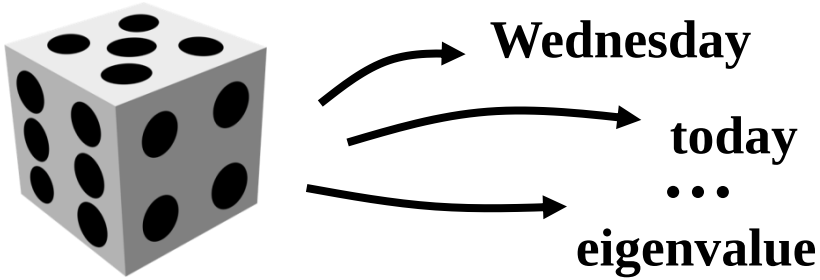
Today Wednesday is

• • •

The eigenvalue is positive

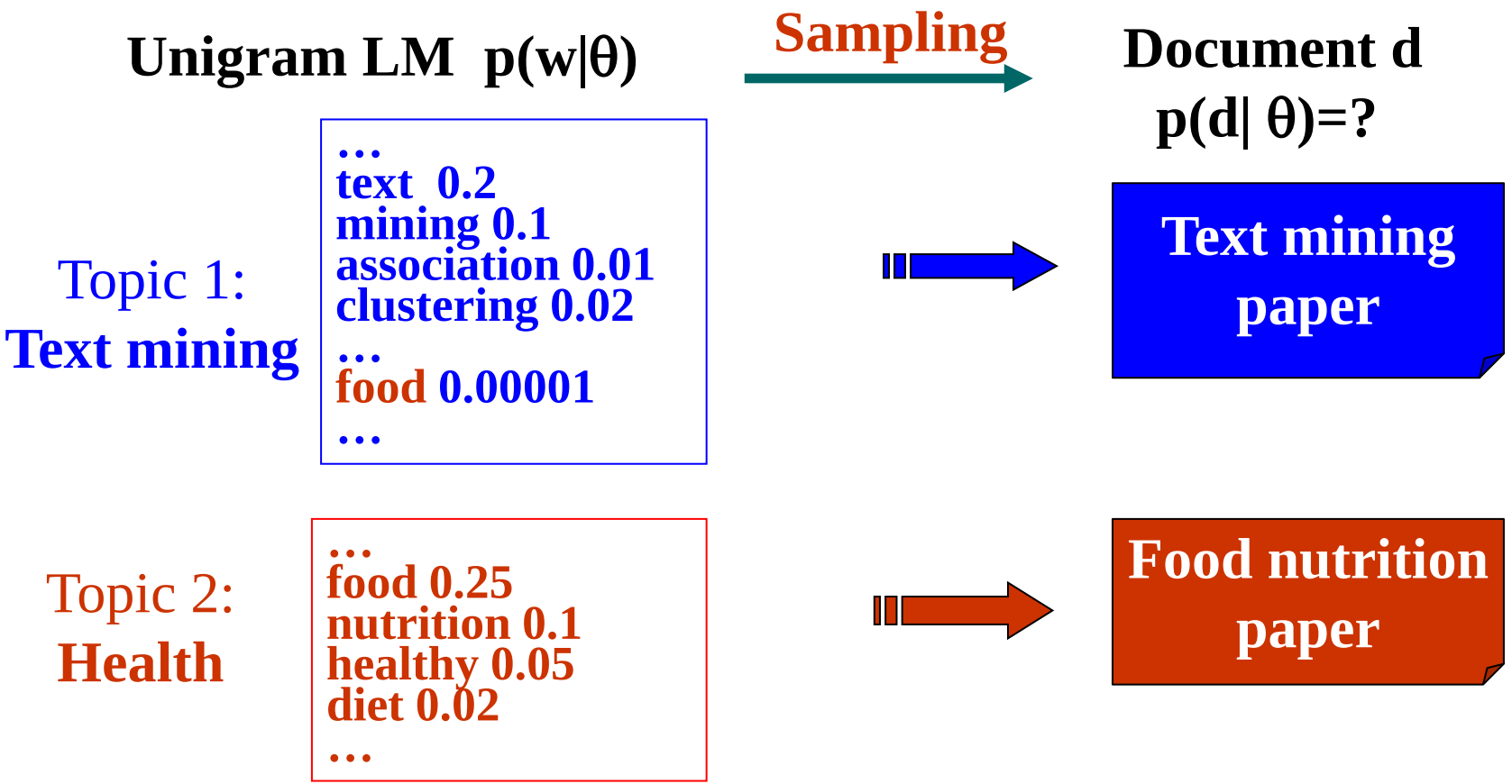
The Simplest Language Model: Unigram LM

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

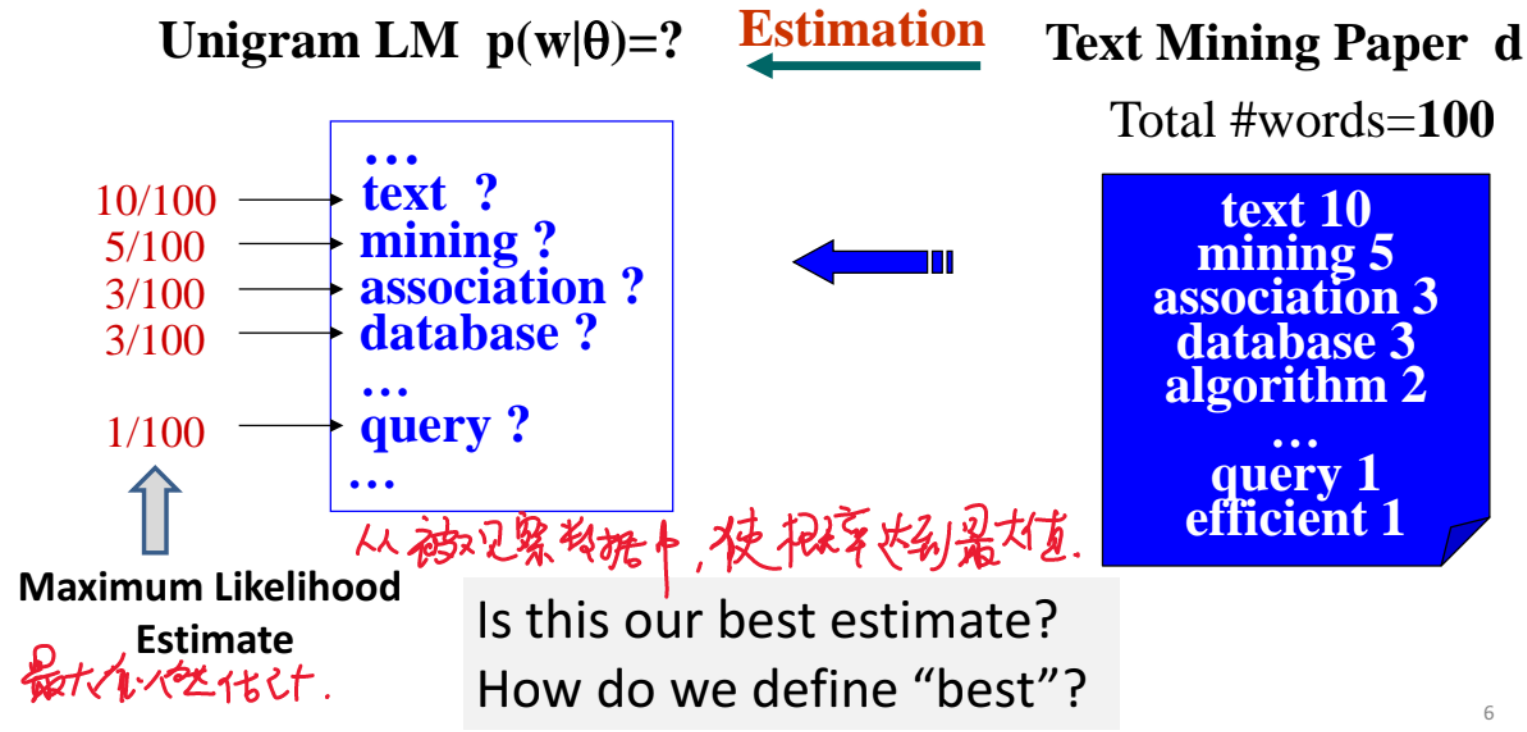


p("today is Wed")
= p("today")p("is")p("Wed")
= 0.0002 × 0.001 × 0.000015

Text Generation with Unigram LM



Estimation of Unigram LM



Is this our best estimate?

How do we define "best"?

Maximum Likelihood vs. Bayesian

- Maximum likelihood estimation
  - "Best" means "data likelihood reaches maximum"

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

  - Problem: Small sample
- Bayesian estimation:

Bayes Rule

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

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

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

  - Problem: How to define prior?

Maximum a Posteriori (MAP) estimate

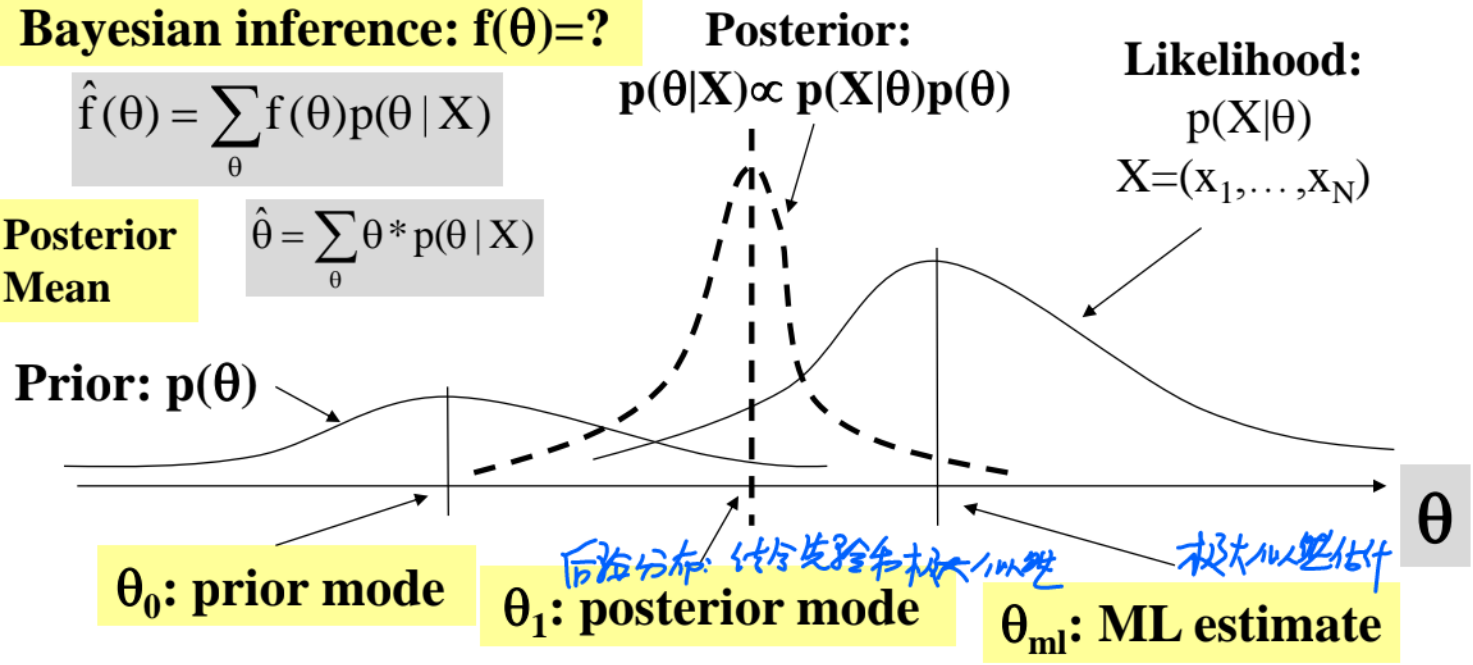
Illustration of Bayesian Estimation

Bayesian inference: f(θ)=?

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

Posterior Mean

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



Summary

- **Language Model** = probability distribution over text = generative model for text data
- **Unigram** Language Model = **word distribution**
- **Likelihood** function: **p(X|θ)**
  - **Given θ ➔** which X has a higher likelihood?
  - **Given X ➔** which θ maximizes p(X| θ)? **[ML estimate]**
- **Bayesian** estimation/inference
  - Must define a **prior: p(θ)**
  - **Posterior** distribution**: p(θ|X)∝ p(X|θ)p(θ)**

  ➔ Allows for inferring any "derived value" from θ!