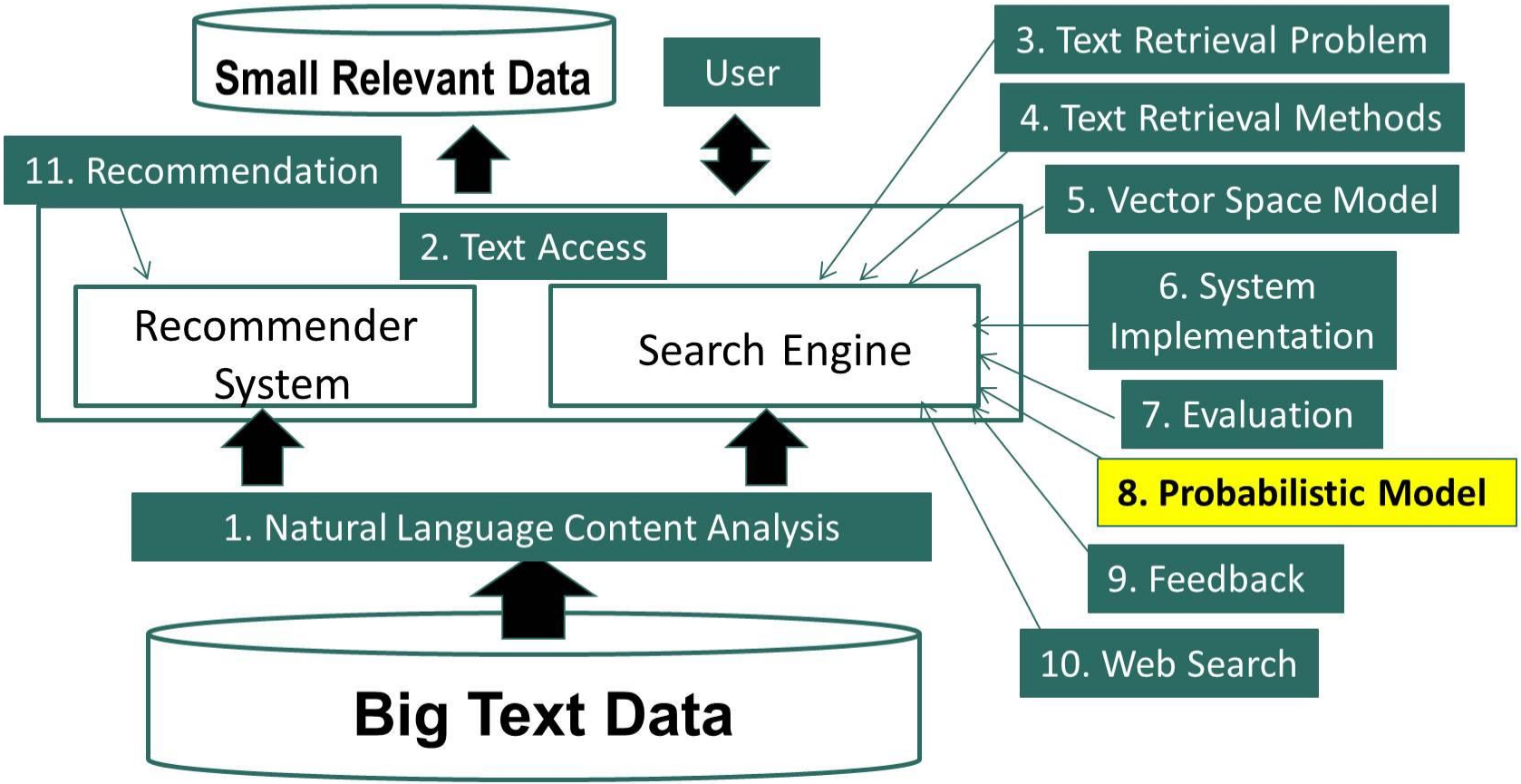
Text Retrieval and Search Engines

Probabilistic Retrieval Model: Statistical Language Model

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

Probabilistic Retrieval Model: Statistical Language Model



Overview

- What is a Language Model?
- Unigram Language Model
- Uses of a Language Model

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
  - p("*The eigenvalue is positive*") ≈ 0.00001
- Context-dependent!
- Can also be regarded as a probabilistic mechanism for "generating" [illegible] so called a "generative" model

Today is Wednesday

Today Wednesday is

• • •

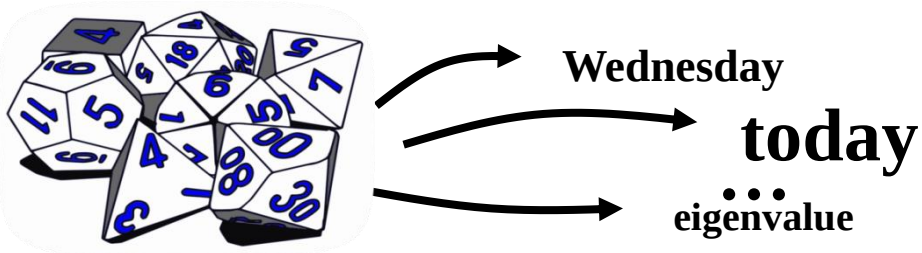
The eigenvalue is positive

Why is a LM Useful?

- Quantify the uncertainties in natural language
- Allows us to answer questions like:
  - Given that we see "*John*" and "*feels*", how likely will we see "*happy*" as opposed to "*habit*" as the next word? (speech recognition)
  - Given that we observe "baseball" three times and "game" once in a news article, how likely is it about "sports"? (text categorization, information retrieval)
  - Given that a user is interested in sports news, how likely would the user use "baseball" in a query? (information retrieval)

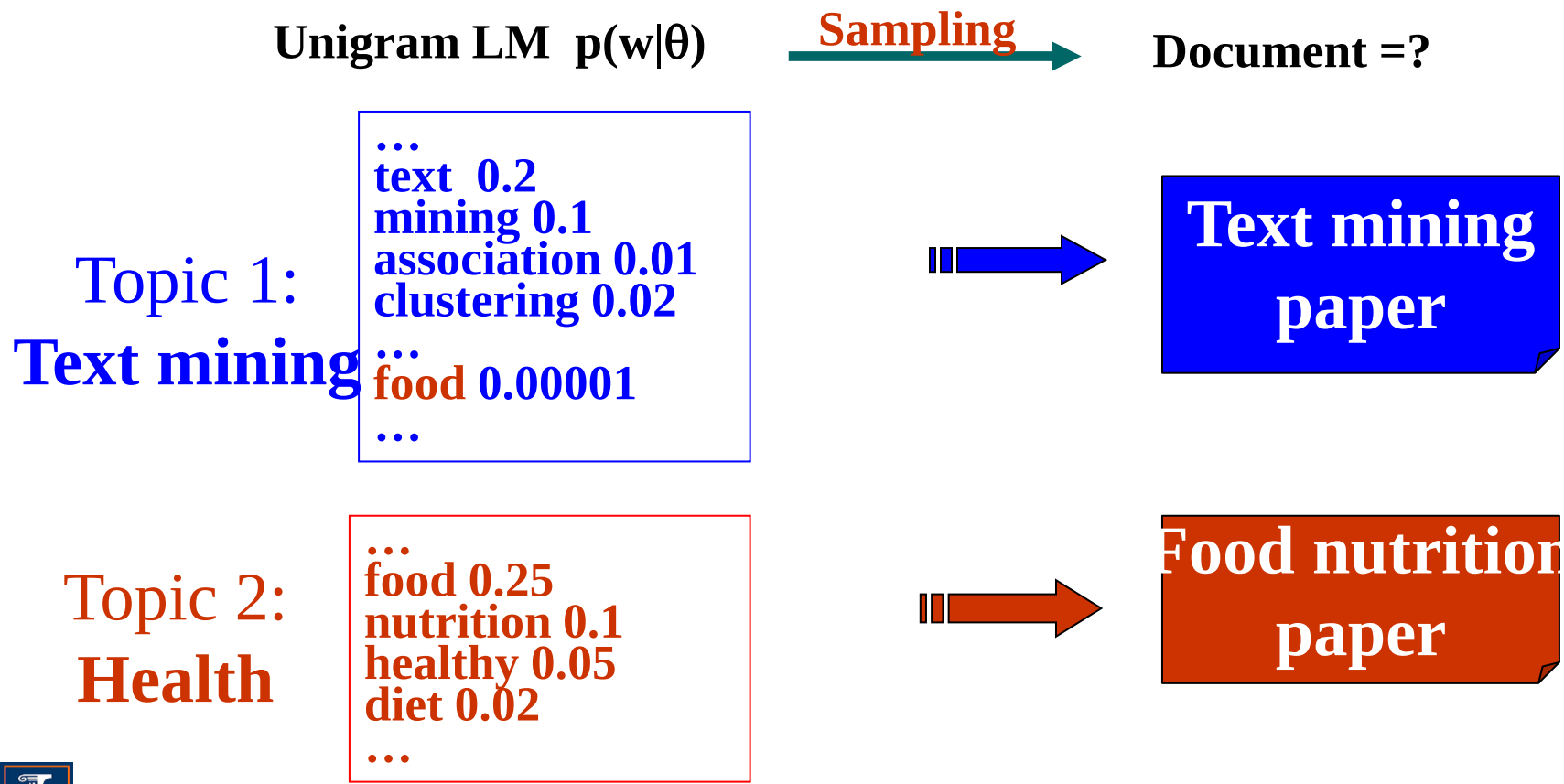
The Simplest Language Model: Unigram LM

- Generate text by generating each word INDEPENDENTLY
- Thus, $p(w_1\ w_2\ \ldots\ w_n)=p(w_1)p(w_2)\ldots p(w_n)$
- Parameters: $\{p(w_i)\}$ $p(w_1)+\ldots+p(w_N)=1$ (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



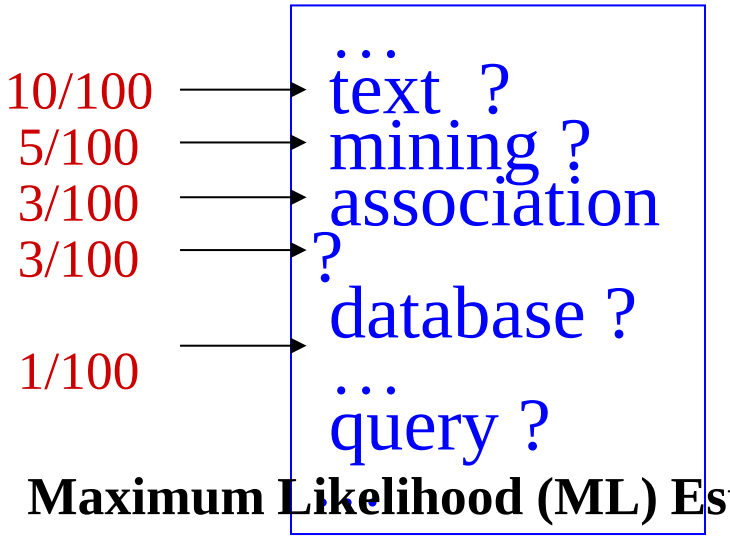
Estimation of Unigram LM

Unigram LM p(w|θ)=?

Estimation

Text Mining Paper d

Total #words=**100**



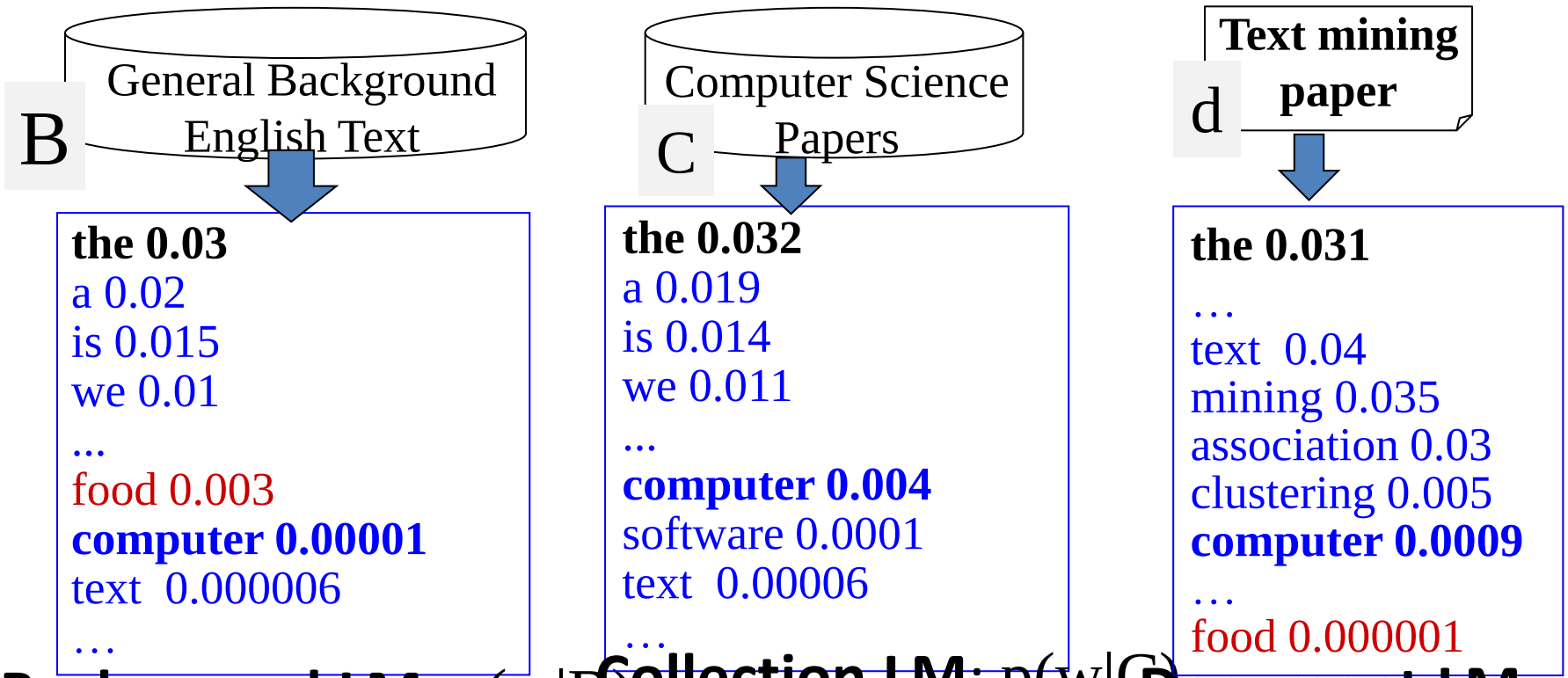
text 10
mining 5
association 3
database 3
algorithm 2
…
query 1
efficient 1

Maximum Likelihood (ML) Estimator:

$$p(w \mid \theta) = p(w \mid d) = \frac{c(w,d)}{|d|}$$

Is this the best estimate?

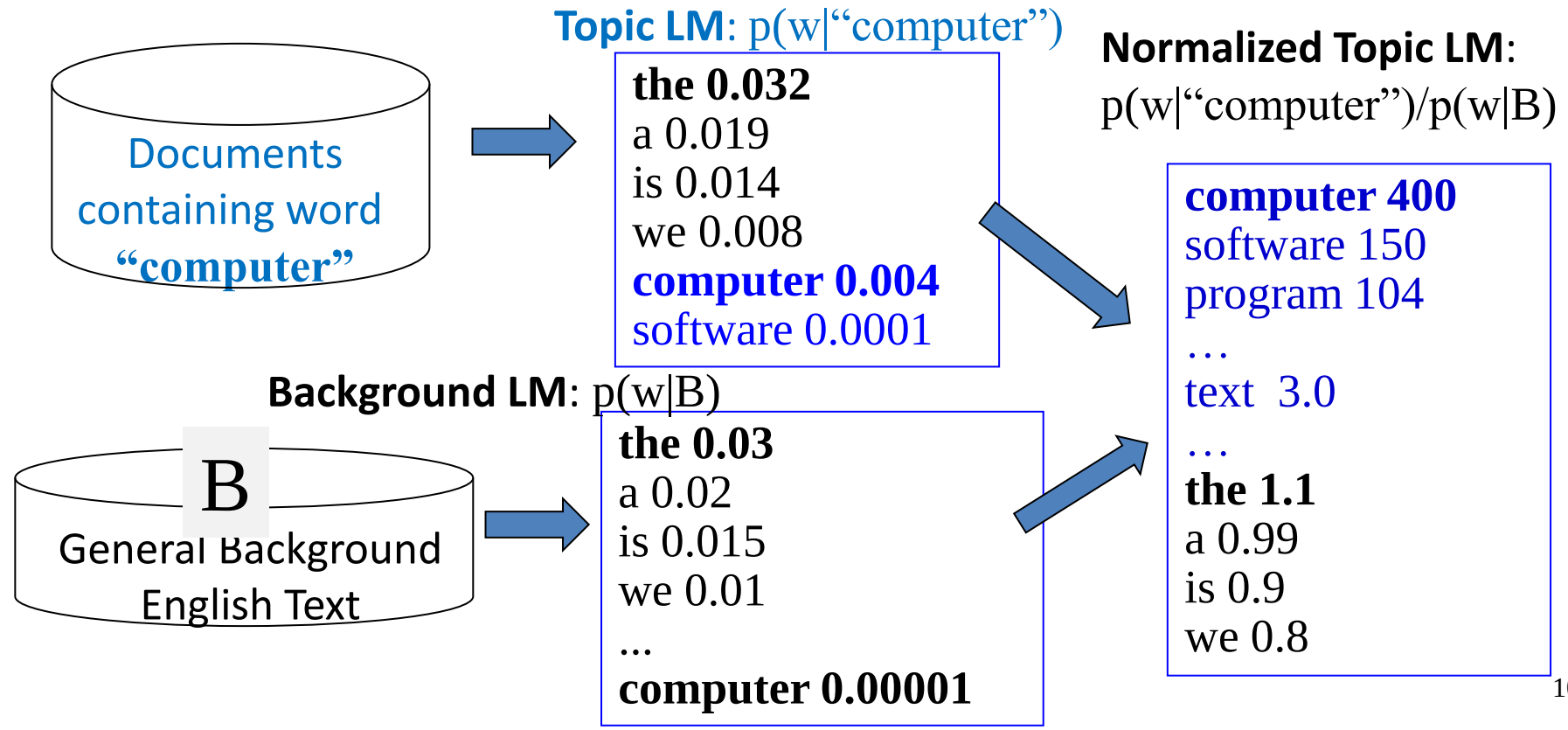
LMs for Topic Representation



Background LM: p(w|B) **Collection LM**: p(w|C) **Document LM**: p(w|

LMs for Association Analysis

What words are semantically related to "computer"?



Summary

- Language Model = probability distribution over text
- Unigram Language Model = word distribution
- Uses of a Language Model
  - Representing topics
  - Discovering word associations

Additional Readings

- Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.
- Rosenfeld, R., "Two decades of statistical language modeling: where do we go from here?," *Proceedings of the IEEE* , vol.88, no.8, pp.1270,1278, Aug. 2000