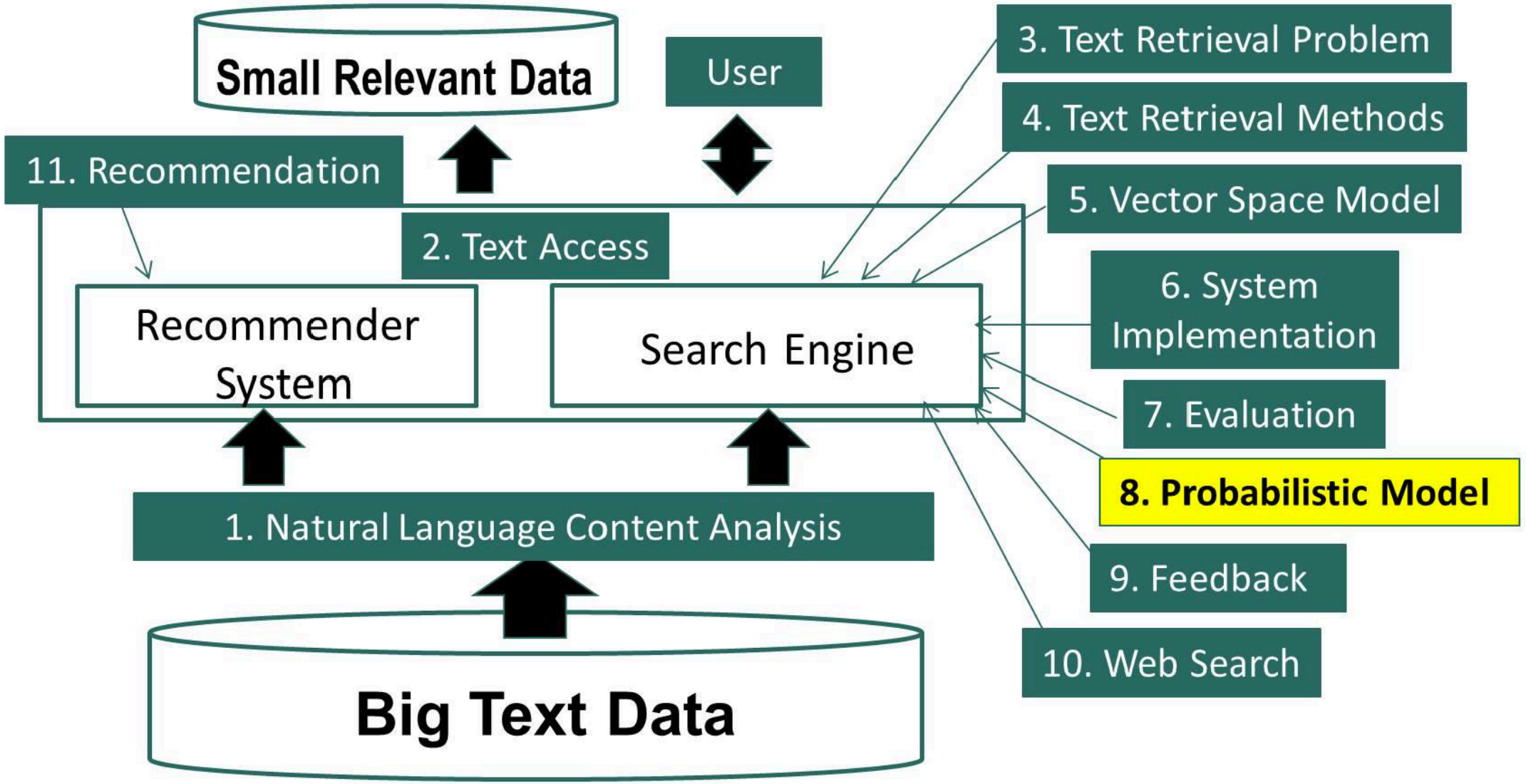
Information Retrieval & Text Mining

Probabilistic Retrieval Model:
Statistical Language Model

Dr. Iqra Safder
Information Technology University

Probabilistic Retrieval Model: Basic Idea



Overview

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

What is a Statistical Language Model (LM)?

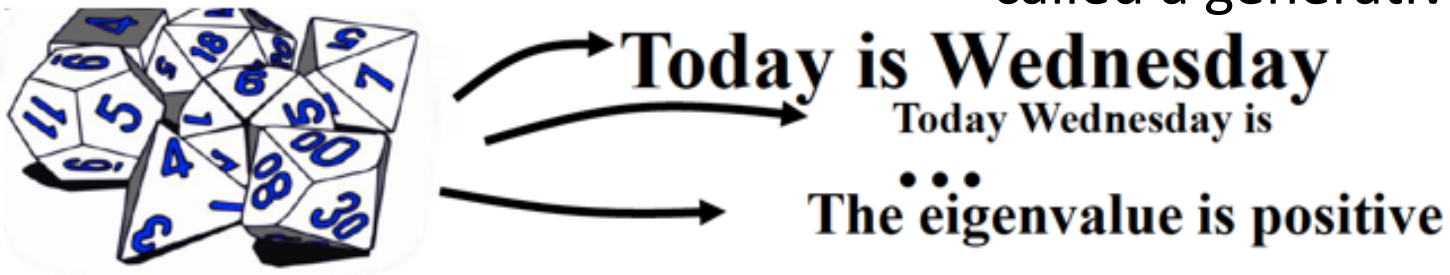
- A probability distribution over word sequences
  - p("*Today is Wednesday*") ≈ 0.001
  - p("*Today Wednesday is*") ≈ 0.0000000000001
  - p("*The eigenvalue is positive*") ≈ 0.00001
- Context-dependent!

This may be a reasonable language model for describing general conversations, but it may be inaccurate for describing conversations happening at a mathematics conference, where the sequence *The eigenvalue is positive* may occur more frequently than *Today is Wednesday.*

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!

Given a language model, we can sample word sequences according to the distribution to obtain a text sample. In this sense, we may use such a model to "generate" text. Thus, a language model is also often called a generative model for text.



Why is a LM Useful?

- Quantify the uncertainties in natural language

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)

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)

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

The Simplest Language Model: Unigram LM

- Generate text by generating each word INDEPENDENTLY
- Thus, $p(w_1\ w_2\ \ldots\ w_n)=p(w_1)p(w_2)\ldots p(w_n)$

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)

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)

Given probabilities of each word, the sum of all is = 1

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**

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**

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

Unigram LM p(w|θ) → **Sampling** → **Document =?**

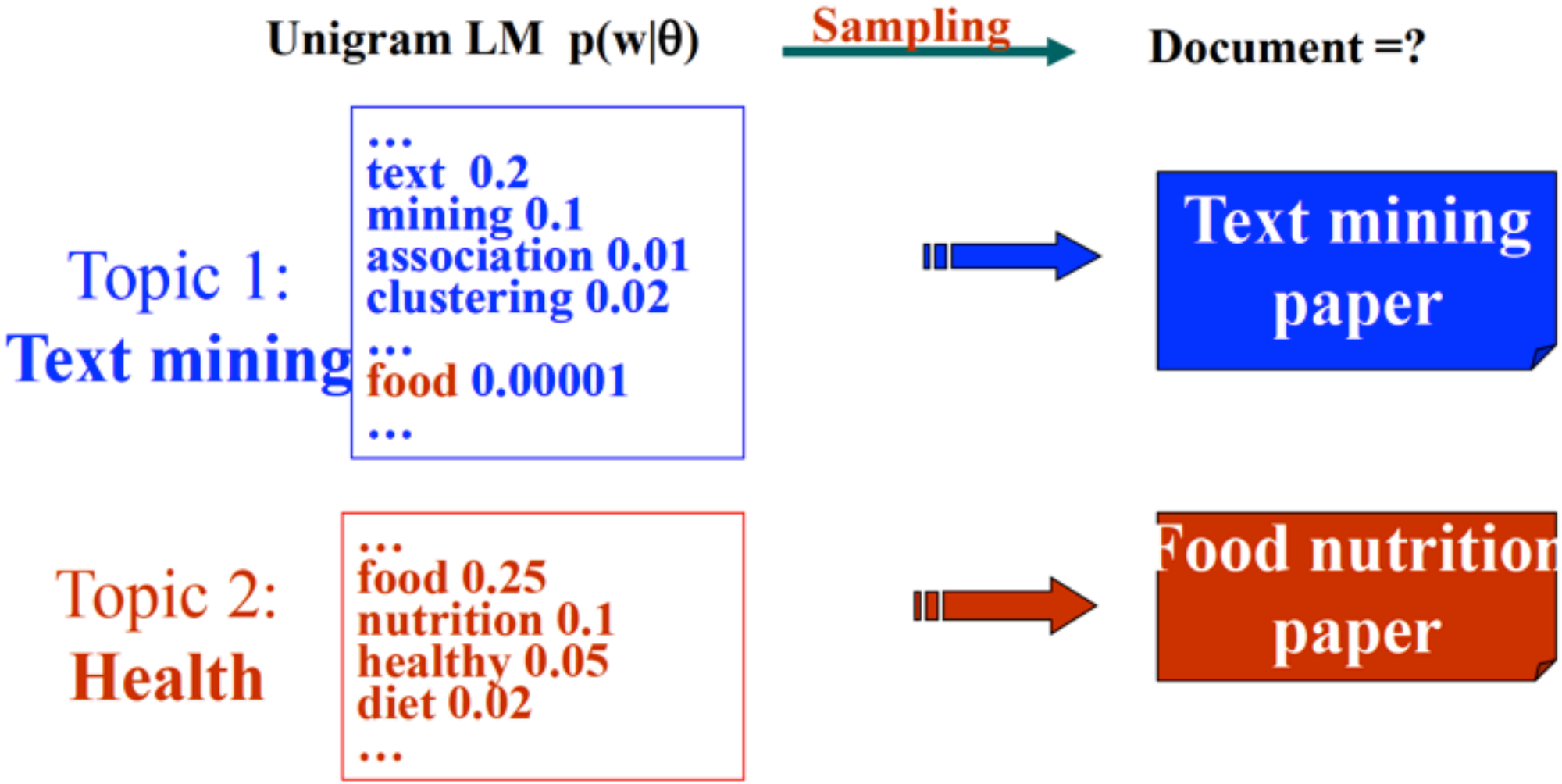
Topic 1: **Text mining**

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

Topic 2: **Health**

…
food 0.25
nutrition 0.1
healthy 0.05
diet 0.02
…

Text Generation with Unigram LM



Estimation of Unigram LM

Unigram LM p(w|θ)=?

Estimation

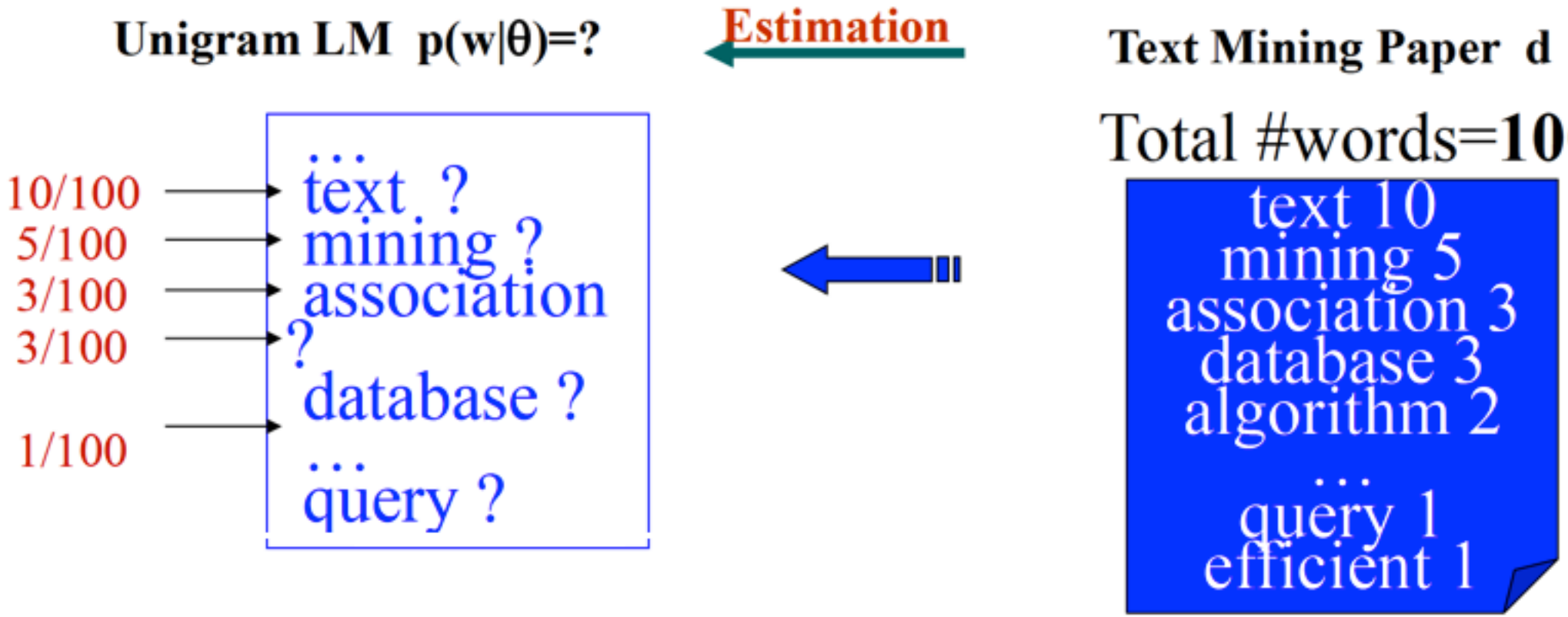
Text Mining Paper d

Total #words=**100**

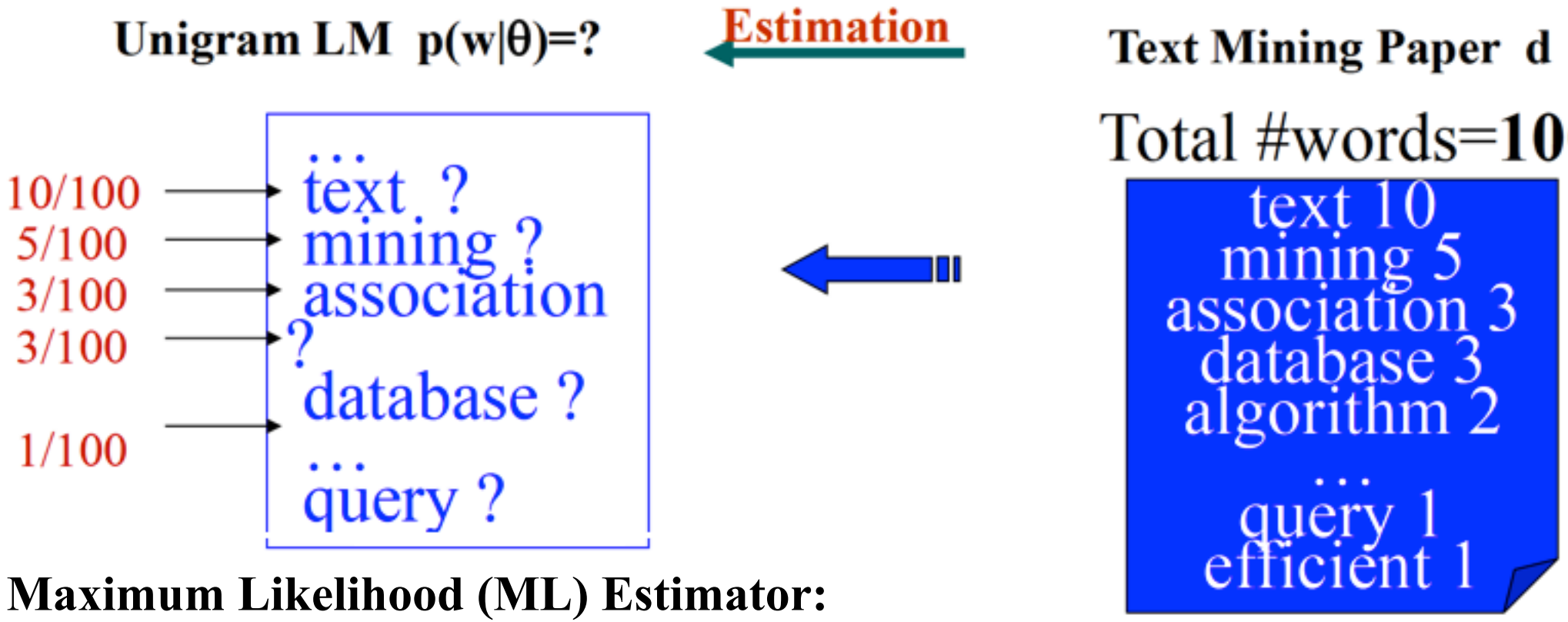
…
text ?
mining ?
association ?
database ?
…
query ?

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

Estimation of Unigram LM



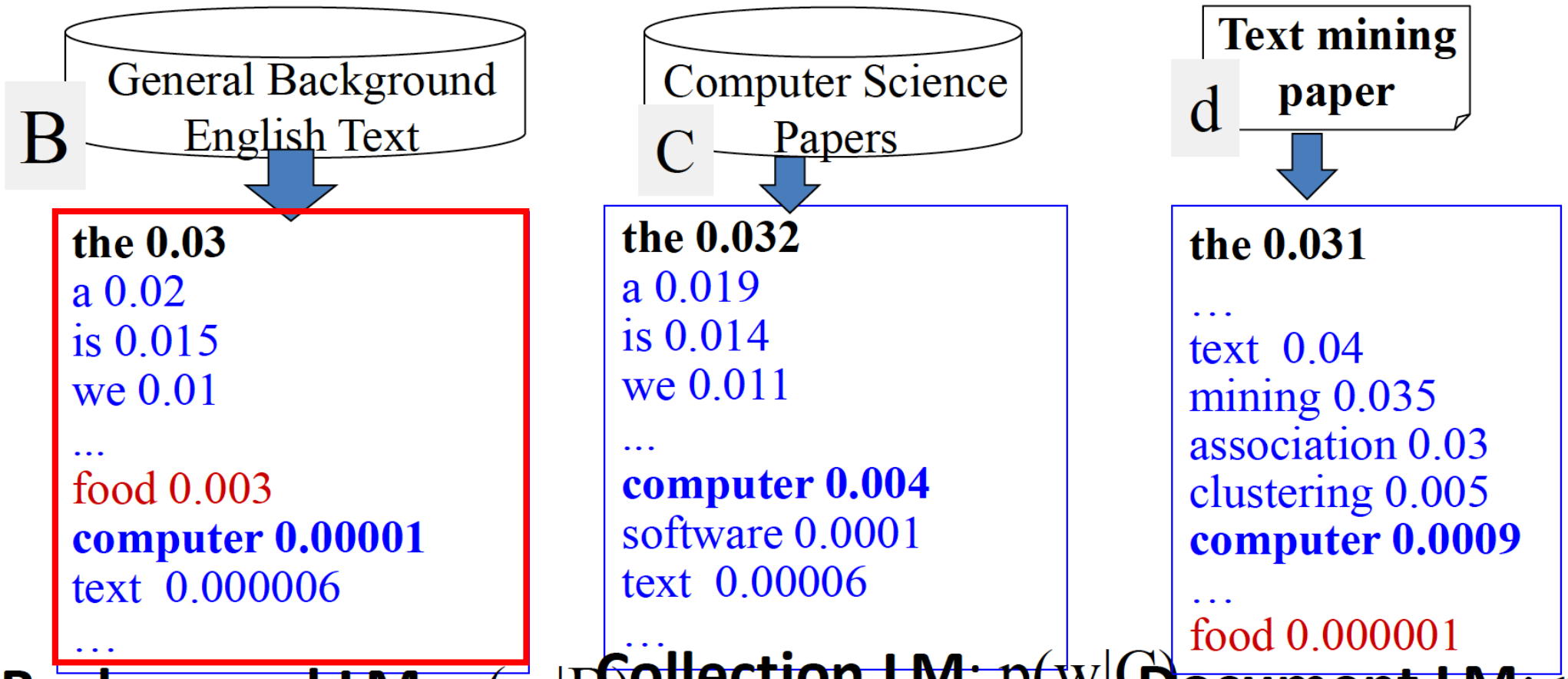
Estimation of Unigram LM



Maximum Likelihood (ML) Estimator:

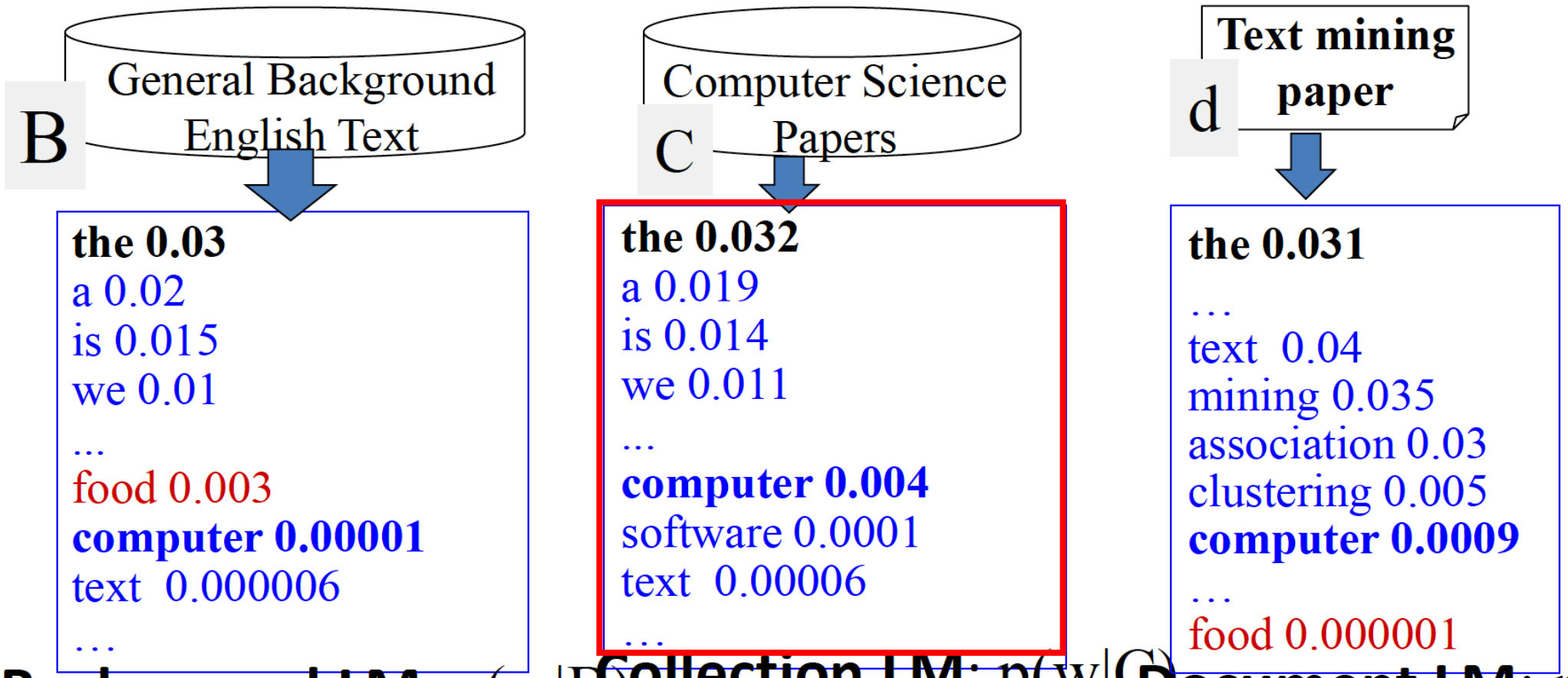
$$p(w \mid \theta) = p(w \mid d) = \frac{c(w,d)}{|d|}$$

LMs for Topic Representation



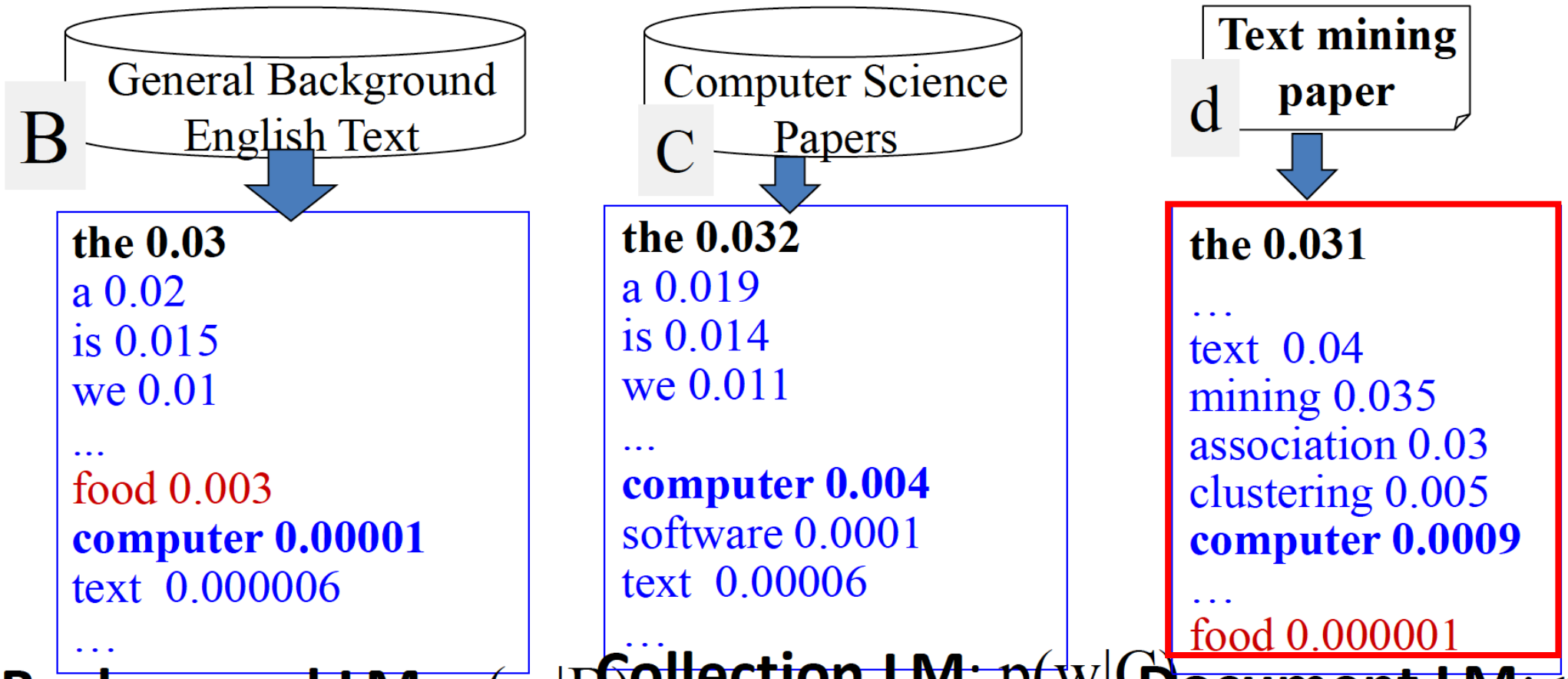
Background LM: p(w|B) **Collection LM**: p(w|C) **Document LM**: p(w

LMs for Topic Representation



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

LMs for Topic Representation

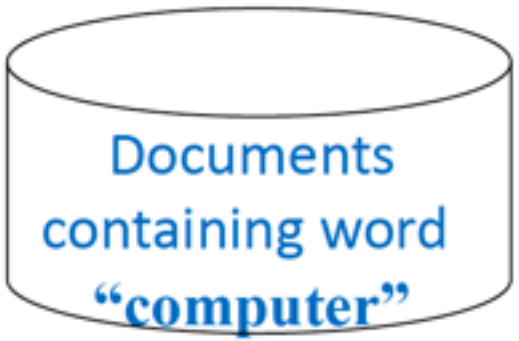


Background LM: p(w|B) **Collection LM**: p(w|C) **Document LM**: p(w

LMs for Association Analysis

What words are semantically related to "computer"?

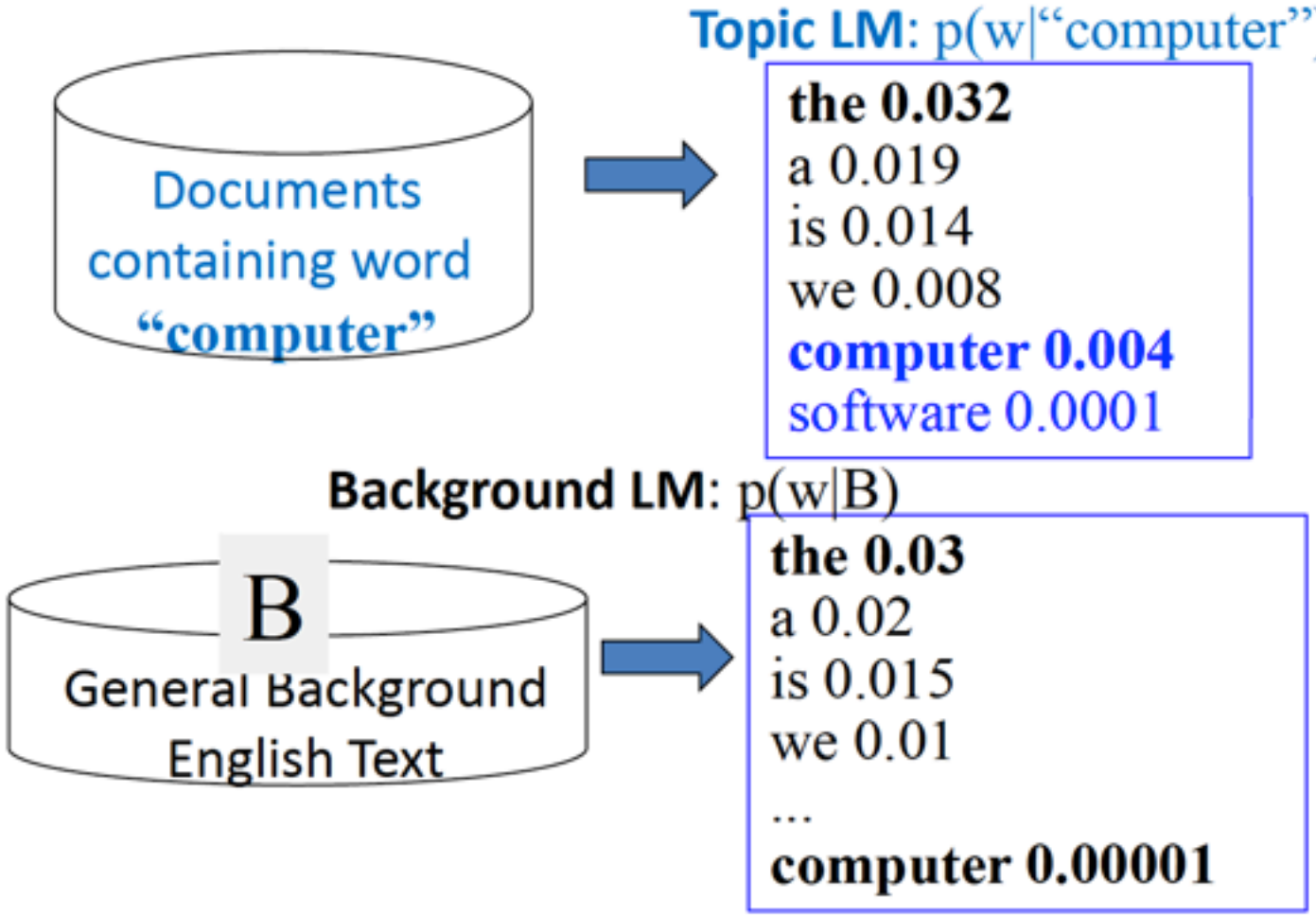
Topic LM: p(w|"computer")



the 0.032
a 0.019
is 0.014
we 0.008
computer 0.004
software 0.0001

LMs for Association Analysis

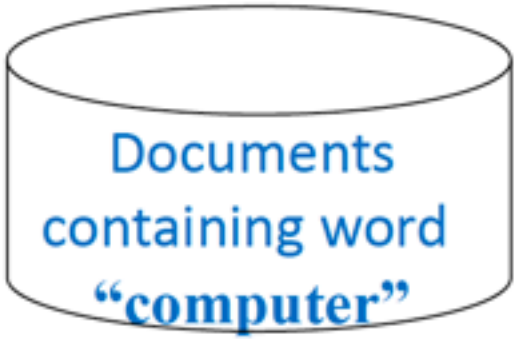
What words are semantically related to "computer"?



LMs for Association Analysis

What words are semantically related to "computer"?

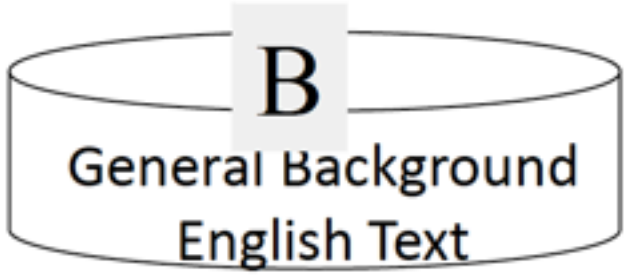
Topic LM: p(w|"computer")



the 0.032
a 0.019
is 0.014
we 0.008
computer 0.004
software 0.0001

Normalized Topic LM:
p(w|"computer")/p(w|B)

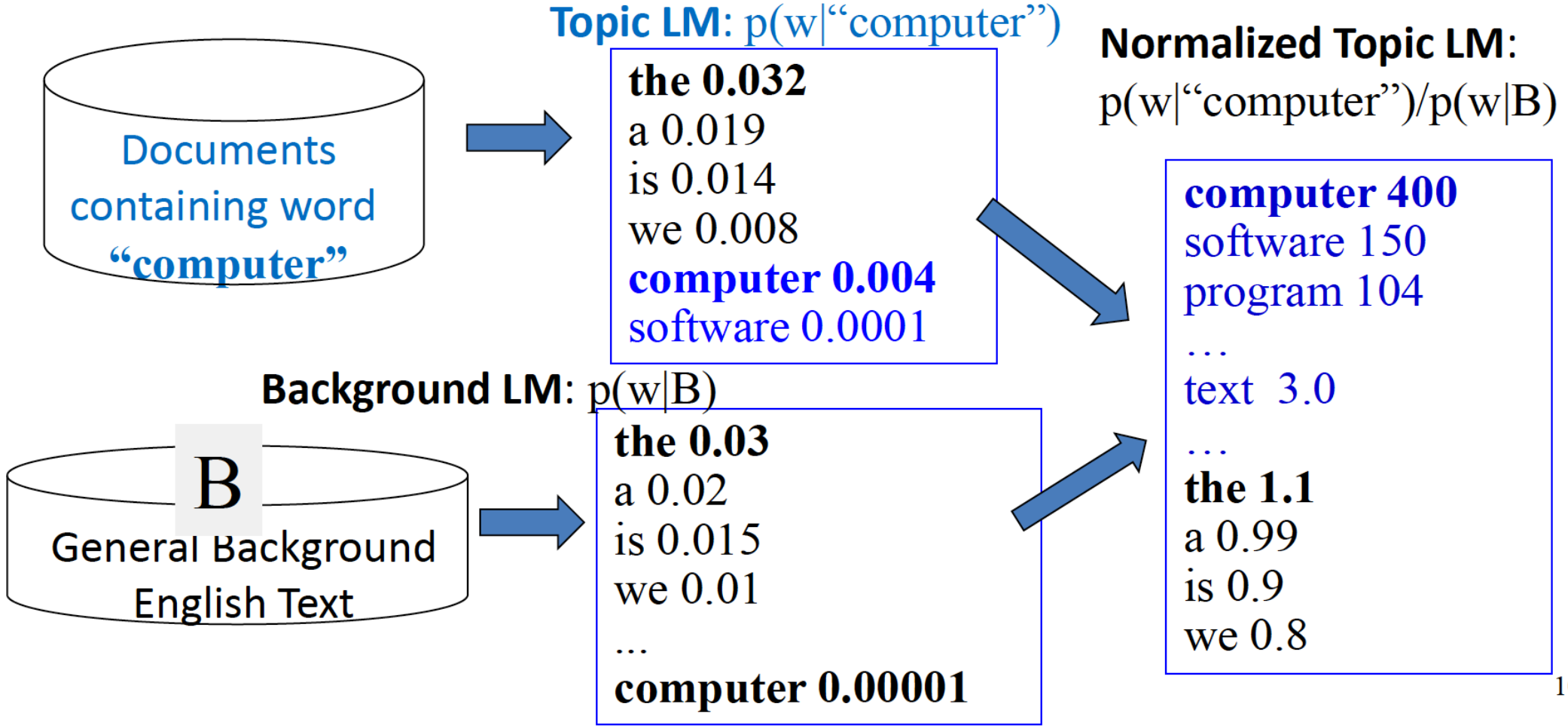
Background LM: p(w|B)



the 0.03
a 0.02
is 0.015
we 0.01
...
computer 0.00001

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