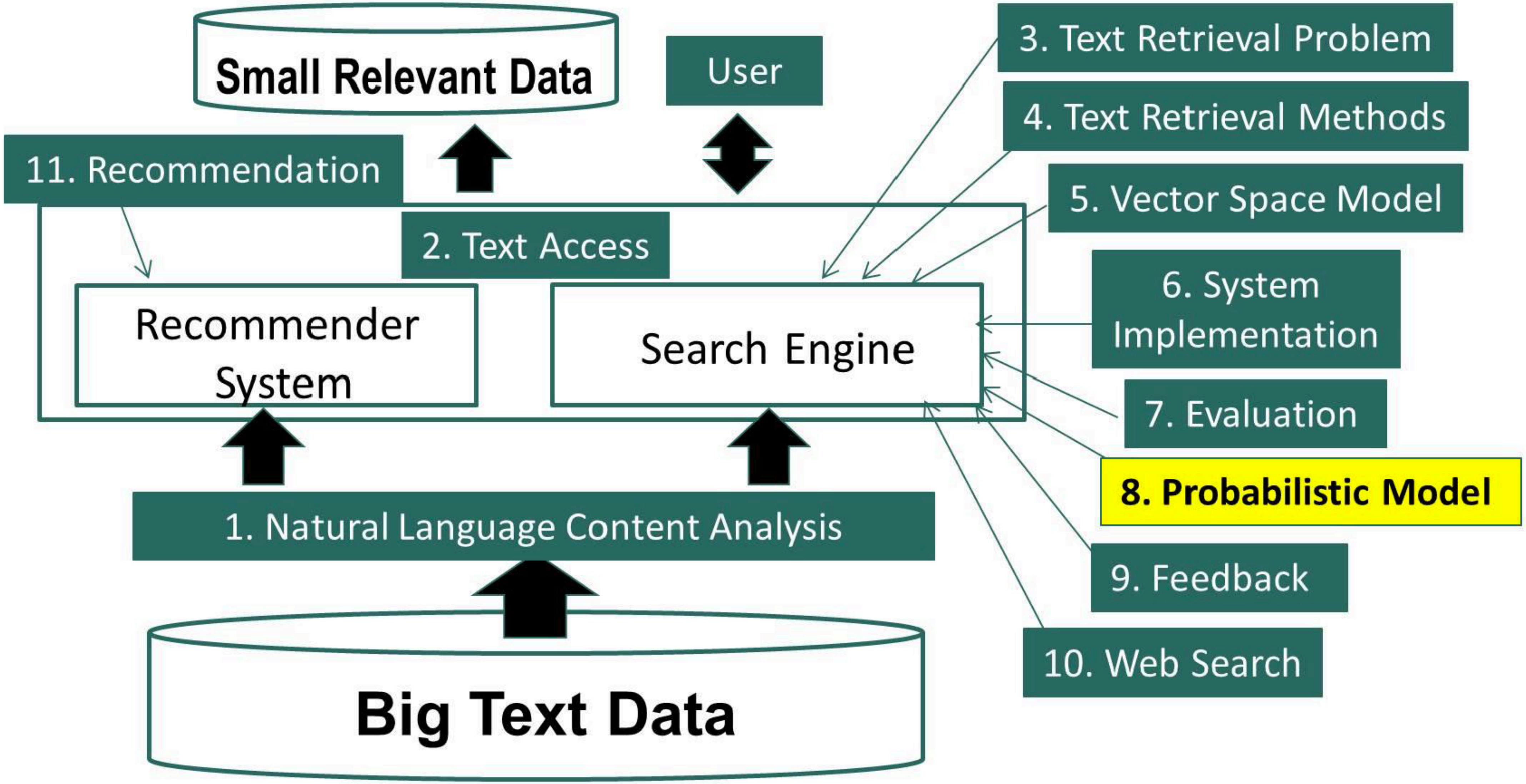
Information Retrieval & Text Mining

Probabilistic Retrieval Model:
Statistical Language Model

Dr. Saeed Ul Hassan
Information Technology University

Probabilistic Retrieval Model: Basic Idea



Overview

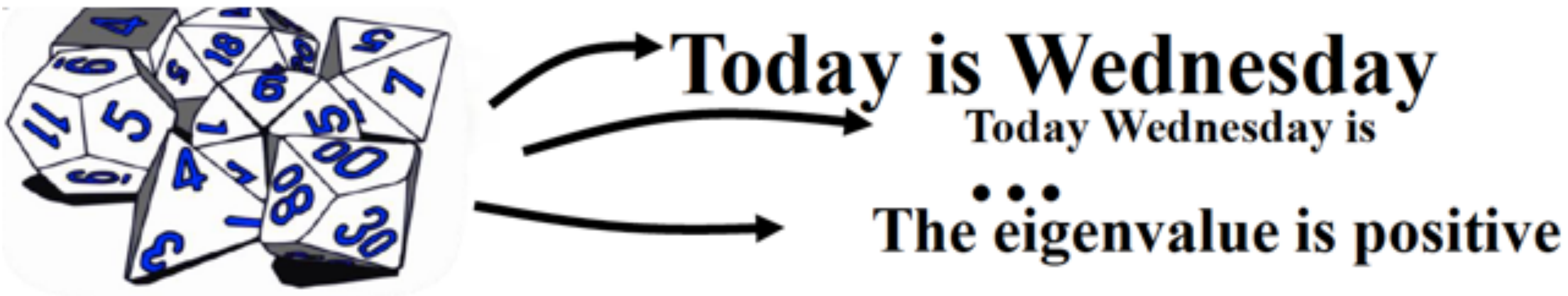
- What is a Language Model?
- Unigram Language Model
- Uses of a Language Model

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!

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!



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, $p(w_1\ w_2\ \dots\ w_n)=p(w_1)p(w_2)\dots p(w_n)$
- Parameters: $\{p(w_i)\}$ $p(w_1)+\dots+p(w_N)=1$ (N is voc. size)

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, $p(w_1\ w_2\ \ldots\ w_n)=p(w_1)p(w_2)\ldots p(w_n)$
- 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**

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**

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

Text Generation with Unigram LM

Unigram LM $p(w|\theta)$ —**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

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
…

→ **Text mining paper**

Topic 2:
Health

…
food 0.25
nutrition 0.1
healthy 0.05
diet 0.02
…

→ **Food nutrition paper**

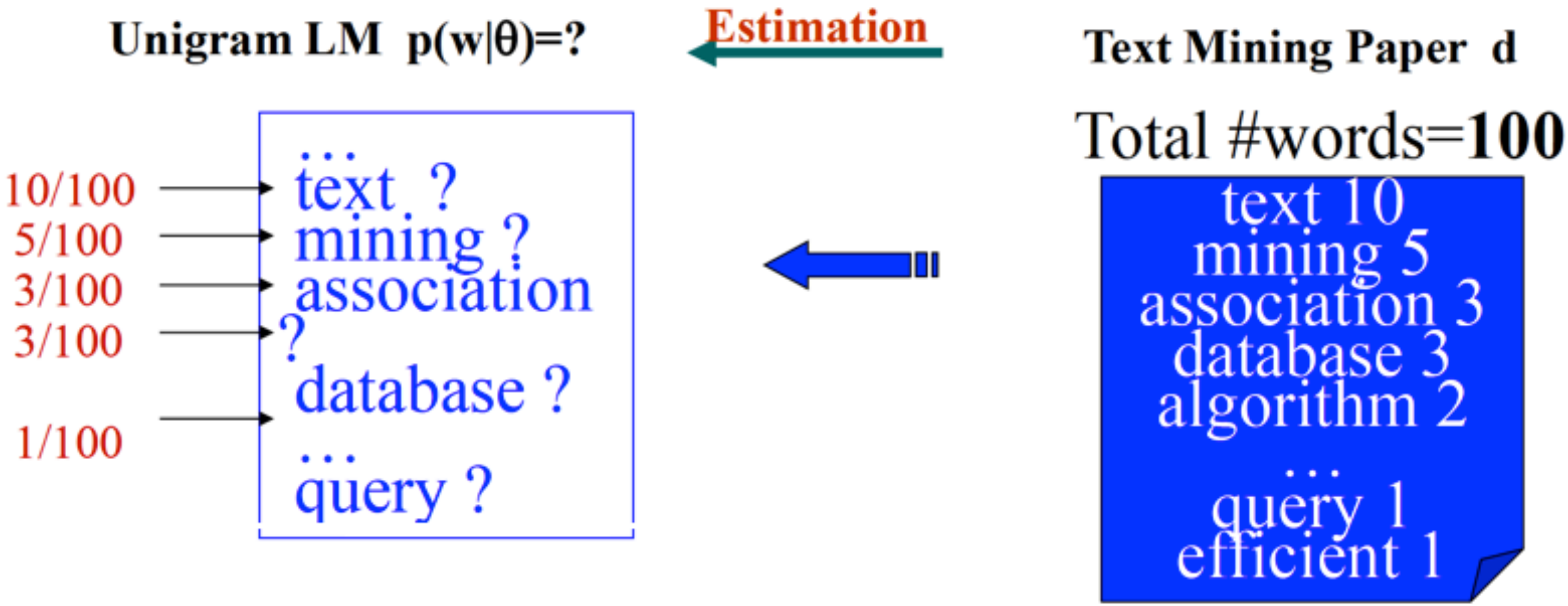
Estimation of Unigram LM

Unigram LM p(w|θ)=? **Estimation** **Text Mining Paper d**

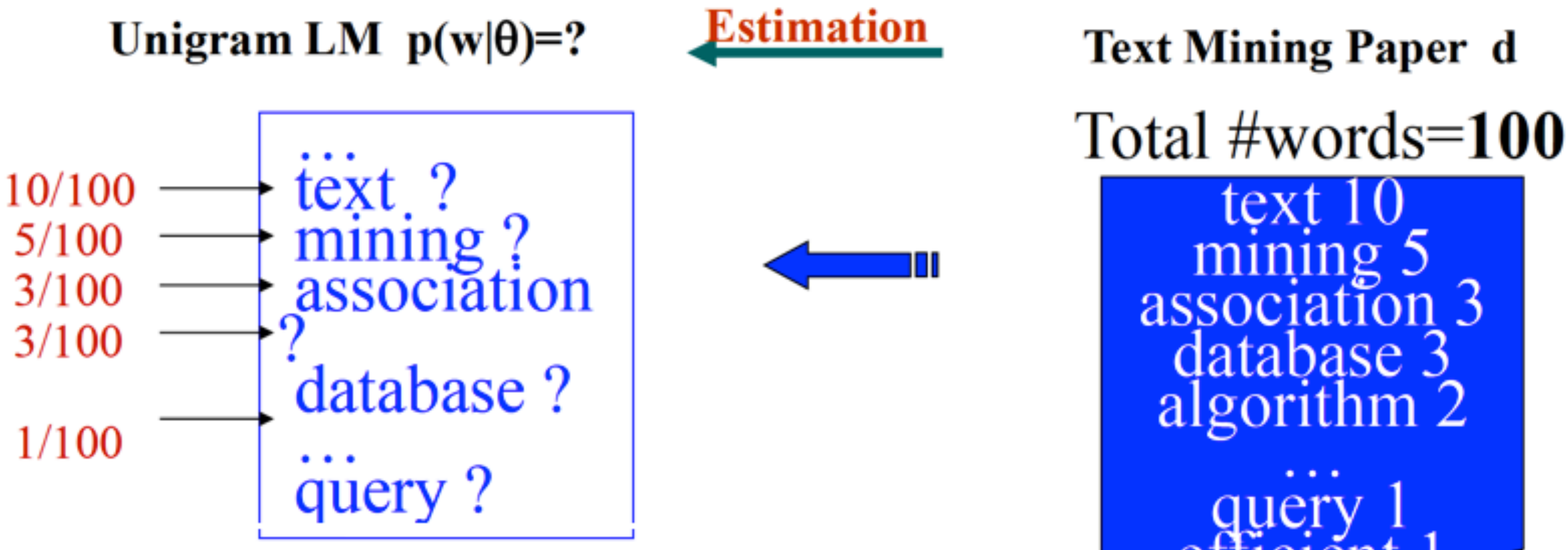
Total #words=**100**

Unigram LM p(w|θ)	Text Mining Paper d
…	
text ?	text 10
mining ?	mining 5
association ?	association 3
database ?	database 3
	algorithm 2
…	…
query ?	query 1
	efficient 1

Estimation of Unigram LM



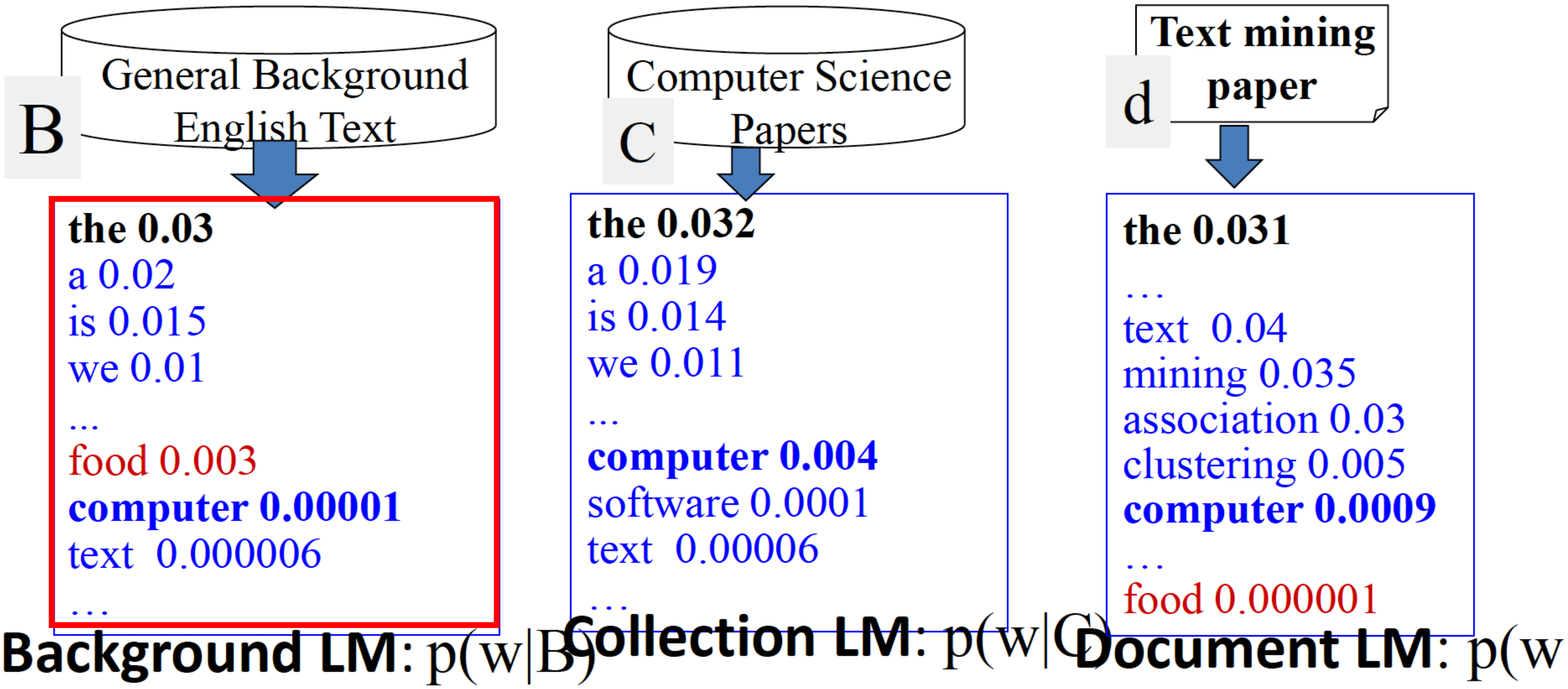
Estimation of Unigram LM



Maximum Likelihood (ML) Estimator:

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

LMs for Topic Representation



LMs for Topic Representation

B — General Background English Text

the 0.03
a 0.02
is 0.015
we 0.01
...
food 0.003
computer 0.00001
text 0.000006
…

Background LM: p(w|B)

C — Computer Science Papers

the 0.032
a 0.019
is 0.014
we 0.011
...
computer 0.004
software 0.0001
text 0.00006
…

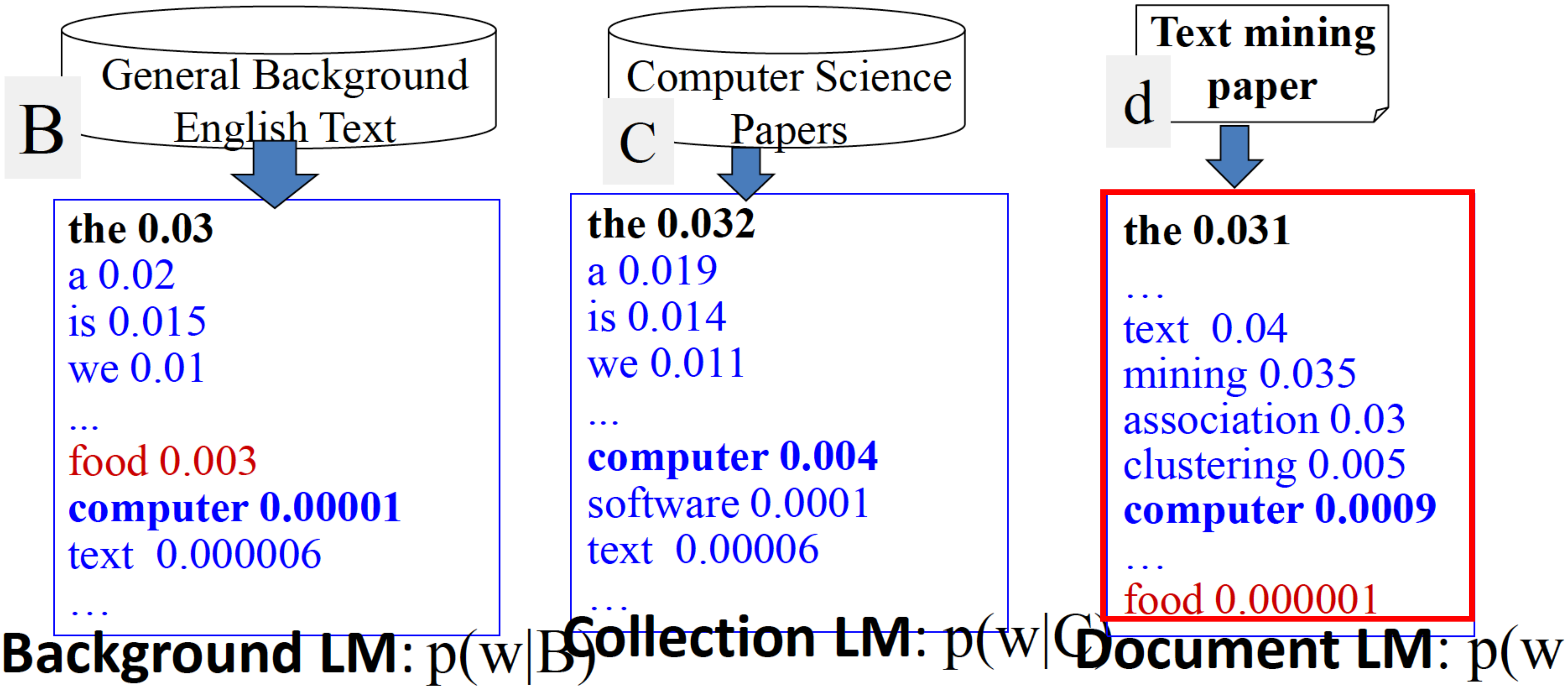
Collection LM: p(w|C)

d — **Text mining paper**

the 0.031
…
text 0.04
mining 0.035
association 0.03
clustering 0.005
computer 0.0009
…
food 0.000001

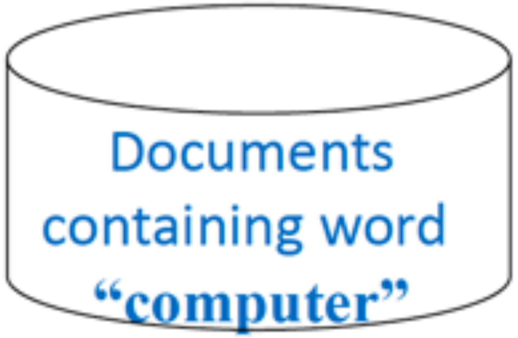
Document LM: p(w

LMs for Topic Representation



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”?

Documents containing word **“computer”**

Topic LM: p(w|“computer”)

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

B

General Background English Text

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"?

Topic LM: p(w|"computer")

Documents containing word "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)

B
General Background English Text

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"?

Documents containing word "computer"

Topic LM: p(w|"computer")

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

General Background English Text

B

Background LM: p(w|B)

the 0.03
a 0.02
is 0.015
we 0.01
…
computer 0.00001

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

computer 400
software 150
program 104
…
text 3.0
…
the 1.1
a 0.99
is 0.9
we 0.8

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