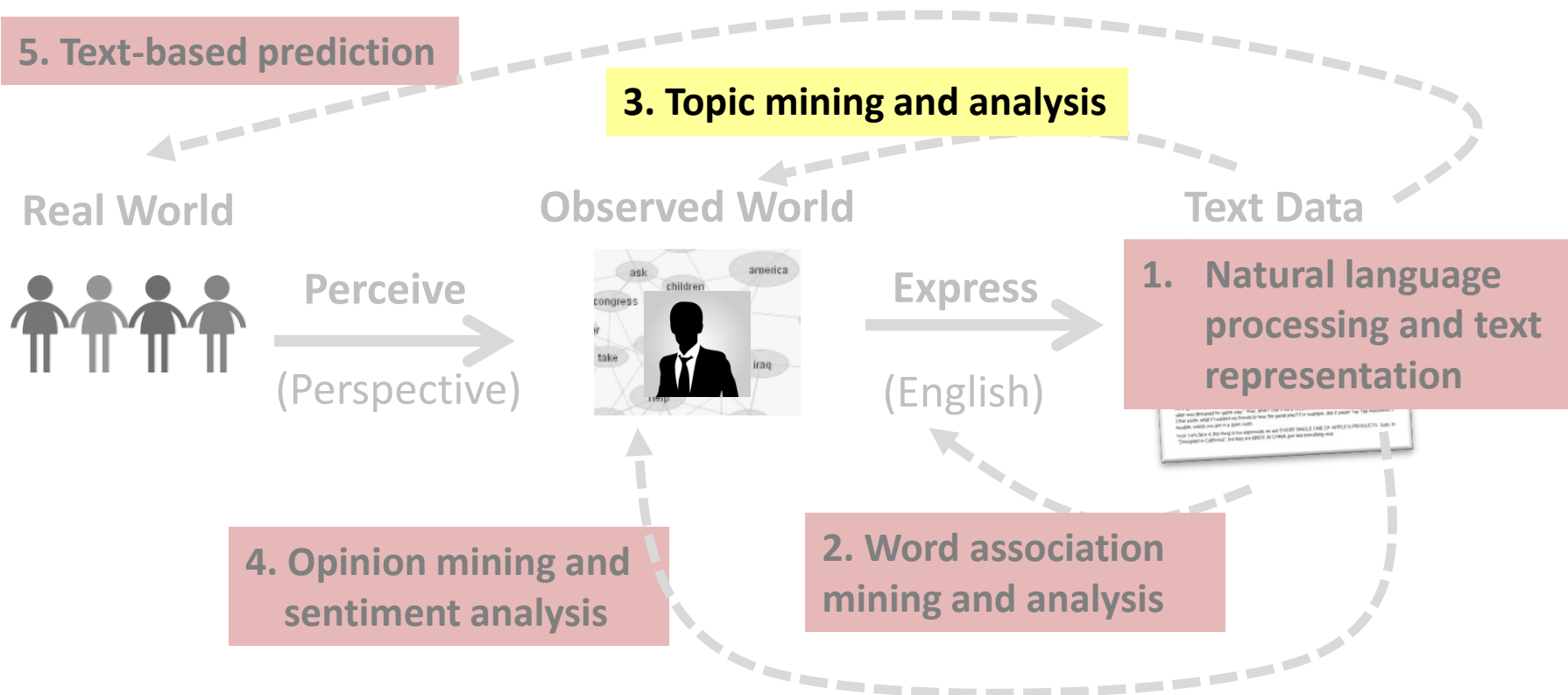


Text Clustering: Generative Probabilistic Models

Part 1

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

Text Clustering: Generative Probabilistic Models (Part 1)



Overview

- What is text clustering?
- Why text clustering?
- How to do text clustering?
 - **Generative probabilistic models**
 - Similarity-based approaches
- How to evaluate clustering results?

Topic Mining Revisited

INPUT: C, k, V

OUTPUT: $\{ \theta_1, \dots, \theta_k \}, \{ \pi_{i1}, \dots, \pi_{ik} \}$

Text Data

θ_1

sports 0.02
game 0.01
basketball 0.005
football 0.004
...

θ_2

travel 0.05
attraction 0.03
trip 0.01
...

...

θ_k

science 0.04
scientist 0.03
spaceship 0.006
...

Doc 1

30%

π_{11}

Doc 2

$\pi_{21}=0\%$

...

Doc N

$\pi_{N1}=0\%$

12%

π_{12}

π_{22}

π_{N2}

8%

π_{1k}

π_{2k}

π_{Nk}

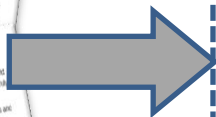
One Topic(=cluster) Per Document

INPUT: C, k, V

OUTPUT: $\{ \theta_1, \dots, \theta_k \},$

$\{ c_1, \dots, c_N \} c_i \in [1, k]$

Text Data



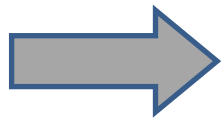
	Doc 1	Doc 2	...	Doc N
θ_1	<div>sports 0.02 game 0.01 basketball 0.005 football 0.004 ...</div>	<div>$\pi_{11}=100\%$ [Bar chart showing 100% for topic 1]</div>	<div>$\pi_{21}=0\%$ [Bar chart showing 0% for topic 1]</div>	<div>$\pi_{N1}=100\%$ [Bar chart showing 100% for topic 1]</div>
θ_2	<div>travel 0.05 attraction 0.03 trip 0.01 ...</div>	<div>$\pi_{12}=0$ [Bar chart showing 0% for topic 2]</div>	<div>$\pi_{22}=100\%$ [Bar chart showing 100% for topic 2]</div>	<div>$\pi_{N2}=0$ [Bar chart showing 0% for topic 2]</div>
...				
θ_k	<div>science 0.04 scientist 0.03 spaceship 0.006 ...</div>	<div>$\pi_{1k}=0$ [Bar chart showing 0% for topic k]</div>	<div>$\pi_{1k}=0$ [Bar chart showing 0% for topic k]</div>	<div>$\pi_{Nk}=0$ [Bar chart showing 0% for topic k]</div>

Mining One Topic Revisited

INPUT: $C=\{d\}$, V

OUTPUT: $\{\theta\}$

Text Data



$P(w|\theta)$

Doc d

100%

text ?
mining ?
association ?
database ?

θ

(1 Doc, 1 Topic)

→ (N Docs, N Topics)

$k < N$

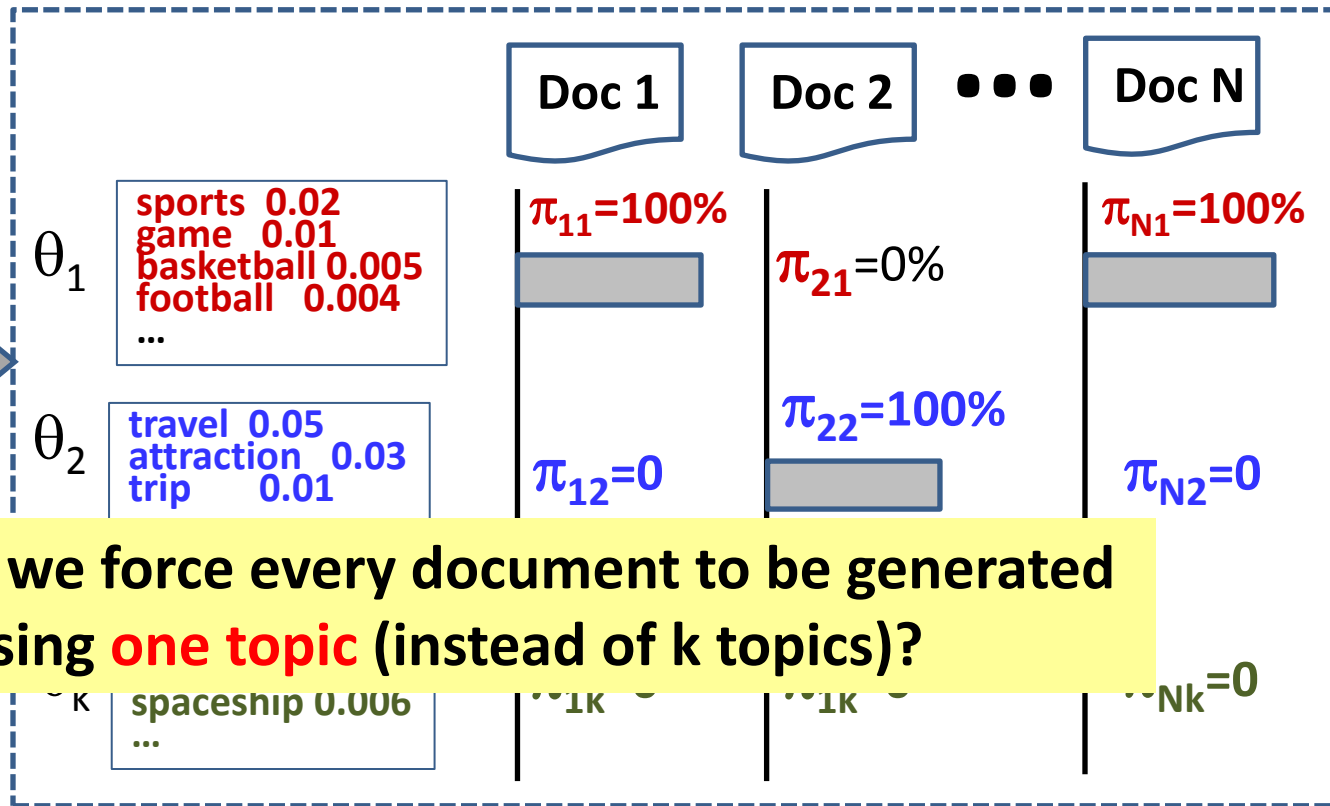
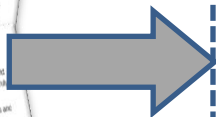
→ (N Docs, k Shared Topics)=Clustering!

What Generative Model Can Do Clustering?

INPUT: C, k, V

OUTPUT: $\{\theta_1, \dots, \theta_k\}, \{c_1, \dots, c_N\} c_i \in [1, k]$

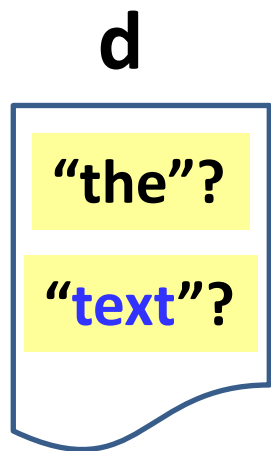
Text Data



How can we force every document to be generated using **one topic** (instead of k topics)?

Generative Topic Model Revisited

Why can't this model be used for clustering?



$P(w | \theta_1)$

$P(w | \theta_2)$

θ_1

t 0.04
mining 0.035
association 0.03
clustering 0.005
...
the 0.000001

θ_2

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

$$p(\theta_1) + p(\theta_2) = 1$$

$$P(\theta_1) = 0.5$$

$$P(\theta_2) = 0.5$$

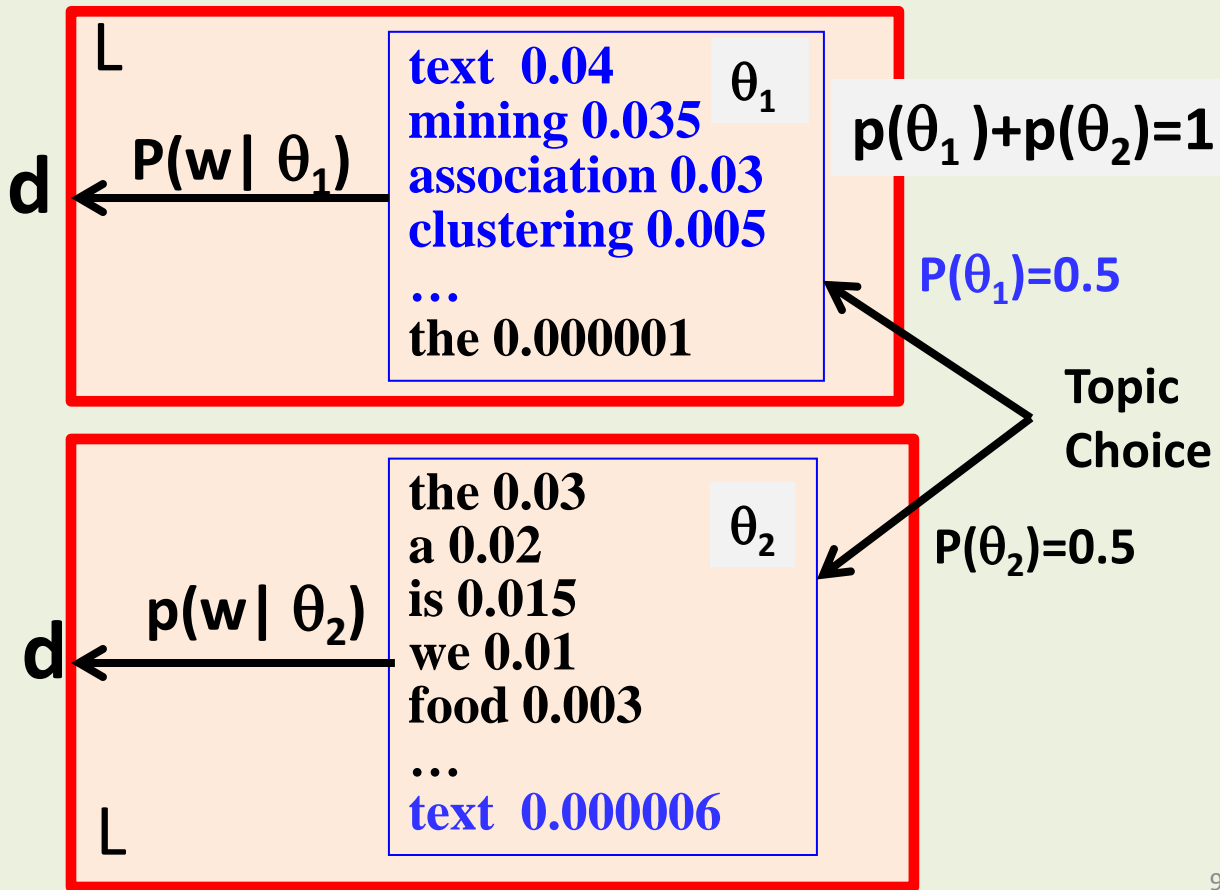
Topic
Choice

Mixture Model for Document Clustering

Difference from
topic model?

$d = x_1 x_2 \dots x_L$

What if $P(\theta_1)=1$
or $P(\theta_2)=1$?



Likelihood Function: $p(d)=?$

$$p(d) = p(\theta_1)p(d | \theta_1) + p(\theta_2)p(d | \theta_2)$$

$$= p(\theta_1) \prod_{i=1}^L p(x_i | \theta_1) + p(\theta_2) \prod_{i=1}^L p(x_i | \theta_2)$$

$d = x_1 x_2 \dots x_L$

How is this different from a topic model?

topic model: $p(d) = \prod_{i=1}^L [p(\theta_1)p(x_i | \theta_1) + p(\theta_2)p(x_i | \theta_2)]$

