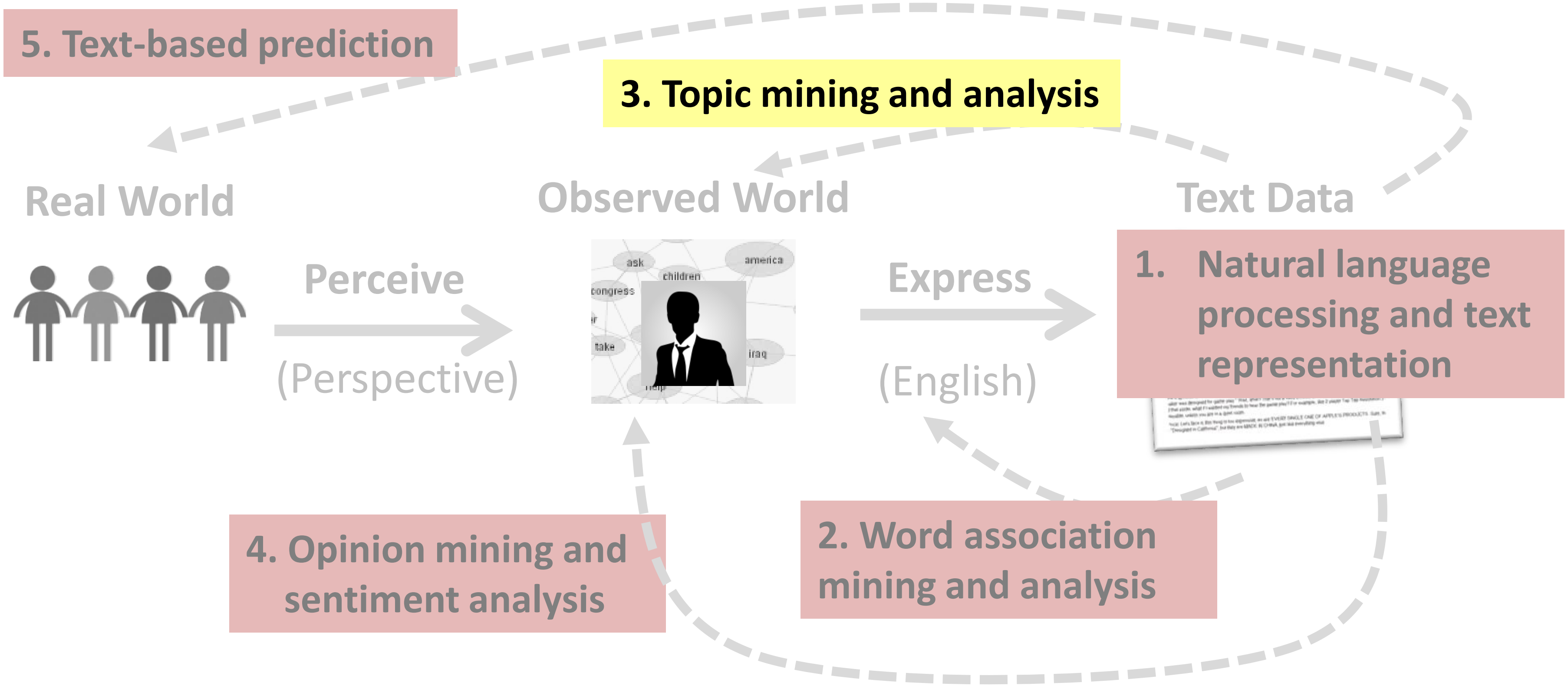
Text Clustering: Generative Probabilistic Models

Part 1

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

Text Clustering: Generative Probabilistic Models (Part 1)



5. Text-based prediction

3. Topic mining and analysis

Real World

Observed World

Text Data

Perceive

(Perspective)

Express

(English)

1. Natural language processing and text representation

4. Opinion mining and sentiment analysis

2. Word association mining and analysis

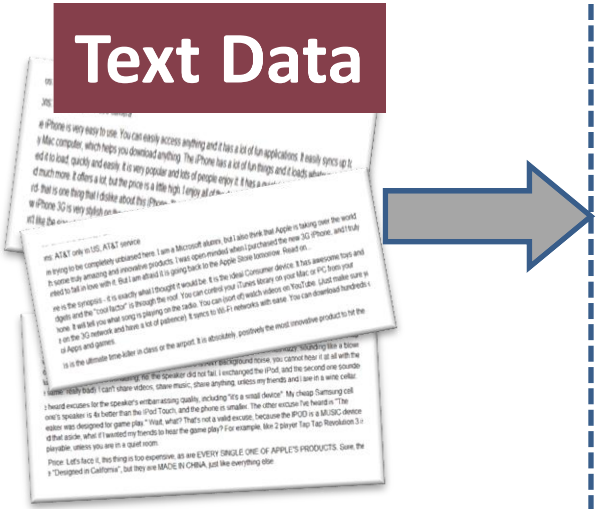
Overview

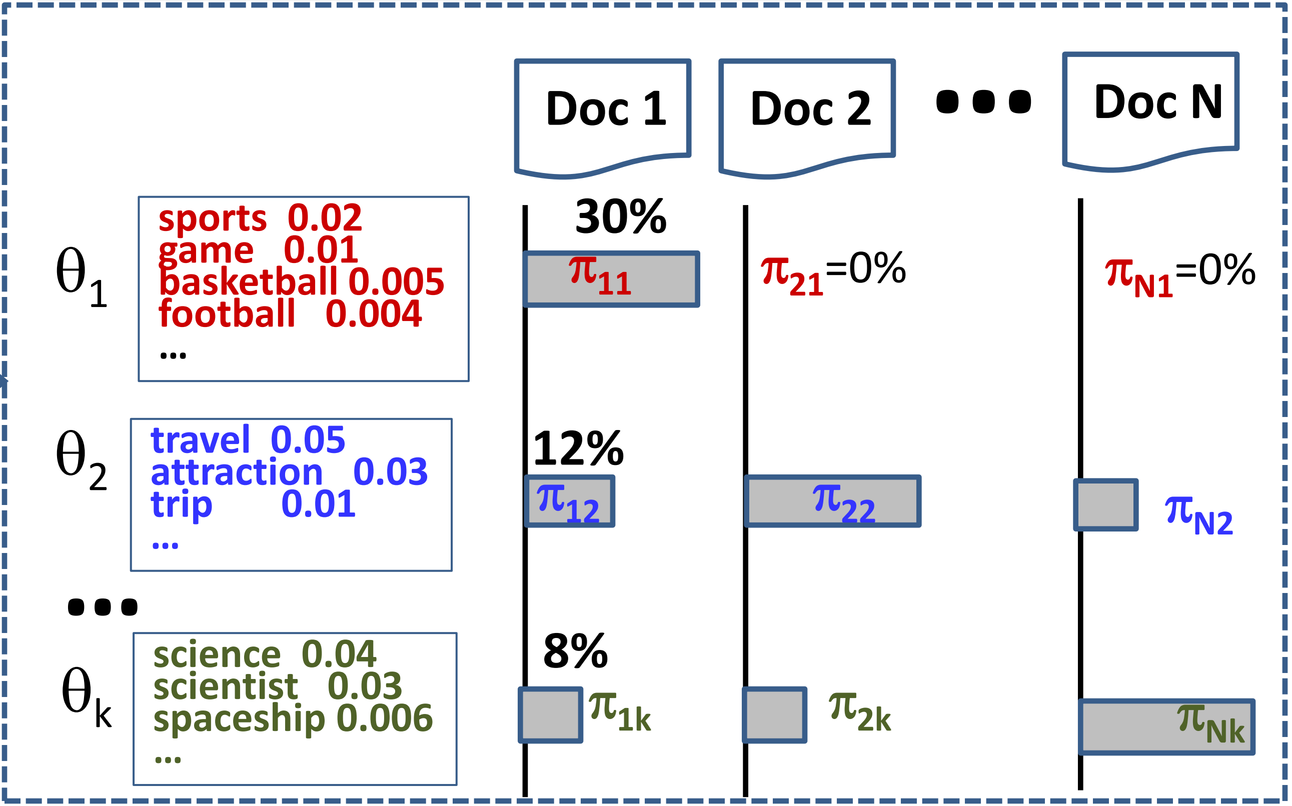
- What is text clustering?
- Why text clustering?
- How to do text clustering?
  - **Generative probabilistic models** 生成概率模型
  - Similarity-based approaches 相似、度为基础的方法.
- How to evaluate clustering results?

Topic Mining Revisited

INPUT: C, k, V

OUTPUT: { θ_1, ..., θ_k }, { π_{i1}, ..., π_{ik} }





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 … Doc 2 … Doc N

Doc 1: 30% π_{11}; 12% π_{12}; 8% π_{1k}

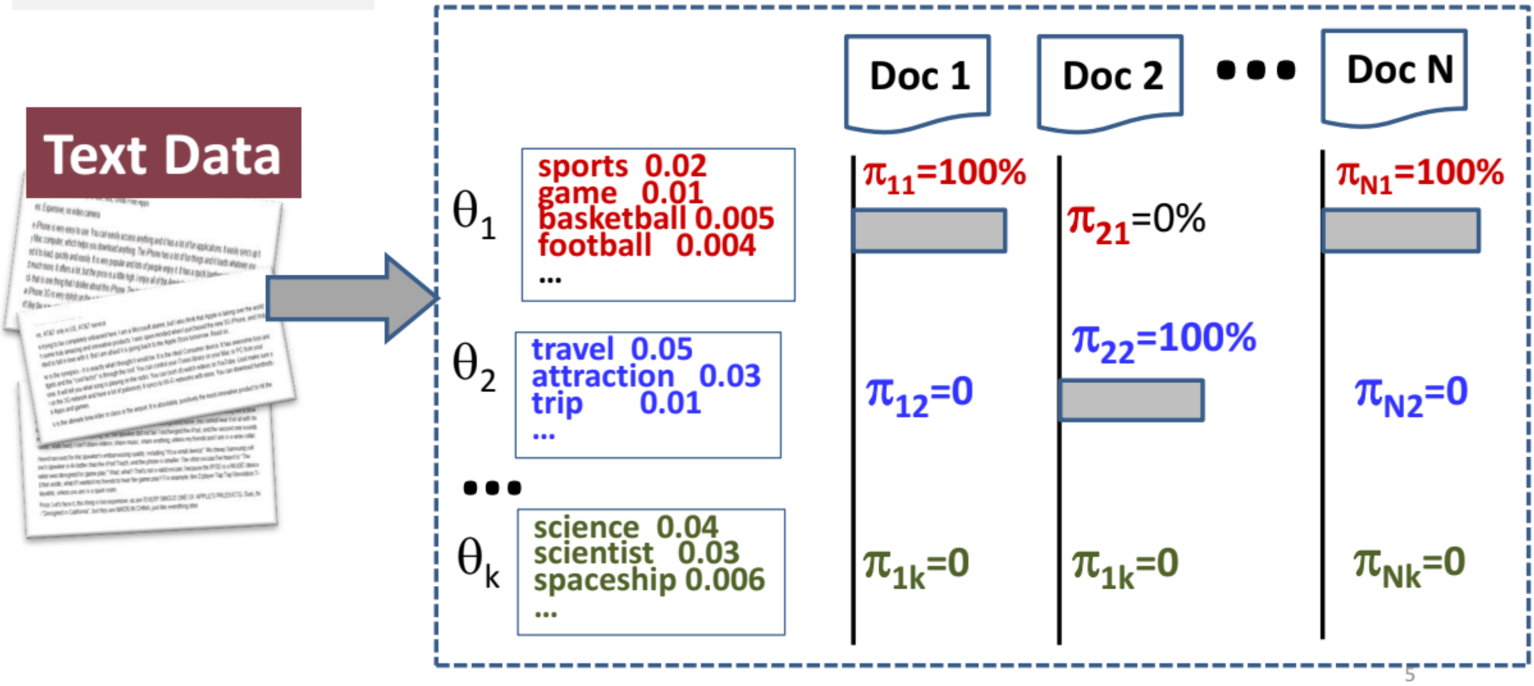
Doc 2: π_{21}=0%; π_{22}; π_{2k}

Doc N: π_{N1}=0%; π_{N2}; π_{Nk}

One Topic(=cluster) Per Document

INPUT: C, k, V

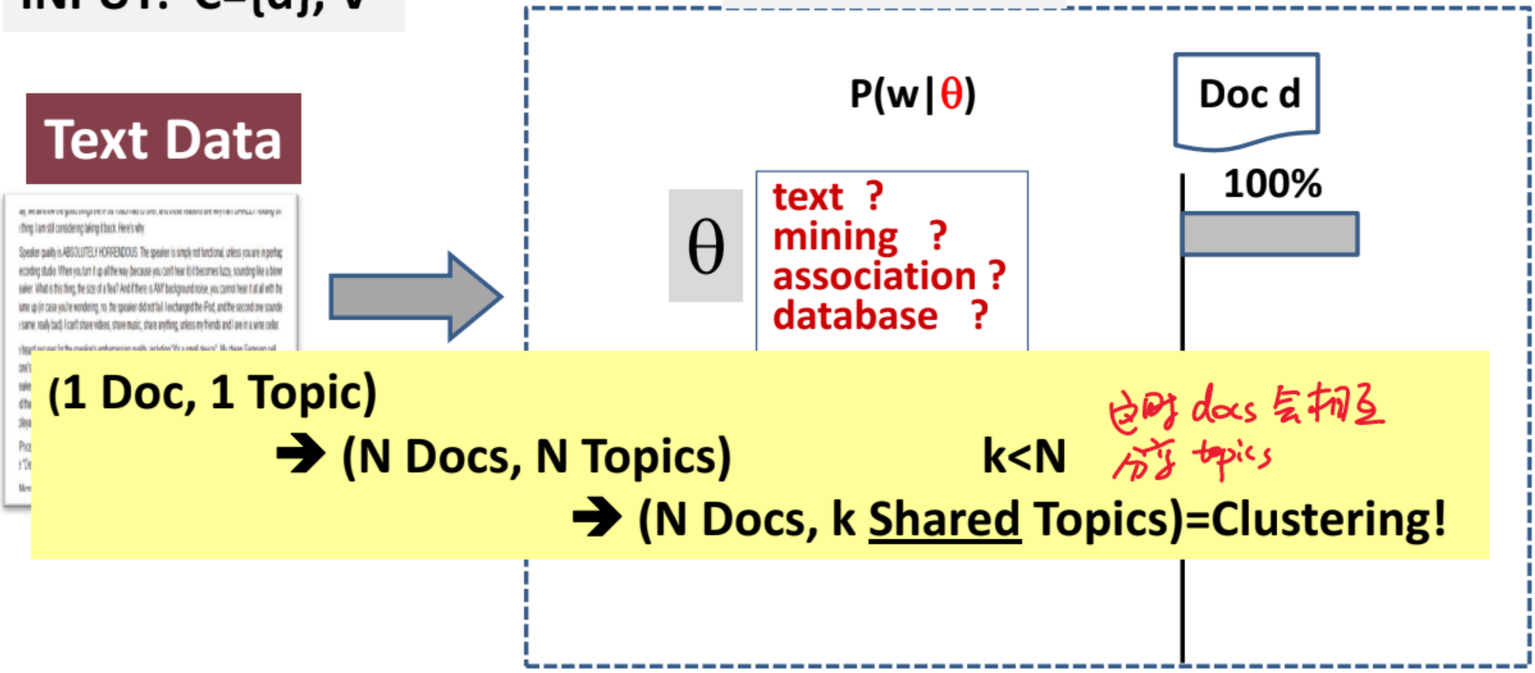
OUTPUT: { θ_1, …, θ_k }, { c_1, …, c_N } $c_i \in [1,k]$ ← 表示属于哪个聚类



Mining One Topic Revisited

INPUT: C={d}, V

OUTPUT: { θ}



(1 Doc, 1 Topic)

➔ (N Docs, N Topics) **k<N**

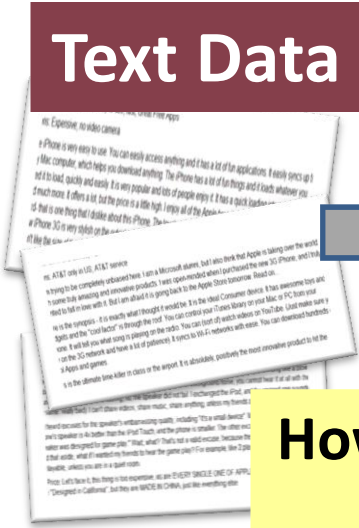
➔ (N Docs, k Shared Topics)=Clustering!

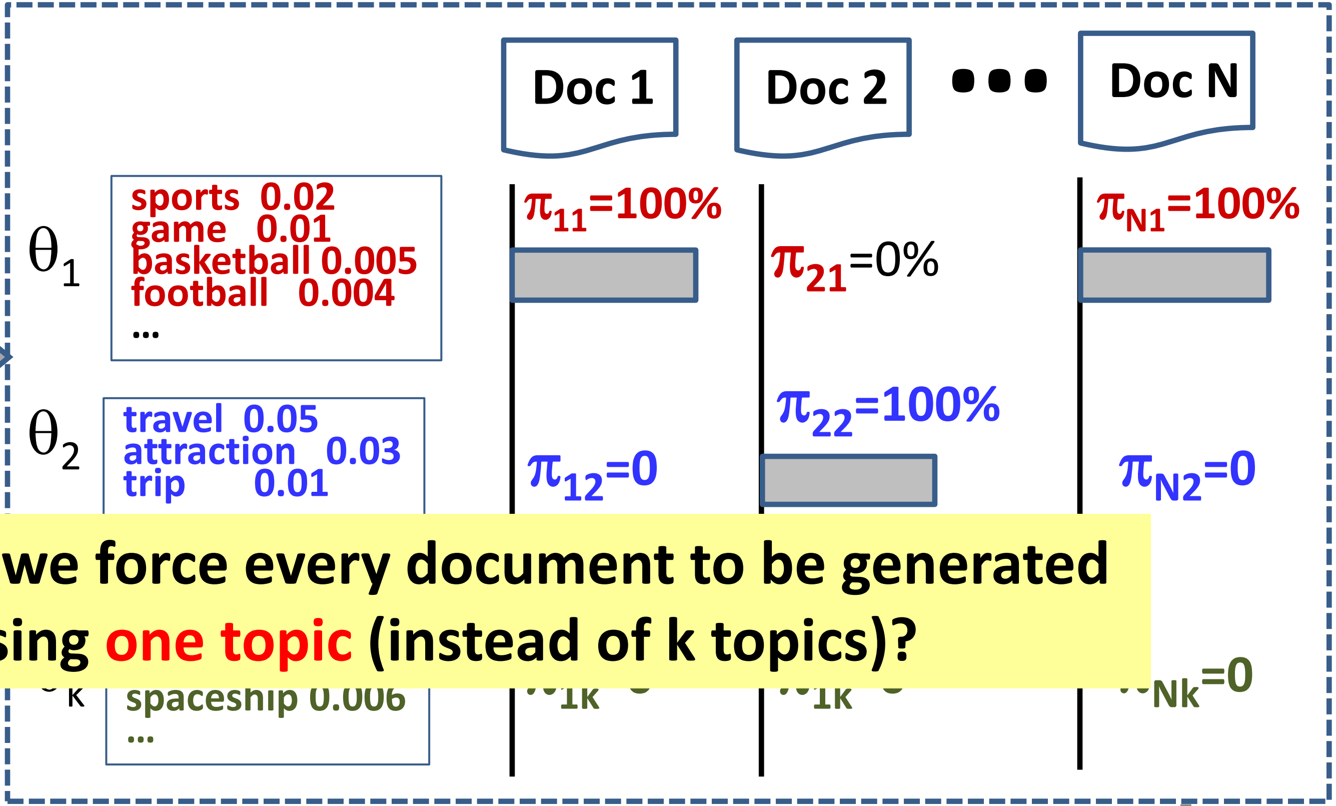
这时 docs 会相互
分享 topics

What Generative Model Can Do Clustering?

INPUT: C, k, V

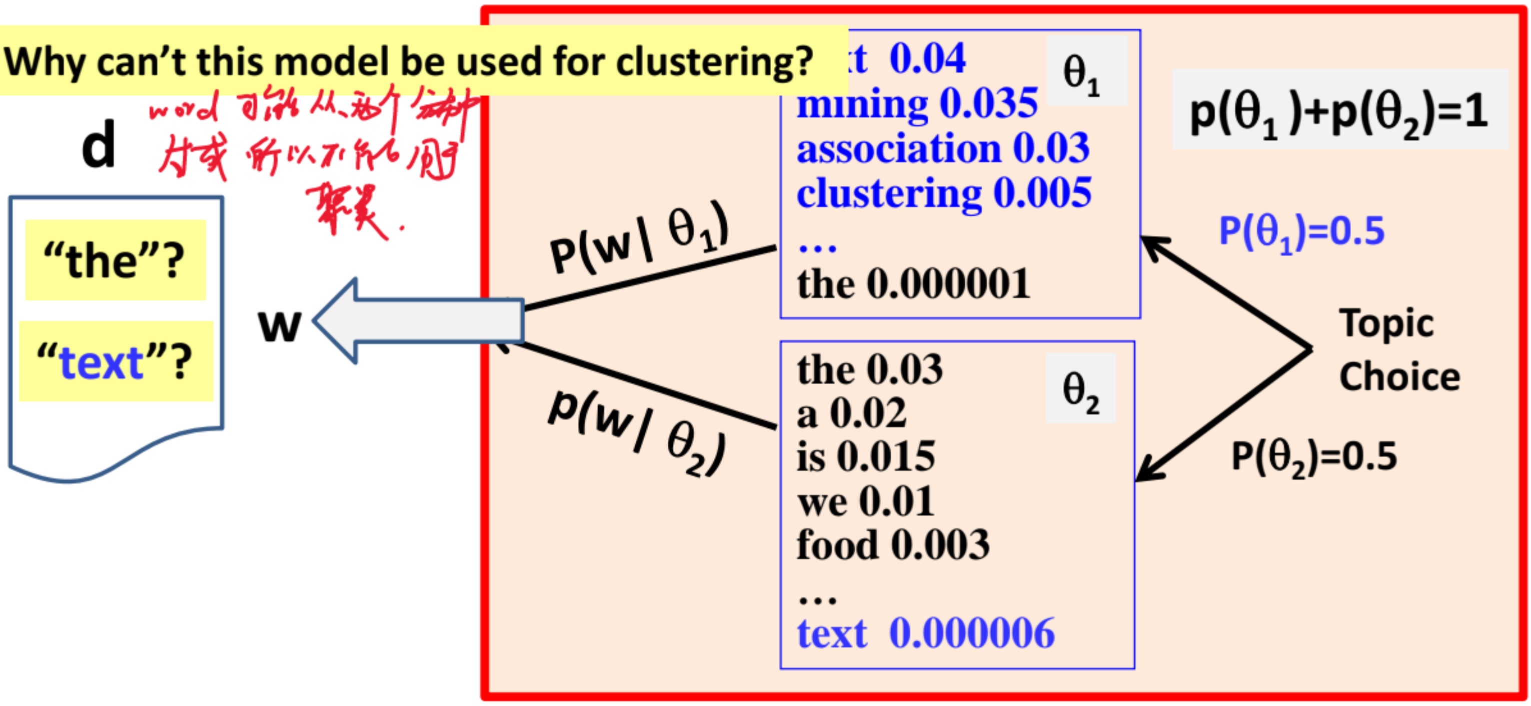
OUTPUT: { θ_1, …, θ_k }, **{ c_1, …, c_N } $c_i \in [1,k]$**





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

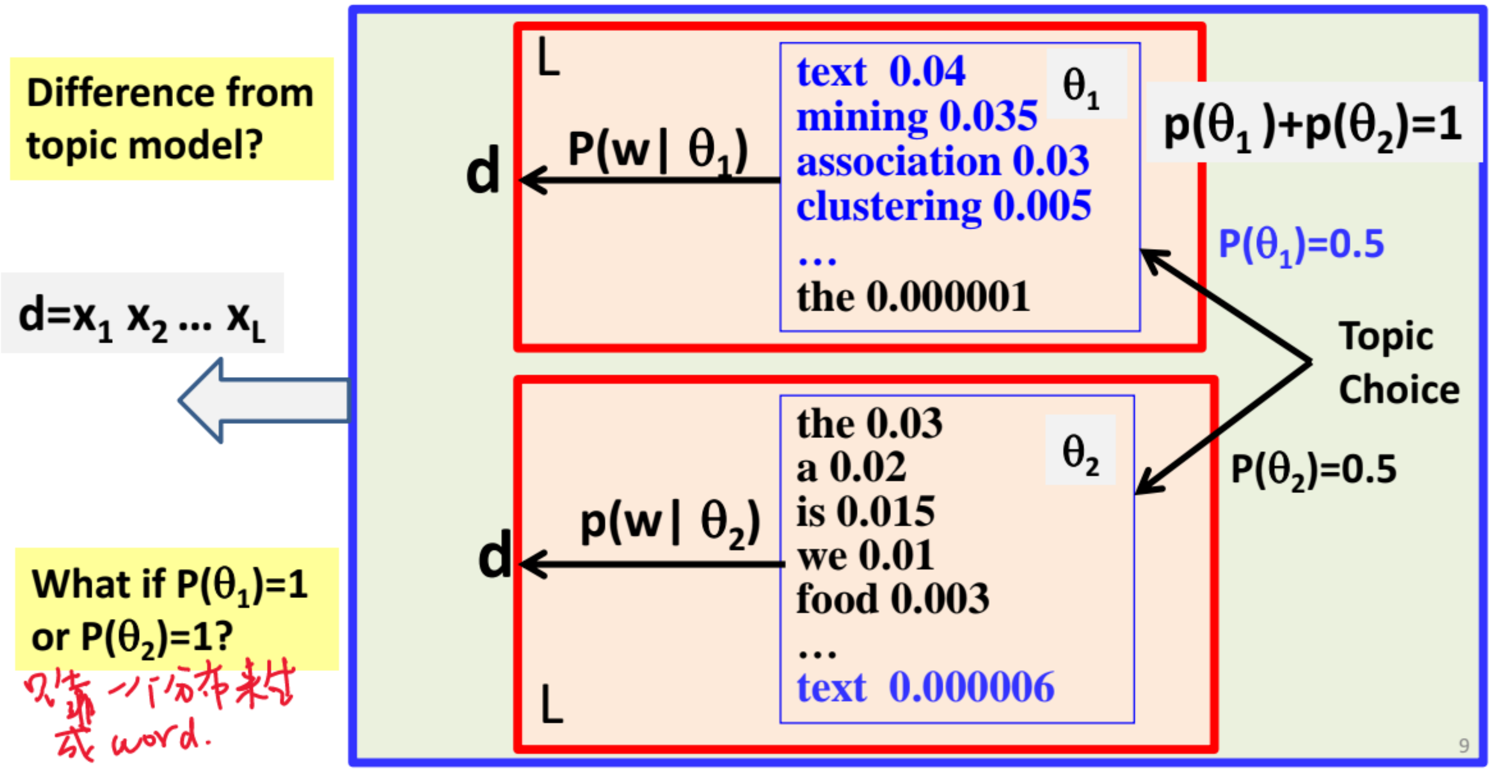
Generative Topic Model Revisited



Why can't this model be used for clustering?

word 可能从两个分布生成 所以不能用 聚类.

Mixture Model for Document Clustering



Difference from topic model?

$d = x_1\ x_2 \ldots x_L$

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

只靠一个分布来生成word.

Likelihood Function: p(d)=?

( )( ) + ( )( )

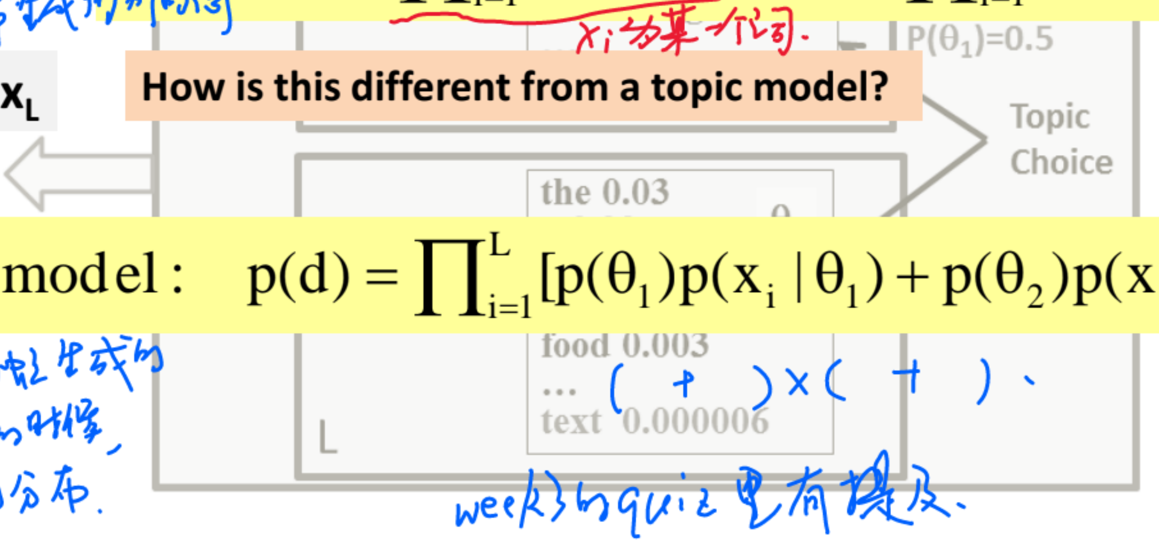
$$p(d) = p(\theta_1)p(d \mid \theta_1) + p(\theta_2)p(d \mid \theta_2)$$

$$= p(\theta_1)\underbrace{\prod_{i=1}^{L} p(x_i \mid \theta_1)}_{x_i为某一个词.} + p(\theta_2)\prod_{i=1}^{L} p(x_i \mid \theta_2)$$

先选择一个分布，
后一直用个分布生成所有的词

d=x_1 x_2 … x_L

How is this different from a topic model?



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

每个词是独立生成的
当考虑每一个词的时候，
再选择需要的分布.

(+)×(+)、

week3的quiz里有提及.