

# Topic Mining and Analysis: Term as Topic

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

# Formal Definition of Topic Mining and Analysis

- Input
  - A **collection** of **N** text documents  **$C = \{d_1, \dots, d_N\}$**
  - **Number of topics:  $k$**
- Output
  - **$k$  topics:  $\{\theta_1, \dots, \theta_k\}$**
  - **Coverage of topics in each  $d_i$ :  $\{\pi_{i1}, \dots, \pi_{ik}\}$**
  - $\pi_{ij}$  = prob. of  $d_i$  covering topic  $\theta_j$

$$\sum_{j=1}^k \pi_{ij} = 1$$

**How to define  $\theta_i$  ?**

# Initial Idea: Topic = Term 平均短語

## Text Data



$\theta_1$  "Sports"

$\theta_2$  "Travel"

...

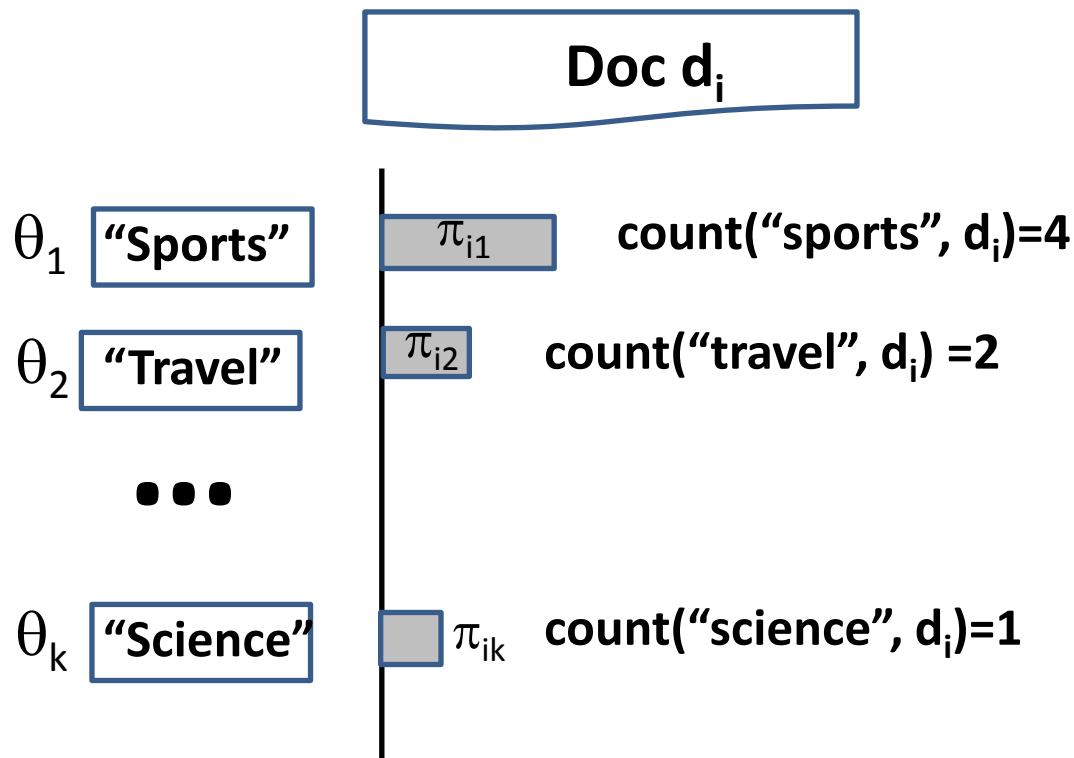
$\theta_k$  "Science"

Doc 1	Doc 2	...	Doc N
30%			
$\pi_{11}$	$\pi_{21}=0$		$\pi_{N1}=0$
$\pi_{12}$	$\pi_{22}$		$\pi_{N2}$
12%			
$\pi_{1k}$	$\pi_{2k}$		$\pi_{Nk}$
8%			

# Mining k Topical Terms from Collection C

- Parse text in C to obtain candidate terms (e.g., term = word).
- Design a scoring function to measure how good each term is as a topic.
  - Favor a representative term (high frequency is favored)
  - Avoid words that are too frequent (e.g., "the", "a").
  - TF-IDF weighting from retrieval can be very useful.
  - Domain-specific heuristics are possible (e.g., favor title words, hashtags in tweets). #
- Pick k terms with the highest scores but try to minimize redundancy.
  - △ – If multiple terms are very similar or closely related, pick only one of them and ignore others.

# Computing Topic Coverage: $\pi_{ij}$



$$\pi_{ij} = \frac{\text{count}(\theta_j, d_i)}{\sum_{L=1}^k \text{count}(\theta_L, d_i)}$$

# How Well Does This Approach Work?

Doc  $d_i$

Cavaliers vs. Golden State Warriors: NBA playoff finals ...  
basketball game ... **travel** to Cleveland ... **star** ...

$\theta_1$  "Sports"

$$\pi_{i1} \propto c(\text{"sports"}, d_i) = 0$$

1. Need to count  
related words also!

$\theta_2$  "Travel"

$$\pi_{i2} \propto c(\text{"travel"}, d_i) = 1 > 0$$

...

2. "Star" can be ambiguous (e.g., star in the sky).

$\theta_k$  "Science"

$$\pi_{ik} \propto c(\text{"science"}, d_i) = 0$$

3. Mine complicated topics?

# Problems with “Term as Topic”

- Lack of expressive power
  - Can only represent simple/general topics
  - Can't represent complicated topics
- Incompleteness in vocabulary coverage
  - Can't capture variations of vocabulary (e.g., related words)
- Word sense ambiguity
  - A topical term or related term can be ambiguous (e.g., basketball star vs. star in the sky)