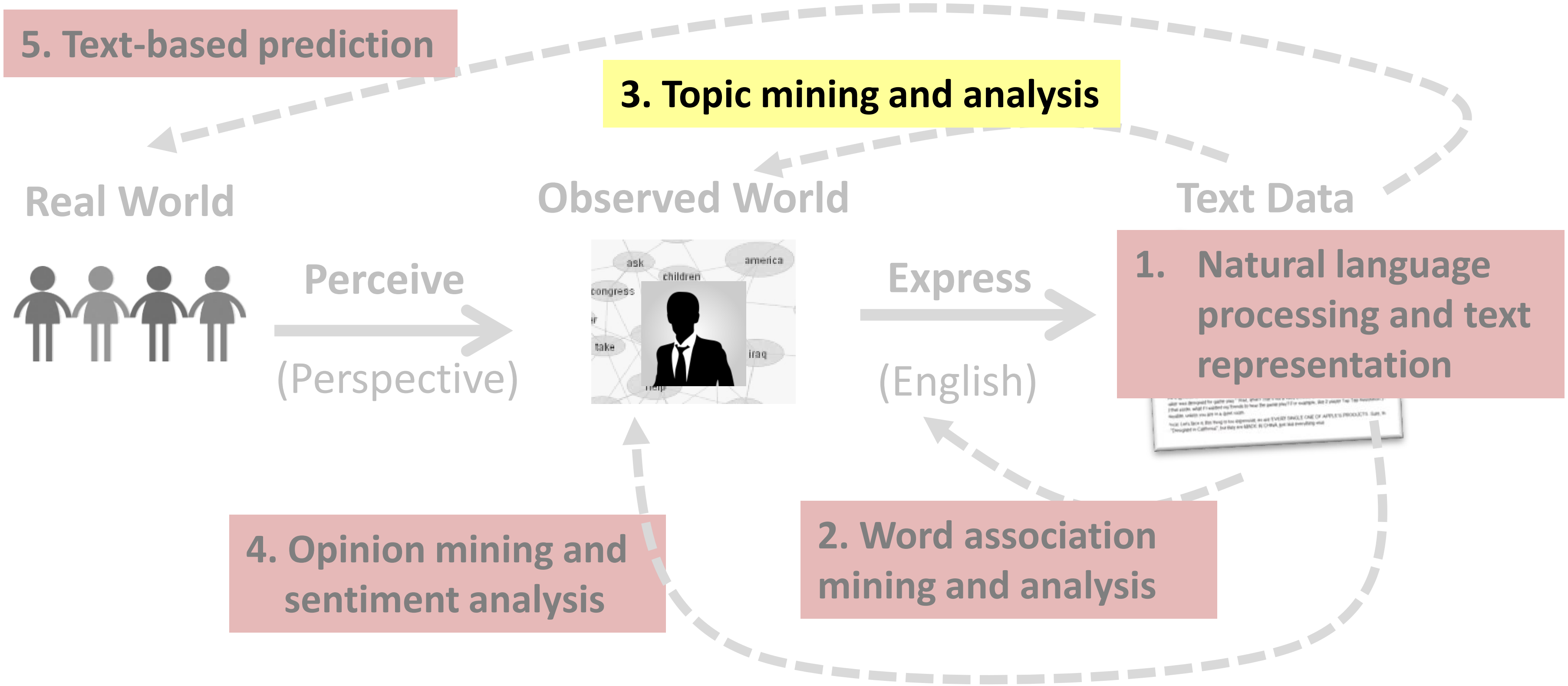
Topic Mining and Analysis: Probabilistic Topic Models

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

Topic Mining and Analysis: Probabilistic Topic Models



Problems with “Term as Topic”

- Lack of expressive power

➔ Topic = {Multiple Words}

  - Can only represent simple/general topics
  - Can’t represent complicated topics
- Incompleteness in vocabulary coverage

+ weights on words

  - Can’t capture variations of vocabulary (e.g., related words)
- Word sense ambiguity

➔ Split an ambiguous word

  - A topical term or related term can be ambiguous (e.g., basketball star vs. star in the sky)

A probabilistic topic model can do all these!

Improved Idea: Topic = Word Distribution

θ_1 **"Sports"**

$P(w|\theta_1)$

sports	0.02
game	0.01
basketball	0.005
football	0.004
play	0.003
star	0.003
…	
nba	0.001
…	
travel	0.0005
…	

θ_2 **"Travel"**

$P(w|\theta_2)$

travel	0.05
attraction	0.03
trip	0.01
flight	0.004
hotel	0.003
island	0.003
…	
culture	0.001
…	
play	0.0002
…	

•••

θ_k **"Science"**

$P(w|\theta_k)$

science	0.04
scientist	0.03
spaceship	0.006
telescope	0.004
genomics	0.004
star	0.002
…	
genetics	0.001
…	
travel	0.00001
…	

$$\sum_{w \in V} p(w|\theta_i) = 1$$

Vocabulary Set: V={w1, w2,….}

Probabilistic Topic Mining and Analysis

- Input
  - A **collection** of **N** text documents **$C=\{d_1, \ldots, d_N\}$**
  - **Vocabulary set: $V=\{w_1, \ldots, w_M\}$**
  - **Number of topics**: **k**
- Output
  - **k topics, each a word distribution**: **$\{ \theta_1, \ldots, \theta_k \}$**
  - **Coverage of topics in each d_i**: **$\{ \pi_{i1}, \ldots, \pi_{ik} \}$**
  - π_{ij}=prob. of d_i covering topic θ_j

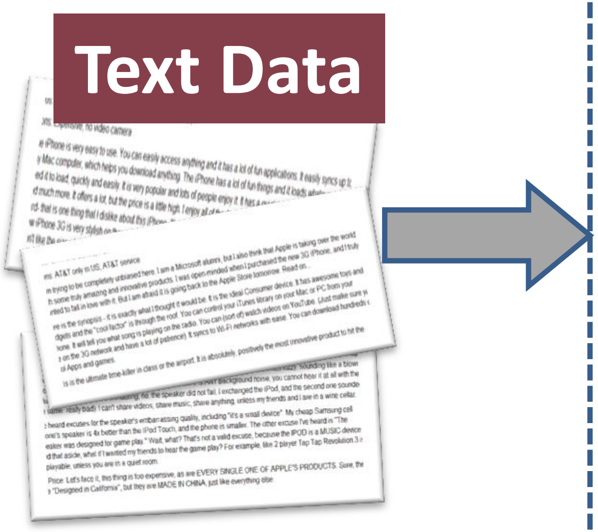
$$\sum_{w \in V} p(w \mid \theta_i) = 1$$

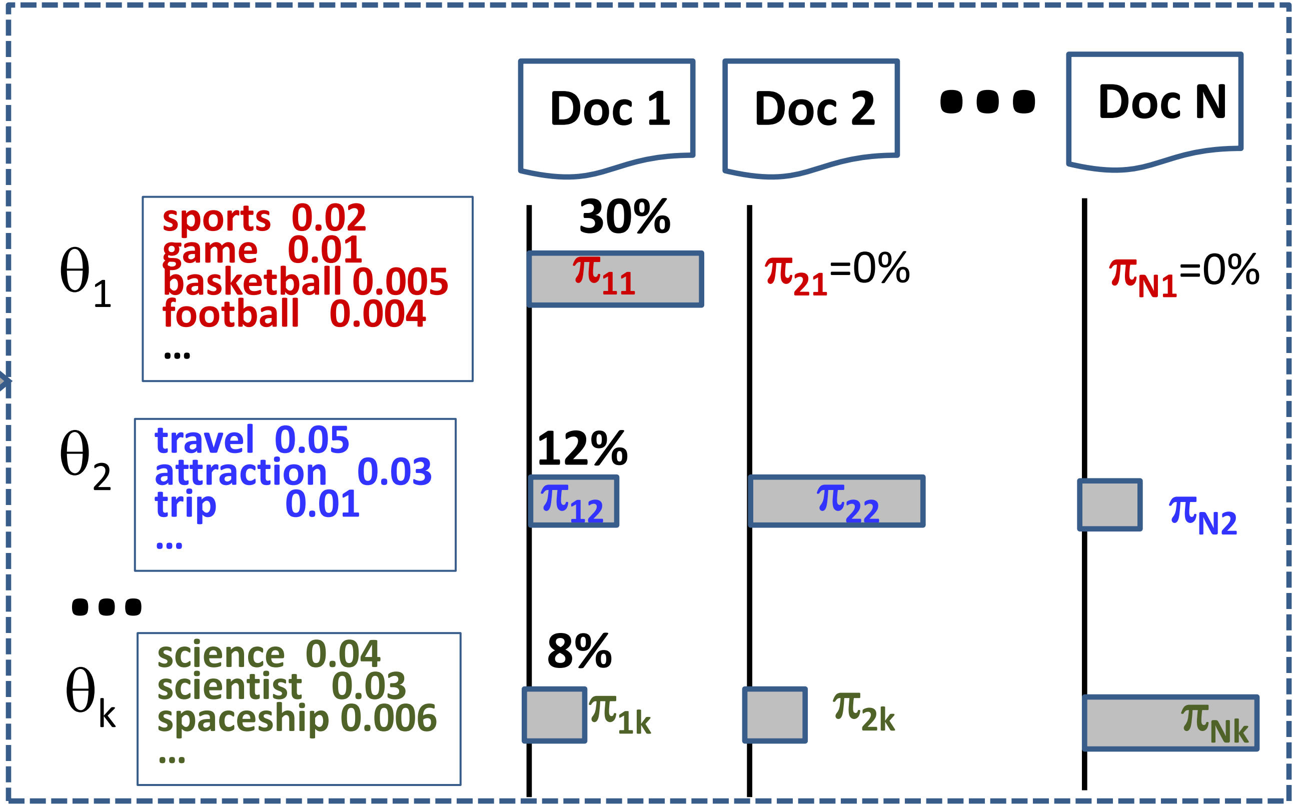
$$\sum_{j=1}^{k} \pi_{ij} = 1$$

The Computation Task

INPUT: C, k, V

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





Generative Model for Text Mining

Modeling of Data Generation: P(Data |Model, Λ)

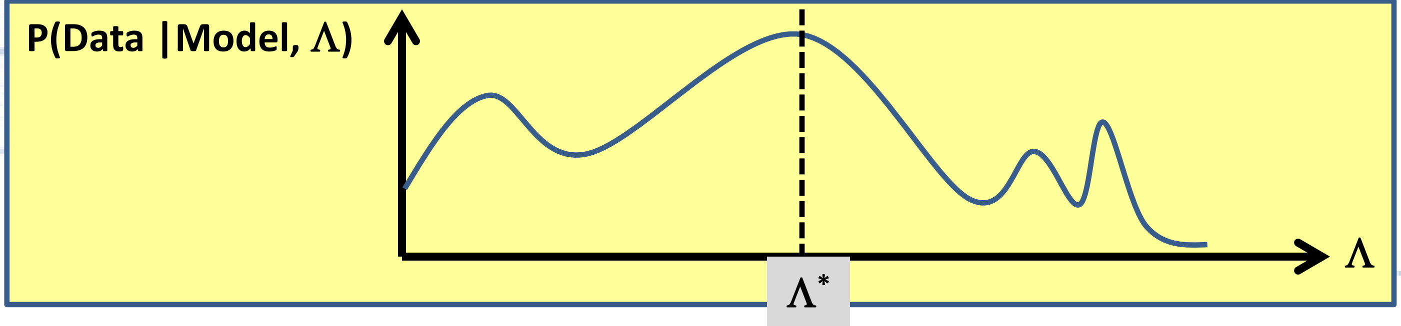
Λ=({ θ_1, …, θ_k }, { π_{11}, …, π_{1k} }, …, { π_{N1}, …, π_{Nk} })

Text Data

How many parameters in total?

Parameter Estimation/ Inferences

$$\Lambda^* = \text{argmax}_{\Lambda}\ p(\text{Data} \mid \text{Model}, \Lambda)$$



Summary

- Topic represented as word distribution
  - Multiple words: allow for describing a complicated topic
  - Weights on words: model subtle semantic variations of a topic
- Task of topic mining and analysis
  - Input: collection C, number of topics k, vocabulary set V
  - Output: a set of topics, each a word distribution; coverage of all topics in each document

$$\Lambda=(\{ \theta_1, \ldots, \theta_k \}, \{ \pi_{11}, \ldots, \pi_{1k} \}, \ldots, \{ \pi_{N1}, \ldots, \pi_{Nk} \})$$

$$\forall j \in [1,k], \sum_{w \in V} p(w \mid \theta_j) = 1$$

$$\forall i \in [1,N], \sum_{j=1}^{k} \pi_{ij} = 1$$

Summary (cont.)

- **Generative model** for text mining
  - **Model data generation** with a prob. model: **P(Data |Model, Λ)**
  - **Infer the most likely parameter values Λ^*** given a particular data set: $\Lambda^* = \arg\max_{\Lambda} p(\text{Data} \mid \text{Model}, \Lambda)$
  - **Take Λ^* as the "knowledge"** to be mined for the text mining problem
  - **Adjust** the design of the model to discover different knowledge