

Probabilistic Topic Mining and Analysis

- Input

- A collection of **N** text documents $C = \{d_1, \dots, d_N\}$
- Vocabulary set: $V = \{w_1, \dots, w_M\}$
- Number of topics: **k**

- Output

- **k** topics, each a word distribution: $\{\theta_1, \dots, \theta_k\}$
- Coverage of topics in each d_i : $\{\pi_{i1}, \dots, \pi_{ik}\}$
- π_{ij} = prob. of d_i covering topic θ_j

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

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