Probabilistic Topic Mining and Analysis

- Input
  - A **collection** of **N** text documents **C={d_1, …, d_N}**
  - **Vocabulary set: V={w_1, …, w_M}**
  - **Number of topics**: **k**
- Output
  - **k topics, each a word distribution**: **{ θ_1, …, θ_k }**
  - **Coverage of topics in each d_i**: **{ π_{i1}, …, π_{ik} }**
  - π_{ij}=prob. of d_i covering topic θ_j

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

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