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

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

$$\forall j \in [1, k], \sum_{w \in V} p(w \mid \theta_j) = 1 \qquad \forall i \in [1, N], \sum_{j=1}^{k} \pi_{ij} = 1$$