Mixture Model for Document Clustering

- Data: a collection of documents $C=\{d_1, \ldots, d_N\}$
- Model: mixture of k unigram LMs: $\Lambda=(\{\theta_i\}; \{p(\theta_i)\})$, $i \in [1,k]$
  - To generate a document, first **choose a θ_i** according to $p(\theta_i)$, and then generate **all** words in the document using $p(w|\theta_i)$
- Likelihood:

$$p(d \mid \Lambda) = \sum_{i=1}^{k} [p(\theta_i) \prod_{j=1}^{|d|} p(x_j \mid \theta_i)]$$

$$= \sum_{i=1}^{k} [p(\theta_i) \prod_{w \in V} p(w \mid \theta_i)^{c(w,d)}]$$

- Maximum Likelihood estimate

$$\Lambda^* = \arg\max_{\Lambda} p(d \mid \Lambda)$$