EM Algorithm for Document Clustering

- Initialization: Randomly set $\Lambda=(\{\theta_i\}; \{p(\theta_i)\})$, $i \in [1,k]$
- **Repeat until likelihood $p(C|\Lambda)$ converges**
  - **E-Step: Infer which distribution has been used to generate document d: hidden variable $Z_d \in [1, k]$**

$$p^{(n)}(Z_d = i \mid d) \propto p^{(n)}(\theta_i) \prod_{w \in V} p^{(n)}(w \mid \theta_i)^{c(w,d)} \qquad \sum_{i=1}^{k} p^{(n)}(Z_d = i \mid d) = 1$$

  - **M-Step: Re-estimation of all parameters**

$$p^{(n+1)}(\theta_i) \propto \sum_{j=1}^{N} p^{(n)}(Z_{d_j} = i \mid d_j) \qquad \sum_{i=1}^{k} p^{(n+1)}(\theta_i) = 1$$

$$p^{(n+1)}(w \mid \theta_i) \propto \sum_{j=1}^{N} c(w, d_j) p^{(n)}(Z_{d_j} = 1 \mid d_j) \qquad \sum_{w \in V} p^{(n+1)}(w \mid \theta_i) = 1, \quad \forall i \in [1,k]$$