Summary of Generative Model for Clustering

- A slight variation of topic model can be used for clustering documents
  - Each **cluster** is represented by a **unigram LM** $\mathbf{p(w|\theta_i)}$ ➔ **Term cluster**
  - A document is generated by first choosing a unigram LM and then generating **ALL words** in the document using this **single LM**
  - Estimated model parameters give both a topic characterization of each cluster and a probabilistic assignment of a document into each cluster
  - "Hard" clusters can be obtained by forcing a document into the cluster corresponding to the unigram LM most likely used to generate the document
- EM algorithm can be used to compute the ML estimate
  - Normalization is often needed to avoid underflow