Document clustering:

Derivation of Expectation and Maximization:

The training algorithm of EM is to find the maximum likelihood.

$$\ln P(X|\theta) = \ln \sum P(X, Z|\theta)$$

For computing likelihood, we will need Z, now we will calculate posterior probability, $P(Z|X,\theta)$.

The Expection is :

$$Q(\theta, \theta^{old}) = \sum_{Z} P(Z|X, \theta^{old}) \cdot \ln P(X, Z|\theta)$$

For Generative Model :

→ toss K-face(dice) with ϕ to choose the face K that n^{th} document belongs to.

→ For each placeholder :

- generate the word by tossing dice with parameter μ_K, corresponding to K.

The parameters of model are :

→ cluster proportion ϕ where $\sum_{K=1}^{K} \phi_K = 1$

→ word proportion μ_K for cluster K, where

$$\sum_{K=1}^{K} \mu_K = 1$$

The probability of generating a pair & cluster (k,d):

$$p(k,d) = p(k)\,p(d|k)$$
$$= \phi_k \cdot \prod_{w \in d} \mu_{k,w}^{c(w,d)}$$

where $c(w,d)$ is no. of occurences.

For incomplete data, the document cluster is not given, the latent variable z_n.

∴ The probability of document is,

$$p(d_1, d_2 \dots d_N) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} p(z_{n,k}=1, d_n)$$
$$= \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \left(\phi_k \prod \mu_{k,w}^{c(w,d)}\right)$$

The expectation function, $Q(\theta, \theta^{old})$:

$$Q(\theta,\theta^{old}) = \sum_{n=1}^{N}\sum_{k=1}^{K} p(z_{n,k}=1|d_n,\theta^{old}) \ln p(z_{n,k}=1, d_n)$$
$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{n,k})\left(\ln\phi_k + \sum_{w \in A} c(w,d) \ln \mu_{k,w}\right)$$

where $\gamma(z_{n,k}) = p(z_{n,k}=1 \mid d_n, \theta^{old})$ represents the responsibility factor, assuming all words came from a dictionary A.

The probability that a doc. belongs to a cluster, gamma,

$$\gamma(z_n, k) = \frac{\prod_{n=1}^{N} \phi_k \prod_{w \in A} \mu_{k,w}^{c(w,d)}}{\prod_{n=1}^{N} \sum_{k=1}^{K} (\phi_k \prod_{w \in A} \mu_{k,w}^{c(w,d)})}$$

To maximize the Q function,

→ Mixing components : $\phi_k = N_k / N$ where $N_k = \sum_{n=1}^{N} \gamma(z_n, k)$

→ The word proportion parameter:

$$\mu_{k,w} = \frac{\sum_{n=1}^{N} \gamma(z_{n,k})\, c(w, d_n)}{\sum_{w' \in A} \sum_{n=1}^{N} \gamma(z_{n,k})\, c(w', d_n)}$$

Hence, soft EM algo for document clustering is:

- choose an initial setting : $\theta^{old} = (\phi^{old}, \mu_1^{old}, \ldots \mu_K^{old})$
- while convergence is not met :
  - E step : set $\forall_n, \forall_k : \gamma(z_{n,k})$ based on θ^{old}
  - M step : set θ^{new} based on above equations.
  - $\theta^{old} \leftarrow \theta^{new}$

Hard EM algo for document clustering is :

- choose an initial setting for θ^{old}
- while convergence not met :
  - E step : set $z_n \leftarrow \text{argmax}_z\, p(z | x, \theta^{old})$
  - M step : set $\theta^{new} \leftarrow \text{argmax}_\theta\, p(X, z^* | \theta)$
  - $\theta^{old} \leftarrow \theta^{new}$