Applications of Mixture Models for Text Mining

Likelihood:

$$p(d \mid \theta_1 \oplus \theta_2) = \prod_{w \in V} [\lambda p(w \mid \theta_1) + (1-\lambda) p(w \mid \theta_2)]^{c(w,d)}$$

$$\log p(d \mid \theta_1 \oplus \theta_2) = \sum_{w \in V} c(w,d) \log[\lambda p(w \mid \theta_1) + (1-\lambda) p(w \mid \theta_2)]$$

Application Scenarios:

-p(w|θ_1) & p(w|θ_2) are known; estimate λ

The doc is about text mining and food nutrition, how much percent is about text mining?

-p(w|θ_1) & λ are known; estimate p(w|θ_2)

30% of the doc is about text mining, what's the rest about?

-p(w|θ_1) is known; estimate λ & p(w|θ_2)

The doc is about text mining, is it also about some other topic, and if so to what extent?

-λ is known; estimate p(w|θ_1)& p(w|θ_2)

30% of the doc is about one topic and 70% is about another, what are these two topics?

-Estimate λ, p(w|θ_1), p(w|θ_2)

The doc is about two subtopics, find out what these two subtopics are and to what extent the doc covers each.