Generative model of texts

Probabilistic Latent Semantic Analysis (PLSA):

$$p(w|d) = \sum_{t \in T} p(w|t,d)\, p(t|d) = \sum_{t \in T} p(w|t)\, p(t|d)$$

Law of total probability

$$p(w) = \sum_{t \in T} p(w|t)\, p(t)$$

Assumption of conditional independence

$$p(w|t,d) = p(w|t)$$

Notation:

- *w – word*
- *d – document*
- *t – topic*