- Гипотеза условной независимости. Появление слов в документе d, относящихся к теме t, описывается общим для всей коллекции распределением p (w I t) и не зависит от документа.
- Вероятностная модель. Согласно формуле полной вероятности и гипотезе условной независимости

$$p(w \mid d) = \sum p(t \mid d)\, p(w \mid t)$$

- Алгоритм 1:
- Вход: распределения p (t I d), p (w I t).

1 для всех d

задать длину n документа d;

2 для всех i=1…n

выбрать случайную тему t из распределения p(t I d),

выбрать случайный термин w из распределения p(w I t).

Добавить в выборку пару (d, w). Тема забывается.

Выход: выборка пар (d_i, w_i), где i=1, …, n.