Sobre Muestreo de Gibbs

En LDA, cada documento es modelado como una mixtura sobre K tópicos latentes, siendo cada uno una distribución multinomial sobre un vocabulario de W palabras.

Para generar un nuevo documento j, calculamos una proporción $\theta_{k|j}$ a partir de una distribución Dirichlet con parámetro α. Para la i-ésima palabra en el documento, una asignación al tópico z_{ij} se realiza con el tópico k escogido con probabilidad $\theta_{k|j}$. Entonces la palabra x_{ij} se genera a partir del tópico z_{ij}, con x_{ij} tomando valores w con probabilidad $\phi_{w|k}$, donde $\phi_{w|k}$ se genera a partir de una distribución Dirichlet con parámetro β. Finalmente el proceso generativo viene dado por:

$$\theta_{k|j} \sim Dir(\alpha) \qquad \phi_{w|k} \sim Dir(\beta) \qquad z_{ij} \sim theta_{k|j} \qquad x_{ij} \sim \phi_{w|z_{ij}}$$

donde $Dir(\alpha)$ representa la distribución de Dirichlet.

Dado el conjunto de entrenamiento de N palabras $x = x_{ij}$, es posible inferir la distribución posterior de variables latentes. Un procedimiento eficiente es usar **Muestreo de Gibbs Colapsado**, que toma una muestra de variables latentes $z = z_{ij}$ mediante $\theta_{k|j}$ y $\phi_{w|k}$.

La probabilidad de z_{ij} es calculada como sigue:

$$p(z_{ij} = k|z^{\neg ij}, x, \alpha, \beta) \propto (\alpha + n_{k|j}^{\neg ij})(\beta + n_{x_{ij}|k}^{\neg ij})(W\beta + n_k^{\neg ij})^{-1}$$

donde el superíndice $\neg ij$ significa que el correspondiente ítem es excluido del conteo de valores, $n_{k|j}$ denota el conteo del documento j asignado al tópico k, $n_{x_{ij}|k}$ denota el conteo de la palabra w asignada al tópico k y $n_k = \sum_w n_{w|k}$