Text Mining – Assignment #3

Roger Cuscó, Matthew-Sudmann-Day and Miquel Torrens

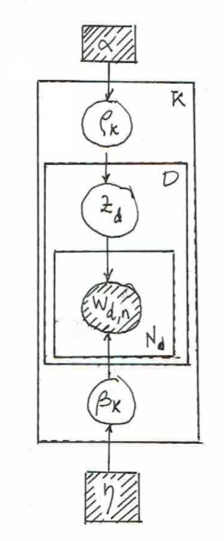
Exercise 1

Text.

Exercise 2

Part (a)

The directed graph is the following:



Part (b)

The Markov blankets of these elemets of the model can be expressed as follows:

- Words in document d: topic assignments z_d (parent) and topics β_k (parent).
- Topic assignment z_d: topic probabilities ρ_k (parent), the set of words $w_{d,n}$ (children) and topics β_k (children's parent).
- Topics β_k: hyperparameter η (parent), the set of words $w_{d,n}$ (children) and topic assignment z_d (children's parent).

Part (c)

An uncollapsed Gibbs algorithm could be the following:

1. Set values for $\boldsymbol{\eta}, \boldsymbol{\alpha} \in \mathbb{R}^K$
2. Draw for each topic $k \in \{1, \ldots, K\}$ a sample $\boldsymbol{\beta}_k \sim \text{Dir}(\boldsymbol{\eta}) \in \Delta^{V-1}$
3. Draw a sample $\boldsymbol{\rho} \sim \text{Dir}(\boldsymbol{\alpha}) \in \Delta^{K-1}$ that specifies the likelihood of each topic
4. Draw for each document $d \in \{1, \ldots, D\}$ a sample $z_d \sim \text{multinom}\,(\boldsymbol{\rho})$
5. Draw for each word $n \in \{1, \ldots, N_d\}$ in document d the word $w_{d,n} \sim \text{multinom}\left(\boldsymbol{\beta}_{z_d}\right)$
6. Update for each k the vector $\boldsymbol{\beta}_k \sim \text{Dir}(\boldsymbol{\eta} + \mathbf{m}_k) \in \Delta^{V-1}$, where element v of vector $\mathbf{m}_k \in \mathbb{R}^V$ is $m_{k,v}$, the number of times topic k generates word v.
7. Update the vector $\boldsymbol{\rho} \sim \text{Dir}(\boldsymbol{\alpha} + \boldsymbol{\delta}) \in \Delta^{K-1}$, where element k of vector $\boldsymbol{\delta} \in \mathbb{R}^K$ is δ_k, the number of documents that are assigned topic k.
8. Return to step 4 and repeat until convergence.