Hierarchical Dirichlet Processes

Yee Whye Teh, Michael I. Jordan
Matthew J. Beal, David M. Blei
Presented By : Qiang Fu

Outline

- Introduction
- Hierarchical Dirichlet Process (HDP)
- Representations of HDP
- Inference
- Experiments
- Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM)

Introduction

- Problem Setting
  - Groups of data
  - Observations within a group = Mixture Model
  - Mixture components are shared
- Assumption :
  - number of mixture components unknown
  - Exchangeability

HDP

- Consider a DP for each group
- One Simple Solution:

$$G_j \mid \alpha_o, G_o(\gamma) \sim DP(\alpha_o, G_o(\gamma)) \quad \text{for each j}$$

- But doesn't work all the time
- Stick-Breaking Construction:

$$G \sim \mathrm{DP}(\alpha_0, G_0) \qquad G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

HDP

- HDP:

$$G_0 \mid \gamma, H \sim \mathrm{DP}(\gamma, H)$$
$$G_j \mid \alpha_0, G_0 \sim \mathrm{DP}(\alpha_0, G_0) \qquad \text{for each } j,$$

- Probability Model (Generative Process):

$$\theta_{ji} \mid G_j \sim G_j \qquad \text{for each } j \text{ and } i,$$
$$x_{ji} \mid \theta_{ji} \sim F(\theta_{ji}) \qquad \text{for each } j \text{ and } i,$$

Stick-Breaking Construction for DP

- Measures drawn from a Dirichlet process are discrete with probability one.

$$\pi'_k \mid \alpha_0, G_0 \sim \text{Beta}(1, \alpha_0) \qquad \phi_k \mid \alpha_0, G_0 \sim G_0$$

$$\pi_k = \pi'_k \prod_{l=1}^{k-1} (1 - \pi'_l) \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

- Notation : $\pi \sim \text{GEM}(\alpha_0)$

Stick-Breaking Construction for HDP

- G_o can be expressed as : $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$
- G_j can be expressed similarly : $G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$
- Let $(A_1, \ldots, A_r)$ be a measurable partition on Θ
- Define $K_l = \{k : \phi_k \in A_l\}$ for $l = 1, \ldots, r$
- $(K_1, \ldots, K_r)$ is a finite partitions of positive integers

Stick-Breaking Construction for HDP

- For each j, we have:

$$(G_j(A_1),\ldots,G_j(A_r)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1),\ldots,\alpha_0 G_0(A_r))$$

$$\Rightarrow \left(\sum_{k\in K_1}\pi_{jk},\ldots,\sum_{k\in K_r}\pi_{jk}\right) \sim \mathrm{Dir}\left(\alpha_0\sum_{k\in K_1}\beta_k,\ldots,\alpha_0\sum_{k\in K_r}\beta_k\right)$$

Stick-Breaking Construction for HDP

- Derive the explicit relationship
- For a partition $\left(\{1,\ldots,k-1\},\{k\},\{k+1,k+2,\ldots\}\right)$.

$$\left(\sum_{l=1}^{k-1}\pi_{jl},\pi_{jk},\sum_{l=k+1}^{\infty}\pi_{jl}\right)\sim \mathrm{Dir}\left(\alpha_0\sum_{l=1}^{k-1}\beta_l,\alpha_0\beta_k,\alpha_0\sum_{l=k+1}^{\infty}\beta_l\right)$$

- Remove the first element:

$$\frac{1}{1-\sum_{l=1}^{k-1}\pi_{jl}}\left(\pi_{jk},\sum_{l=k+1}^{\infty}\pi_{jl}\right)\sim \mathrm{Dir}\left(\alpha_0\beta_k,\alpha_0\sum_{l=k+1}^{\infty}\beta_l\right)$$

Stick-Breaking Construction for HDP

- Define : $\pi'_{jk} = \frac{\pi_{jk}}{1-\sum_{l=1}^{k-1}\pi_{jl}}$
- Observe that : $1 - \sum_{l=1}^{k}\beta_l = \sum_{l=k+1}^{\infty}\beta_l$
- We have :

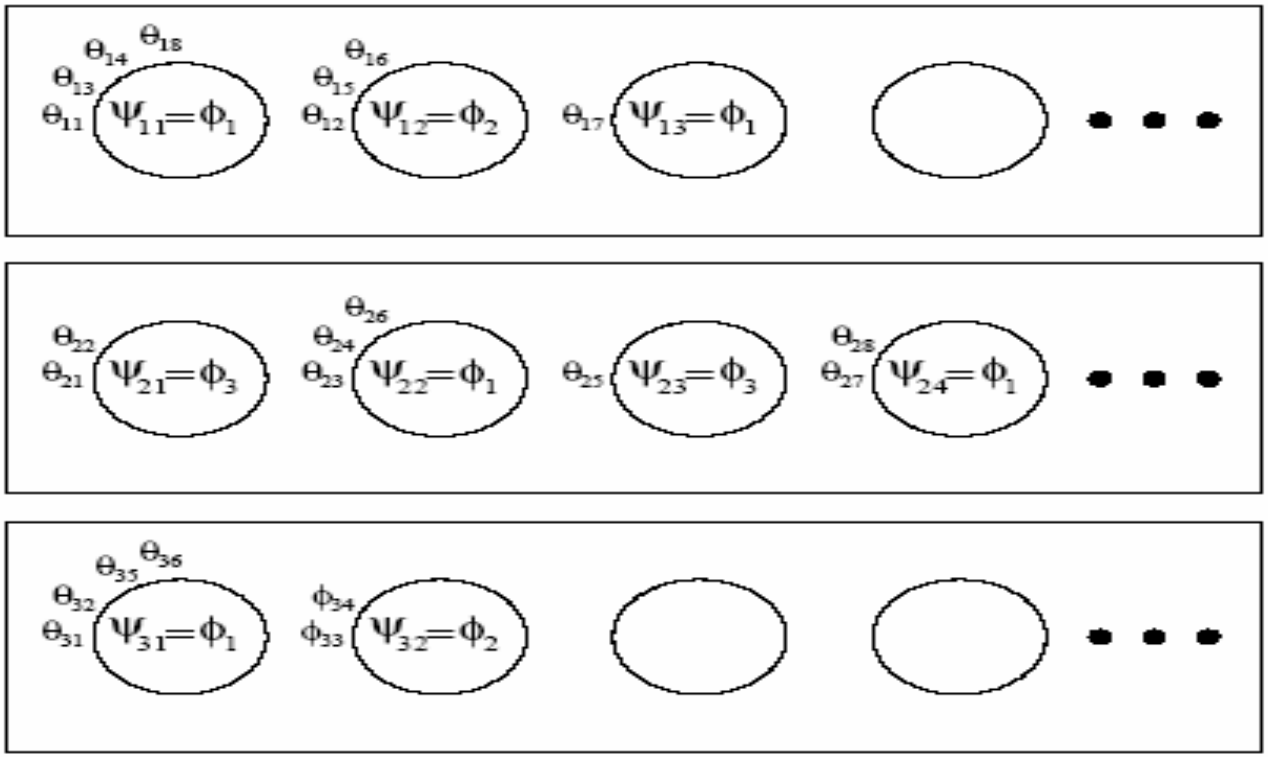
$$\pi'_{jk} \sim \text{Beta}\left(\alpha_0\beta_k, \alpha_0\left(1-\sum_{l=1}^{k}\beta_l\right)\right) \qquad \pi_{jk} = \pi'_{jk}\prod_{l=1}^{k-1}(1-\pi'_{jl})$$

Chinese Restaurant Process

- Clustering effect of DP
- The metaphor
- After integrate out G, we have :

$$\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \alpha_0, G_0 \sim \sum_{k=1}^{K} \frac{m_k}{i-1+\alpha_0} \delta_{\phi_k} + \frac{\alpha_0}{i-1+\alpha_0} G_0$$

Chinese Restaurant Franchise



Chinese Restaurant Franchise

- After G_j is integrated out :

$$\theta_{ji} \mid \theta_{j1}, \ldots, \theta_{j,i-1}, \alpha_0, G_0 \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{i-1+\alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{i-1+\alpha_0} G_0$$

- After G_o is integrated out :

$$\psi_{jt} \mid \psi_{11}, \psi_{12}, \ldots, \psi_{21}, \ldots, \psi_{j\,t-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma} \delta_{\phi_k} + \frac{\gamma}{m_{\cdot\cdot}+\gamma} H$$

Posterior Sampling in the CRF

- Sample **t**
- Integrate out the possible values of k_{it}^{new}

$$p(x_{ji} \mid t^{-ji}, t_{ji} = t^{\text{new}}, k) = \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma} f_k^{-x_{ji}}(x_{ji}) + \frac{\gamma}{m_{\cdot\cdot} + \gamma} f_{k^{\text{new}}}^{-x_{ji}}(x_{ji})$$

- Then :

$$p(k_{jt^{\text{new}}} = k \mid t, k^{-jt^{\text{new}}}) \propto \begin{cases} m_{\cdot k} f_k^{-x_{ji}}(x_{ji}) & \text{if } k \text{ previously used,} \\ \gamma f_{k^{\text{new}}}^{-x_{ji}}(x_{ji}) & \text{if } k = k^{\text{new}}. \end{cases}$$

$$p(t_{ji} = t \mid t^{-ji}, k) \propto \begin{cases} n_{jt\cdot}^{-ji} f_{k_{jt}}^{-x_{ji}}(x_{ji}) & \text{if } t \text{ previously used,} \\ \alpha_0 p(x_{ji} | t^{-ji}, t_{ji} = t^{\text{new}}, k) & \text{if } t = t^{\text{new}}. \end{cases}$$

Posterior Sampling in the CRF

- Sample **k w**ill be similar :

$$p(k_{jt} = k \mid t, k^{-jt}) \propto \begin{cases} m_{\cdot k}^{-jt} f_k^{-\boldsymbol{x}_{jt}}(x_{jt}) & \text{if } k \text{ is previously used,} \\ \gamma f_{k^{\text{new}}}^{-\boldsymbol{x}_{jt}}(x_{jt}) & \text{if } k = k^{\text{new}}. \end{cases}$$

- θ_{ji} and Ψ_{ji} can be reconstructed from these index variables

Posterior sampling with an augmented representation

- Based on the Dirichlet Posterior Distribution:

$$G|\theta_1,\ldots,\theta_n \sim \mathrm{DP}\left(\alpha+n, \frac{\alpha}{\alpha+n}H + \frac{n}{\alpha+n}\frac{\sum_{i=1}^{n}\delta_{\theta_i}}{n}\right)$$

- Rewrite it : G_o is distributed as

$$\mathrm{DP}\left(\gamma+m_{\cdot\cdot}, \frac{\gamma H+\sum_{k=1}^{K} m_{\cdot k}\delta_{\phi_k}}{\gamma+m_{\cdot\cdot}}\right)$$

Posterior sampling with an augmented representation

- Construct G_o :

$$\beta = (\beta_1, \ldots, \beta_K, \beta_u) \sim \mathrm{Dir}(m_{\cdot 1}, \ldots, m_{\cdot K}, \gamma)$$

$$G_u \sim \mathrm{DP}(\gamma, H) \qquad G_0 = \sum_{k=1}^{K} \beta_k \delta_{\phi_k} + \beta_u G_u$$

- Sampling for **t** and **k** will be similar to the previous algorithm

Posterior Sampling by Direct Assignment

- No Bookkeeping
- Sample **z**

$$p(z_{ji} = k \mid z^{-ji}, m, \beta) = \begin{cases} (n_{j\cdot k}^{-ji} + \alpha_0 \beta_k) f_k^{-x_{ji}}(x_{ji}) & \text{if } k \text{ previously used,} \\ \alpha_0 \beta_u f_{k^{\text{new}}}^{-x_{ji}}(x_{ji}) & \text{if } k = k^{\text{new}}. \end{cases}$$

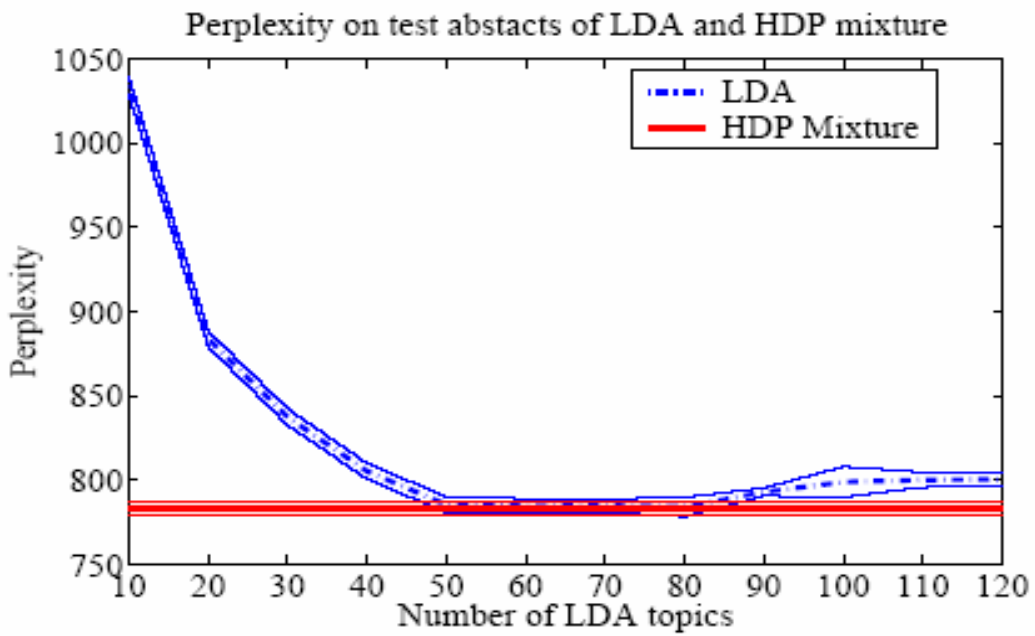
- Sample **m**

$$p(t_{ji} = t | k_{jt} = k, t^{-ji}, \boldsymbol{k}, \beta) \propto n_{jt\cdot}^{-ji}$$

$$p(t_{ji} = t^{\text{new}} | k_{jt^{\text{new}}} = k, t^{-ji}, \boldsymbol{k}, \beta) \propto \alpha_0 \beta_k$$

Experiment – Document Modeling

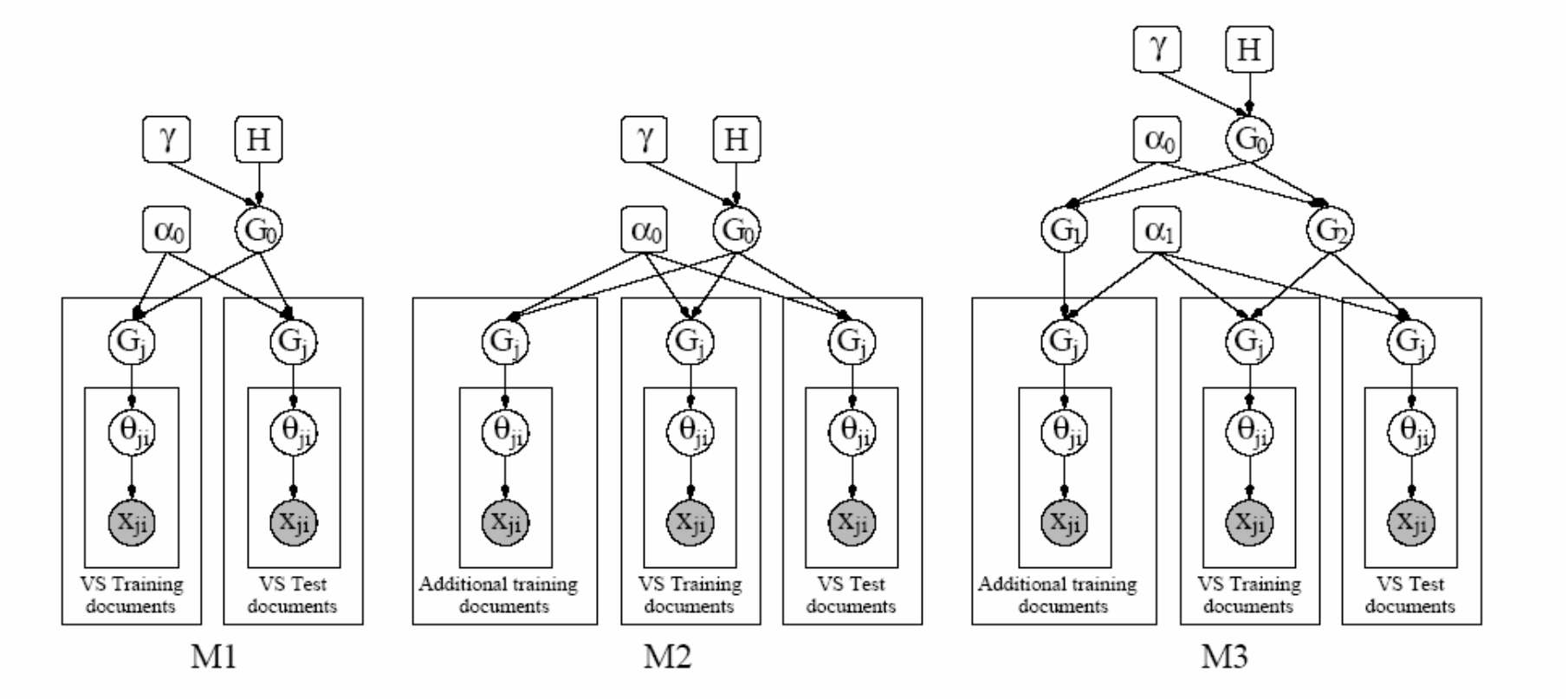
- HDP picks the number of topics for LDA



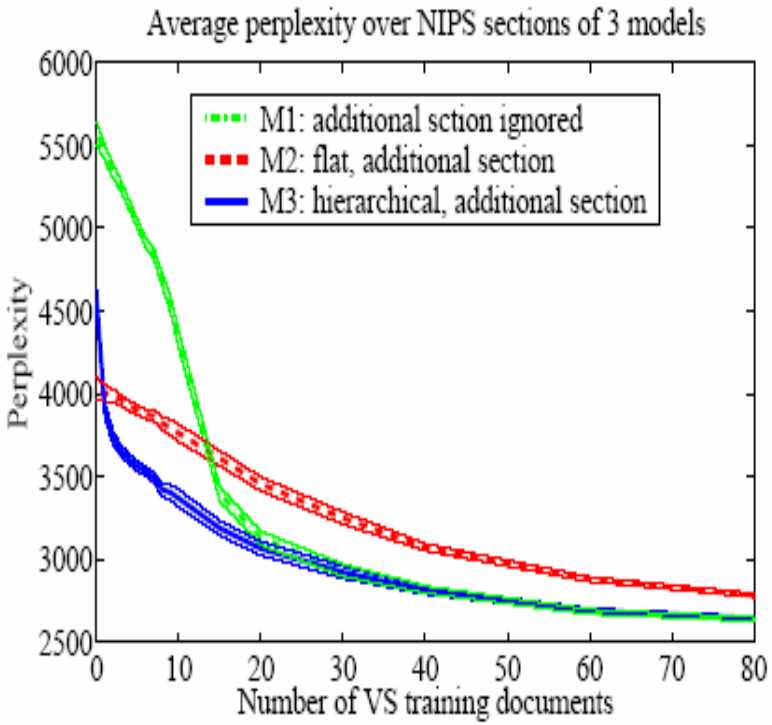
Experiment – Multiple Corpora

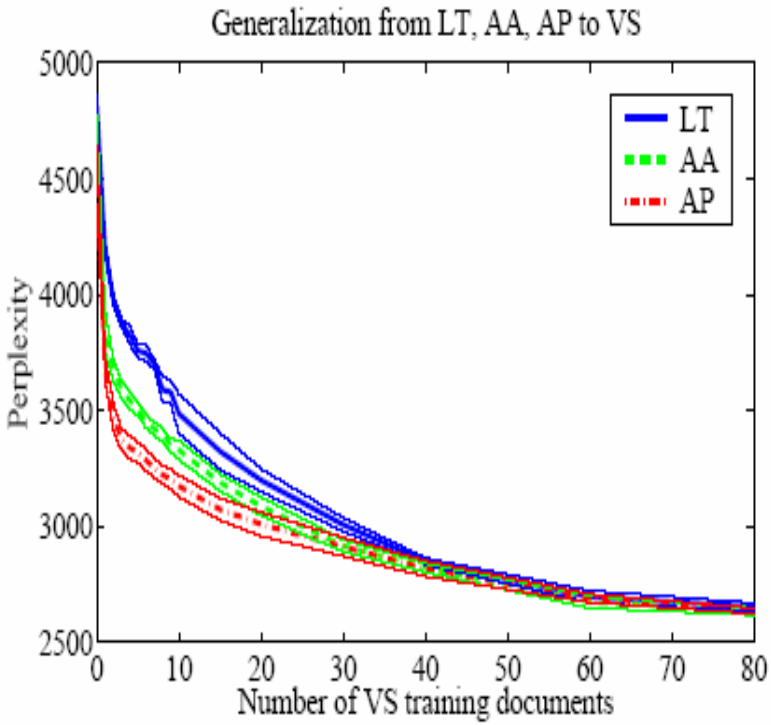
- Articles from the conference are divided into sections
- HDP is used to discover the shared topics among the articles within each section
- Want to exam relationships among the sections

Experiment – Multiple Corpora



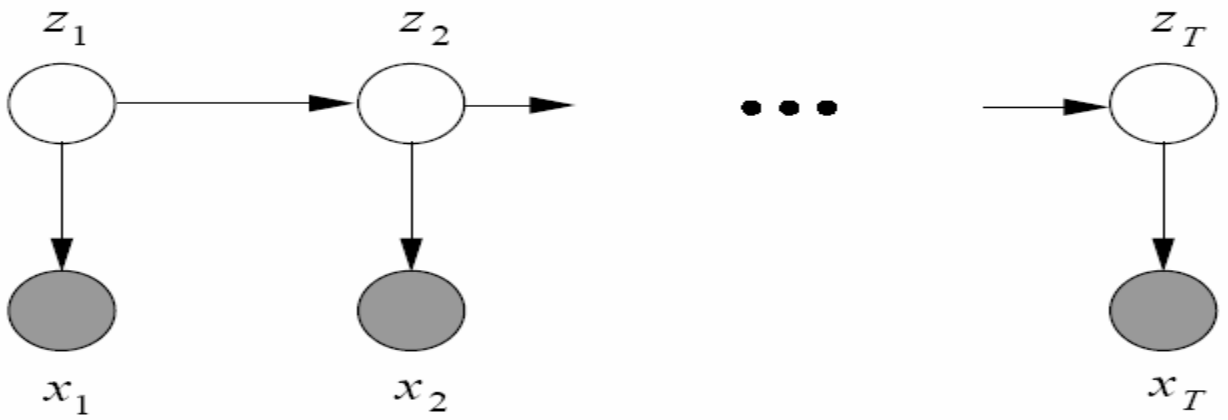
Experiment – Multiple Corpora





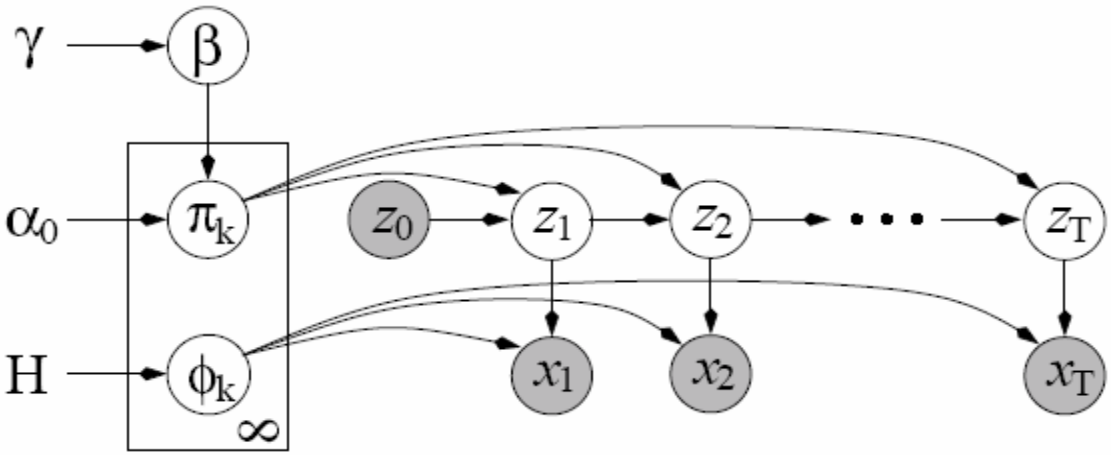
Hidden Markov Models

- HMM is a dynamic variant of a mixture model : each row of the transition matrix is a set of mixing proportions for the choice of the next state



HDP-HMM

- An HMM can be viewed as a set of mixture models : one mixture model for each value of the current state
- When a new state arises, HDP shares this new state among of the current states



Experiments- Alice in Wonderland