Nested Hierarchical Dirichlet Processes

John Paisley, Chong Wang, David M. Blei, and Michael I. Jordan

Review by David Carlson

Overview

- Dirichlet process (DP)
- Nested Chinese restaurant process topic model (nCRP)
- Hierarchical Dirichlet process topic model (HDP)
- Nested Hierarchical Dirichlet process topic model (nHDP)
- Outline of stochastic variational Bayesian procedure
- Results

Dirichlet Process

In general, we can write a that a distribution G drawn from a Dirichlet process can be written as:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha G_0) \quad (1) \\ G &= \sum_{i=1}^{\infty} p_i \delta_{\theta_i} \quad (2) \end{aligned}$$

where p_i is a probability and each θ_i is an atom. We can construct a Dirichlet process mixture model over data $W_1, ..., W_N$:

$$\begin{aligned} W_n | \varphi_n &\sim F_W(\varphi_n) \quad (3) \\ \varphi_n | G &\sim G \quad (4) \end{aligned}$$

Generating the Dirichlet process

There are two common methods for generating the Dirichlet process. The first is the *Chinese restaurant process*, where we integrate out G to get the a distribution for φ_{n+1} given the previous values as:

$$\varphi_{n+1}|\varphi_1, ..., \varphi_n \sim \frac{\alpha}{\alpha + n} G_0 + \sum_{i=1}^{n} \frac{1}{\alpha + n} \delta_{\varphi_i} \tag{5}$$

The second commonly used method is a *stick-breaking construction*. in this case, one can construct G as:

$$G = \sum_{i=1}^{\infty} V_i \left[\prod_{j=1}^{i-1} (1 - V_j) \right] \delta_{\theta_i}, \quad V_i \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha), \quad \theta_i \overset{\text{iid}}{\sim} G_0 \tag{6}$$

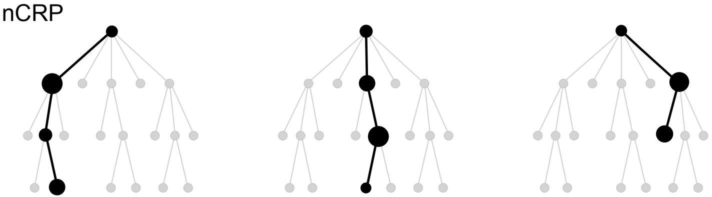
Because the stick-breaking construction maintains the independence among $\varphi_1, ..., \varphi_N$ is has advantages over the CRP during mean-field variational inference.

Nested Chinese restaurant processes

The CRP (or DP) is a flat model. Often, it is of interest to organize the topics (or atoms) hierarchically to have subcategories of larger categories in a tree-structure. One way to construct such a hierarchical data structure is through the *nested Chinese restaurant process* (nCRP).

Nested Chinese restaurant processes

As an analogy, consider an extension of the CRP analogy. Each customer selects a table (parameter) according to the CRP. From that table, the customer chooses a restaurant accessible only from the table, where he/she chooses a table from that restaurant specific CRP.



As shown in the image, each customer (document) that draws from the CRP chooses a single path down the tree.

Modeling the nCRP

Let $\boldsymbol{i}_l = (i_1, ..., i_l)$ be a path to a node at level l of the tree. Then we can define the DP at the end of this path as:

$$G_{\boldsymbol{i}_l} = \sum_{j=1}^{\infty} V_{(\boldsymbol{i}_l,j)} \prod_{m=1}^{j-1} (1 - V_{(\boldsymbol{i}_l,m)}) \delta_{\theta_{(\boldsymbol{i}_l,j)}} \tag{7}$$

If the next node is child j, then the nCRP transitions to the DP $G_{\boldsymbol{i}_{l+1}}$, where we define $\boldsymbol{i}_{l+1} = (\boldsymbol{i}_l, j)$

Nested CRP topic models

We can use the nCRP to define a path down a shared tree, but we want to use this tree to model the data. One application of the tree-structure is a topic model, where we would define each atom $\theta_{\boldsymbol{i}_l,j}$ defines a topic.

$$\theta_{\boldsymbol{i}_l,j} \sim \text{Dir}(\boldsymbol{\eta}) \tag{8}$$

Each document in the nCRP would choose one path down the tree according to a Markov process, and the path provides a sequence of topics $\varphi_d = (\varphi_{d,1}, \varphi_{d,2}, \ldots)$ which we can use to generate the words in the document. The distribution over these topics is provided by a new document-specific stick-breaking process:

$$G^{(d)} = \sum_{j=1}^{\infty} U_{d,j} \prod_{m=1}^{j-1} (1 - U_{d,m}) \delta_{\varphi_{d,j}}, \quad U_{d,j} \overset{\text{iid}}{\sim} \text{Beta}(\gamma_1, \gamma_2) \tag{9}$$

Problems with nCRP

There are several problems with the nCRP, including:

- Each document is only allowed to follow one path down the tree, limiting the number of topics for each document to the number of levels (typically ≤ 4), which can force topics to blend (have less specificity)
- Topics are often repeated on many different parts of the tree if they appear as random effects in documents
- The tree is shared, but very few topics are shared between a set of documents because they each follow independent single paths down the tree

We would like to be able to learn a distribution over the entire shared tree for each document to give a more flexible modeling structure. The solution given to this problem is the *nested hierarchical Dirichlet process.*

Hierarchical Dirichlet processes

The HDP is a multi-level version of the Dirichlet process. This is described as the hierarchical process:

$$G_d | G \sim \mathrm{DP}(\beta G), \qquad G \sim \mathrm{DP}(\alpha G_0) \tag{10}$$

In this case, we have that each document has it's own DP (G_d) which is drawn from a shared DP G. In this way, the weights on each topic (atom) are allowed to vary smoothly from document to document, but still share statistical strength.

This can be represented as a stick-breaking process as well:

$$G_d = \sum_{i=1}^{\infty} V_i^d \prod_{j=1}^{i-1} (1 - V_j^d) \delta_{\phi_i}, \qquad V_i^d \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \beta), \qquad \phi_i \overset{\text{iid}}{\sim} G \tag{11}$$

Nested Hierarchical Dirichlet Processes

The nHDP formulation allows (*i*) each word to follow its own path to a topic, and (ii) each topic its own distribution over a shared tree.

To formulate the nHDP, let a tree $\mathcal{T}$ be a draw from the global nCRP with stick-breaking construction. Instead of drawing a path for each document, we use each Dirichlet process in $\mathcal{T}$ as a base for a second level DP drawn independently for each document. In order words, each document d has tree $\mathcal{T}_d$, where for each $G_{\boldsymbol{i}_l} \in \mathcal{T}$ we draw:

$$G_{\boldsymbol{i}_l}^{(d)} \sim \mathrm{DP}(\beta G_{\boldsymbol{i}_l}) \tag{12}$$

Nested Hierarchical Dirichlet Processes

We can write the second level DP as:

$$G^{(d)}_{\boldsymbol{i}_l} = \sum_{j=1}^{\infty} V^{(d)}_{\boldsymbol{i}_l,j} \prod_{m=1}^{j-1} (1 - V^{(d)}_{\boldsymbol{i}_l,m}) \delta_{\phi^{(d)}_{\boldsymbol{i}_l,j}}, \;\; V^{(d)}_{(\boldsymbol{i}_i,j} \overset{\text{iid}}{\sim} \text{Beta}(1, \beta), \;\; \phi^{(d)}_{\boldsymbol{i},j} \overset{\text{iid}}{\sim} G_{\boldsymbol{i}_l} \quad (13)$$

However, we would like to maintain the same tree structure in $\mathcal{T}_d$ as in $\mathcal{T}$. To do this, we can map the probabilities, so that the probability of being on node $\theta_{(\boldsymbol{i}_l,j)}$ in document *d* is:

$$G^{(d)}_{\boldsymbol{i}_l}(\{\theta_{(\boldsymbol{i}_l,j)}\}) = \sum_m G^{(d)}_{\boldsymbol{i}_l}(\{\phi^{(d)}_{\boldsymbol{i}_l,m}\}) \mathbb{I}(\phi^{(d)}_{\boldsymbol{i}_l,m} = \theta_{(\boldsymbol{i}_l,j)}) \quad (14)$$

Generating a Document

After generating the tree $\mathcal{T}_d$ for document d, we draw document-specific beta random variables that act as a stochastic switch. I.E. if a word is at node $\boldsymbol{i}_l$, it determines the probability that the word uses the topic at that node or continues down the tree. So we stop at node $\boldsymbol{i}_l$ with probability:

$$U_{d,\boldsymbol{i}_l} \overset{\text{idd}}{\sim} \text{Beta}(\gamma_1, \gamma_2) \tag{15}$$

From the stick-breaking construction, the probability that the topic $\varphi_{d,n} = \theta_{\boldsymbol{i}_l}$ for word $W_{d,n}$ is:

$$Pr(\varphi_{d,n} = \theta_{\boldsymbol{i}_l} | \mathcal{T}_d, U_d) = \left[\prod_{\boldsymbol{i}_m \subset \boldsymbol{i}_l} G^{(d)}_{\boldsymbol{i}_m}(\{\theta_{\boldsymbol{i}_{m+1}}\}) \right] \left[U_{d,\boldsymbol{i}_l} \prod_{m=1}^{l-1} (1 - U_{d,\boldsymbol{i}_m}) \right] \tag{16}$$

Generative Procedure

Algorithm 1 Generating Documents with the Nested Hierarchical Dirichlet Process

```
Step 1. Generate a global tree $\mathcal{T}$ by constructing an nCRP as in Section II-B1.
Step 2. Generate document tree $\mathcal{T}_d$ and switching probabilities $\boldsymbol{U}^{(d)}$. For document $d$,
    a) For each DP in $\mathcal{T}$, draw a second-level DP with this base distribution (Equation 8).
    b) For each node in $\mathcal{T}_d$ (equivalently $\mathcal{T}$), draw a beta random variable (Equation 10).
Step 3. Generate the documents. For word $n$ in document $d$,
    a) Sample atom $\varphi_{n,d} = \theta_{\boldsymbol{i}_l}$ with probability given in Equation (11).
    b) Sample $W_{n,d}$ from the discrete distribution with parameter $\varphi_{d,n}$.
```

Inference

In a large amount of data, it is difficult to use an MCMC algorithm to efficient learn the parameters in the model. To solve this problem, the authors developed a *stochastic variational Bayesian* inference scheme that updates over a sub-batch of documents denoted by C_s

```
for $s = 1, ..., \infty$ do
    for $d \in C_s$ do
        Update all local parameters for document $d$:
        $(z_{\mathbf{i},j}^{(d)},\ c_{d,n},\ V_{\mathbf{i},j}^{(d)},\ U_{d,\mathbf{i}})$ while holding global variables constant
    end
    Stochastic updates for corpus variables:
    Find a noisy estimate for the Dirichlet parameters $\lambda'_{\mathbf{i}}$ of $q(\theta_{\mathbf{i}})$ , and
    then update the global parameters $\lambda_{\mathbf{i},w}^{s+1} = \lambda_0 + (1-\rho_s)\lambda_{\mathbf{i},w}^{s} + \rho_s \lambda'_{\mathbf{i},w}$
    Likewise update the parameters for $q(V_{\mathbf{i}_l,j})$
end
```

Notes on inference

Initialization: A good initialization greatly benefits the stochastic VB algorithm. For a small set of documents, the authors iteratively use k-means through a hierarchical k-means clustering to define an initial tree, with n_1 clusters at the top level, n_2 clusters at the next level, and n_3 clusters at the last level.

In the small experiments a truncated tree with widths of (10,7,5) was used in the inference results to give 430 possible nodes whereas in the ”big data” experiments the tree was truncated to (20,10,5).

To test the hold-out set the authors completely held out a set of documents, and then learned their local parameters on 75% of the held-out words and tested on the remaining 25%. Predictive log-likelihood values are reported.

Results

TABLE II

Comparison of the nHDP with the nCRP on three smaller problems.

Method\Data set	JACM	Psych. Review	PNAS
Variational nHDP	-5.405 ± 0.012	-5.674 ± 0.019	-6.304 ± 0.003
Variational nCRP	-5.433 ± 0.010	-5.843 ± 0.015	-6.574 ± 0.005
Gibbs nCRP	-5.392 ± 0.005	-5.783 ± 0.015	-6.496 ± 0.007

Results

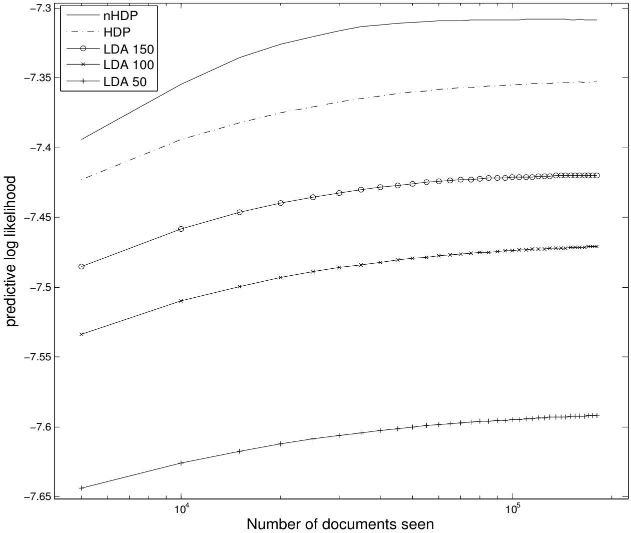
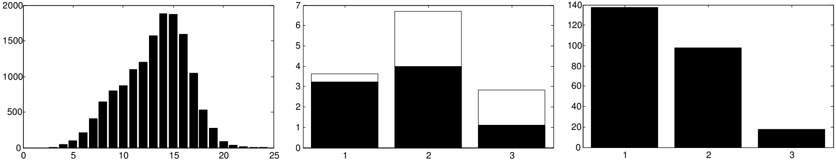


Fig. 2. The New York Times: Average per-word log likelihood on a held-out test set as a function of training documents seen.

Results



(a) Size of subtree

(b) Avg number nodes/level
& Avg number used

(c) Avg number of words/level

Fig. 3. The New York Times: Per-document statistics from the test set using the tree at the final step of the algorithm. (a) A histogram of the size of the subtree selected for a document. (b) The average number of nodes by level within the subtree (white), and the average number by level that have at least one expected observation (black). (c) The average number of words allocated to each level of the tree per document.

Results

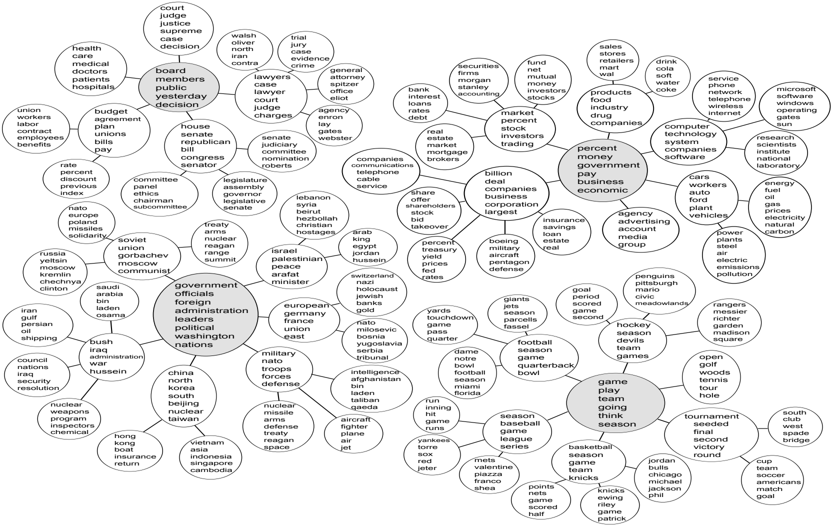

court judge justice supreme case decision
health care medical doctors patients hospitals
board members public yesterday decision
walsh oliver north iran contra
trial jury case evidence crime
lawyers case lawyer court judge charges
general attorney spitzer office eliot
agency enron lay gates webster
union workers labor contract employees benefits
budget agreement plan unions bills pay
house senate republican bill congress senator
senate judiciary committee nomination roberts
rate percent discount previous index
committee panel ethics chairman subcommittee
legislature assembly governor legislative senate
lebanon syria beirut hezbollah christian hostages
nato europe poland missiles solidarity
treaty arms nuclear reagan range summit
soviet union gorbachev moscow communist
arab king egypt jordan hussein
israel palestinian peace arafat minister
russia yeltsin moscow kremlin chechnya clinton
saudi arabia bin laden osama
government officials foreign administration leaders political washington nations
switzerland nazi holocaust jewish banks gold
european germany france union east
iran gulf persian oil shipping
nato milosevic bosnia yugoslavia serbia tribunal
bush iraq administration war hussein
council nations iraq security resolution
military nato troops forces defense
intelligence afghanistan bin laden taliban qaeda
china north korea south beijing nuclear taiwan
nuclear weapons program inspectors chemical
nuclear missile arms defense treaty reagan space
aircraft fighter plane air jet
hong kong boat insurance return
vietnam asia indonesia singapore cambodia
securities firms morgan stanley accounting
fund net mutual money investors stocks
sales stores retailers mart wal
drink cola soft water coke
bank interest loans rates debt
products food industry drug companies
service phone network telephone wireless internet
microsoft software windows operating gates sun
market percent stock investors trading
real estate market mortgage brokers
computer technology system companies software
research scientists institute national laboratory
percent money government pay business economic
companies communications telephone cable service
billion deal companies business corporation largest
cars workers auto ford plant vehicles
energy fuel oil gas prices electricity natural carbon
share offer shareholders stock bid takeover
insurance savings loan estate real
agency advertising account media group
power plants steel air electric emissions pollution
percent treasury yield prices fed rates
boeing military aircraft pentagon defense
penguins pittsburgh mario civic meadowlands
goal period scored game second
giants jets season parcells fassel
rangers messier richter garden madison square
yards touchdown game pass quarter
hockey season devils team games
football season game quarterback bowl
dame notre bowl football season miami florida
open golf woods tennis tour hole
game play team going think season
run inning hit game runs
season baseball game league series
tournament seeded final second victory round
south club west spade bridge
yankees torre sox red jeter
mets valentine piazza franco shea
basketball season game team knicks
jordan bulls chicago michael jackson phil
cup team soccer americans match goal
points nets game scored half
knicks ewing riley game patrick

Fig. 4. Tree-structured topics from The New York Times. The shaded node is the top-level node and lines indicate dependencies within the tree. In general, topics are learning in increasing levels of specificity. For clarity, we have removed grammatical variations of the same word, such as “scientist” and “scientists.”

Results

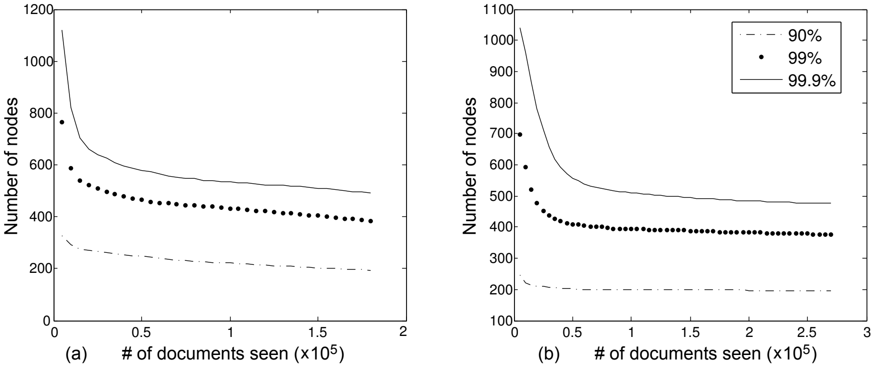


Fig. 5. Tree size: The smallest number of nodes containing 90%, 99% and 99.9% of all paths as a function of documents seen for (a) The New York Times, and (b) Wikipedia.

Results

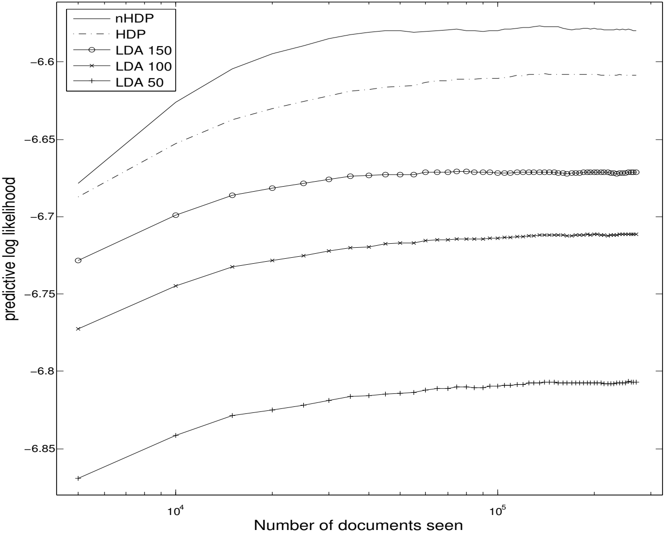
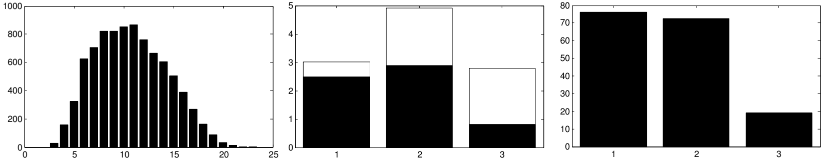


Fig. 6. Wikipedia: Average per-word log likelihood on a held-out test set as a function of training documents seen.

Results



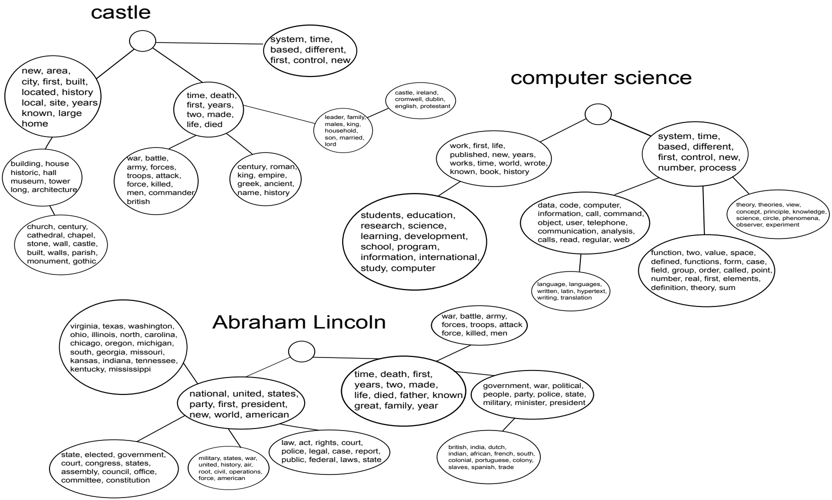
(a) Size of subtree

(b) Avg number nodes/level & Avg number used

(c) Avg number of words/level

Fig. 7. Wikipedia: Per-document statistics from the test set using the tree at the final step of the algorithm. (a) A histogram of the size of the subtree selected for a document. (b) The average number of nodes by level within the subtree (white), and the average number by level that have at least one expected observation (black). (c) The average number of words allocated to each level of the tree per document.

Results



castle

system, time, based, different, first, control, new

new, area, city, first, built, located, history local, site, years known, large home

time, death, first, years, two, made, life, died

leader, family, males, king, household, son, married, lord

castle, ireland, cromwell, dublin, english, protestant

building, house historic, hall museum, tower long, architecture

church, century, cathedral, chapel, stone, wall, castle, built, walls, parish, monument, gothic

war, battle, army, forces, troops, attack, force, killed, men, commander british

century, roman, king, empire, greek, ancient, name, history

computer science

work, first, life, published, new, years, works, time, world, wrote, known, book, history

system, time, based, different, first, control, new, number, process

students, education, research, science, learning, development, school, program, information, international, study, computer

data, code, computer, information, call, command, object, user, telephone, communication, analysis, calls, read, regular, web

theory, theories, view, concept, principle, knowledge, science, circle, phenomena, observer, experiment

language, languages, written, latin, hypertext, writing, translation

function, two, value, space, defined, functions, form, case, field, group, order, called, point, number, real, first, elements, definition, theory, sum

Abraham Lincoln

virginia, texas, washington, ohio, illinois, north, carolina, chicago, oregon, michigan, south, georgia, missouri, kansas, indiana, tennessee, kentucky, mississippi

war, battle, army, forces, troops, attack force, killed, men

national, united, states, party, first, president, new, world, american

time, death, first, years, two, made, life, died, father, known great, family, year

government, war, political, people, party, police, state, military, minister, president

state, elected, government, court, congress, states, assembly, council, office, committee, constitution

military, states, war, united, history, air, root, civil, operations, force, american

law, act, rights, court, police, legal, case, report, public, federal, laws, state

british, india, dutch, indian, african, french, south, colonial, portuguese, colony, slaves, spanish, trade

Fig. 8. Examples of subtrees for three articles from *Wikipedia*. The three sizes of font indicate differentiate the more probable topics from the less probable.

Conclusions

The nHDP provides a way to eliminate some of the constraints of the nCRP and provide a more informative tree that gives a higher predictive log-likelihood.

Using the complete tree is a method explored in the paper ”Tree-Structured Stick Breaking for Hierarchical Data” by Adams et. al, but this method allows documents to share statistical strength for preferences on the tree structure between documents.

The stochastic variational Bayesian algorithm allows for efficient inference on this complicated model that seems to perform well.