Recommendation Engines; Collaborative Filtering; Thematic clustering of large text corpora; Infinite Markov model for statistical NLP

Russell W. Hanson

Dec. 8, 2008

Outline

- Several problems in applied mathematics and approaches to their solutions:
  - Recommendation Engines
  - Collaborative Filtering
  - Thematic clustering of large text corpora
  - Infinite Markov model for statistical NLP

LobeLink.com – social bookmarking; social web annotation; and recommendation engine

Recommendation Engines, $$$

Amazon.com

What Do Customers Ultimately Buy After Viewing This Item?

72% buy the item featured on this page:
Agent-Based Models (Quantitative Applications in the Social Sciences) ★★★★½ (2)
$15.25

12% buy
Complex Adaptive Systems: An Introduction to Computational Models of Social Life (Princeton Studies in Complexity) ★★★★½ (7)
$23.35

7% buy
Generative Social Science: Studies in Agent-Based Computational Modeling (Princeton Studies in Complexity) ★★★★★ (5)
$42.00

5% buy
Evolutionary Dynamics: Exploring the Equations of Life ★★★★★ (8)
$31.96

Compare these items **Explore similar items**

NetFlix.com

THE SCREENS ISSUE

If You Liked This, You're Sure to Love That

By CLIVE THOMPSON
Published: November 21, 2008

THE "NAPOLEON DYNAMITE" problem is driving Len Bertoni crazy. Bertoni is a 51-year-old "semiretired" computer scientist who lives an hour outside Pittsburgh. In the spring of 2007, his sister-in-law e-mailed him an intriguing bit of news: Netflix, the Web-based DVD-rental company, was holding a contest to try to improve Cinematch, its "recommendation engine." The prize: $1 million.

A Recommendation Engine

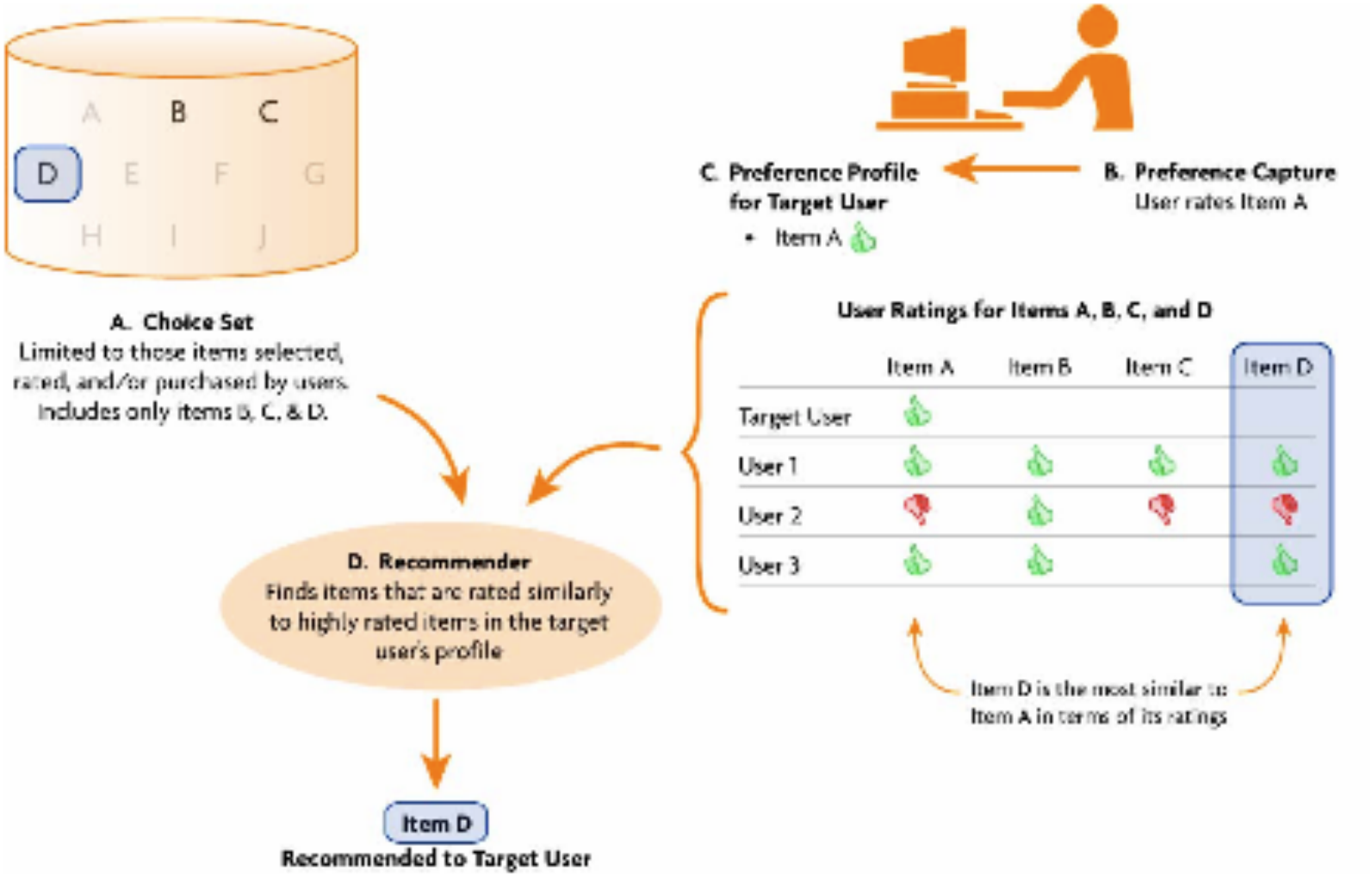


Figure 3. Item-based Collaborative Filtering

Attributized Bayesian Choice Modeling

Attributized content items, i, are stored as vectors in the choice-set database such that:

$\mathbf{A_i} = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iN})$,
where $\alpha_{i1}, \ldots \alpha_{iN}$ are the scores on the N attributes for item i.

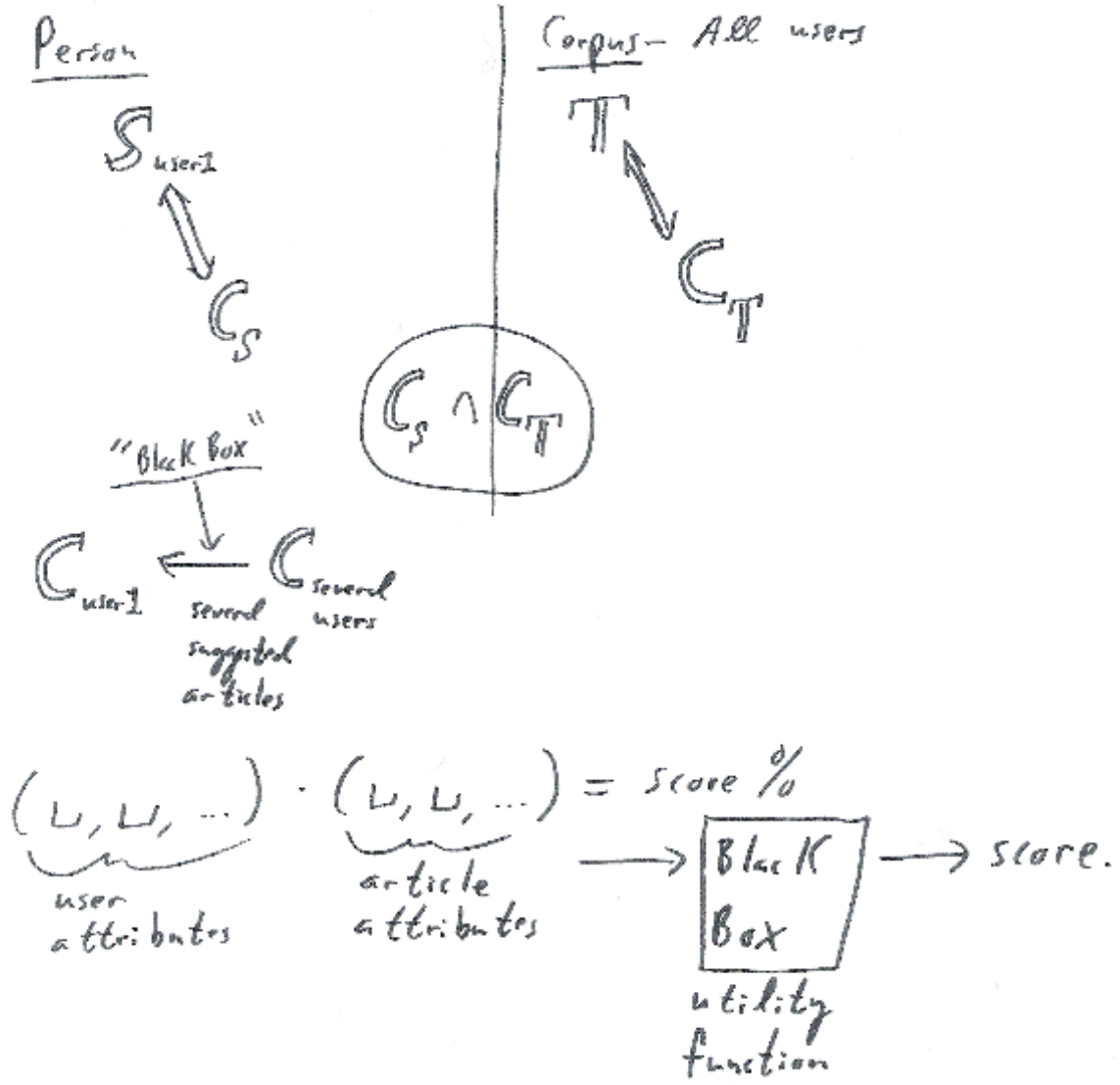
Summary of tastes, T:

$\mathbf{T_u} = (\beta_{u1}, \beta_{u2}, \ldots, \beta_{uN})$,
where $\beta_{u1}, \ldots, \beta_{uN}$ are user u's weights on N attributes.

- Collaborative Filtering for text and "news":
  - Cold Start Problem (it isn't collaborative until it's collaborative)
  - Past Experience: Some people want the most popular ("Dodgers make offer to Manny Ramirez - Boston.com"); some don't ("Non-Abelian Anyons and Topological Quantum Computation")
  - By weight in whole network; by weight in user's network; by weight in thematic cluster

Thematic Clustering

- Want to have more fine-grained recommendations than connectivity in user network — weight in a given thematic cluster.



Document Universe. Let Ω be the *document universe*, i. e. the set of all documents known to the IR system:

$$\Omega = \{D_i \mid i \in \mathbf{N}\}. \tag{1.1}$$

$\mathbf{N}$ is the set of natural numbers.

Document Set. Let $\mathcal{S}$ be the *document set*, i. e. those n documents that form the input to the clustering process:

$$\mathcal{S} = \{D_1, D_2, \ldots, D_n\} \subseteq \Omega. \tag{1.2}$$

Note: For many—but not all—applications $\mathcal{S}$ equals Ω.

Feature Set. Let

$$\mathcal{F} = \{f_1, f_2, \ldots, f_m\}, \tag{1.3}$$

with $\mathcal{F}$ a *set of m features* and f_i an individual feature i. Each feature stands for a concrete or abstract document property.

Document Vector. Let

$$\mathbf{d}_i = (d_{i1}, d_{i2}, \ldots, d_{im}), \tag{1.4}$$

with $\mathbf{d}_i$ the *document vector* of document D_i in an m-dimensional feature space $\mathcal{F}$. The jth component of $\mathbf{d}_i$ (written as d_{ij}) corresponds to the *value* or *strength* of feature f_j in document D_i. d_{ij} is usually a non-negative real number:

$$d_{ij} \in \mathbf{R}_0^+. \tag{1.5}$$

Document Feature Matrix. Let

$$H = \begin{pmatrix} \mathbf{d_1} \\ \mathbf{d_2} \\ \vdots \\ \mathbf{d_n} \end{pmatrix}, \tag{1.6}$$

with H the *document feature matrix*, defined by the individual vector representations $\mathbf{d}$ of all $D \in \mathcal{S}$. The document feature matrix is usually the input to the clustering algorithm.

Document Vectorisation Function. Let τ be a function which transforms a text document into an m-dimensional vector representation in feature space $\mathcal{F}$:

$$\mathbf{d}_i = \tau(D_i), \text{ with } \tau : \Omega \to \mathbf{R}^m, \tag{1.7}$$

with $\mathbf{R}$ the set of real numbers.

Feature Transformation Function. Let ϕ be a function which transforms a document vector from one feature space ($\mathcal{F}_1$) into another ($\mathcal{F}_2$), sometimes making use of additional information from the document feature matrix H:

$$\mathbf{d}'_i = \phi(\mathbf{d}_i, H), \text{ with } \phi : \mathbf{R}^{m_{\mathcal{F}_1}}, \mathbf{R}^{n \times m} \to \mathbf{R}^{m_{\mathcal{F}_2}}. \tag{1.8}$$

Document Frequency. Let

$$\mathrm{df}(j, H) = \sum_i |\mathrm{sgn}(h_{ij})|, \tag{1.9}$$

with the *document frequency* $\mathrm{df}(j, H)$ the number of documents with a non-zero value for feature f_j.

Cluster. Let a *cluster* C_i be an subset of $\mathcal{S}$:

$$C_i \subseteq \mathcal{S}, \tag{1.10}$$

and let n_i be the number of objects in cluster C_i:

$$n_i = |C_i|. \tag{1.11}$$

Cluster Solution. Let

$$\mathcal{C} = \{C_1, C_2, \ldots, C_k \mid C_i \subseteq \mathcal{S} \;\; \forall i \in 1 \ldots k\}. \tag{1.12}$$

A *cluster solution* $\mathcal{C}$ is thus defined as a set of k clusters.

Cluster Algorithm. Let

$$\mathcal{C} = \kappa(H), \quad \text{with} \quad \kappa : \mathbf{R}^{n \times m} \rightarrow \mathcal{P}(\mathcal{S}) \tag{1.13}$$

and with κ denoting the cluster algorithm, $\mathcal{P}(\mathcal{S})$ the power set of $\mathcal{S}$ and $\mathbf{R}$ the set of real numbers.

Cluster Representative. Let

$$\mathbf{r}_i = (r_{i1}, r_{i2}, \ldots, r_{im}), \tag{1.14}$$

with $\mathbf{r}_i$ a *representative vector* for cluster C_i in an m-dimensional feature space $\mathcal{F}$.

Individual Cluster Criterion Function. An individual *criterion function* $E(C)$ measures the quality of a single cluster:

$$E : \mathcal{P}(\mathcal{S}) \rightarrow \mathbf{R}, \tag{1.15}$$

with $\mathcal{P}(X)$ the power set of X and $\mathbf{R}$ the set of real numbers.

Overall Cluster Criterion Function. An overall *criterion function* $\Psi(\mathcal{C})$ measures the quality of an entire cluster solution:

$$\Psi : \mathcal{P}(\mathcal{P}(\mathcal{S})) \rightarrow \mathbf{R}, \tag{1.16}$$

with $\mathcal{P}(\mathcal{P}(\mathcal{S}))$ the set of all possible cluster solutions.

Type and Token. Within documents it is common to refer to word *types* and word *tokens*. The former refer abstractly to features in a document or a corpus, while the latter refer to individual occurrences. Formally speaking, the tokens of a document are a *bag* (which allows multiple occurrences of the same element). The types are the *set* created by eliminating all duplicates from the token bag.

Recall and Precision. In IR two widespread performance measures are defined by the set of documents in a collection that are *relevant* to a particular query ($\mathcal{A}$) and those documents that are actually *retrieved* by the system ($\mathcal{B}$):

$$\textit{Recall}\ (R) \quad = \quad \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A}|}, \tag{1.17}$$

$$\textit{Precision}\ (P) \quad = \quad \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{B}|}. \tag{1.18}$$

Their weighted arithmetic mean, the so-called F-Measure is also used frequently (see Equation 2.77 for an example).

Latent Dirichlet Allocation/Analysis

We describe *latent Dirichlet allocation* (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.[1]

LDA assumes the following generative process for each document $\mathbf{w}$ in a corpus D:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Several simplifying assumptions are made in this basic model, some of which we remove in subsequent sections. First, the dimensionality k of the Dirichlet distribution (and thus the dimensionality of the topic variable z) is assumed known and fixed. Second, the word probabilities are parameterized by a $k \times V$ matrix β where $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$, which for now we treat as a fixed quantity that is to be estimated. Finally, the Poisson assumption is not critical to anything that follows and more realistic document length distributions can be used as needed. Furthermore, note that N is independent of all the other data generating variables (θ and $\mathbf{z}$). It is thus an ancillary variable and we will generally ignore its randomness in the subsequent development.

Latent Dirichlet Allocation/Analysis (p3)

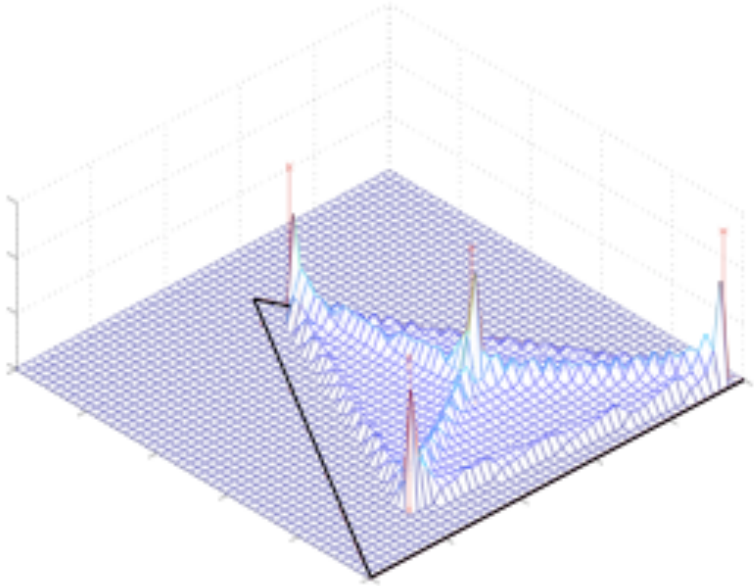


Figure 2: An example density on unigram distributions $p(w|\theta,\beta)$ under LDA for three words and four topics. The triangle embedded in the x-y plane is the 2-D simplex representing all possible multinomial distributions over three words. Each of the vertices of the triangle corresponds to a deterministic distribution that assigns probability one to one of the words; the midpoint of an edge gives probability 0.5 to two of the words; and the centroid of the triangle is the uniform distribution over all three words. The four points marked with an x are the locations of the multinomial distributions $p(w|z)$ for each of the four topics, and the surface shown on top of the simplex is an example of a density over the $(V-1)$-simplex (multinomial distributions of words) given by LDA.

Latent Dirichlet Allocation/Analysis (p2)

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Infinite Markov Models

Language models and parsers
N-gram (bigram, trigram) vs. ∞-gram
The supercalifragilisticexpialidocious-problem

Figure 2: Probabilistic suffix tree of an infinite depth. $(1-q_i)$ is a "penetration probability" of a descending customer at each node i, defining a stick-breaking process over the infinite tree.

Consequently, in the hierarchical Pitman-Yor language model (HPYLM), the predictive probability of a symbol $s = s_t$ in context $h = s_{t-n} \cdots s_{t-1}$ is recursively computed by

$$p(s|h) = \frac{c(s|h) - d \cdot t_{hs}}{\theta + c(h)} + \frac{\theta + d \cdot t_{h\cdot}}{\theta + c(h)} p(s|h'), \quad (1)$$

where $h' = s_{t-n+1} \cdots s_{t-1}$ is a shortened context with the farthest symbol dropped. $c(s|h)$ is the count of s at node h, and $c(h) = \sum_s c(s|h)$ is the total count at node h. t_{hs} is the number of times symbol s is estimated to be generated from its parent distribution $p(s|h')$ rather than $p(s|h)$ in the training data: $t_{h\cdot} = \sum_s t_{hs}$ is its total. θ and d are the parameters of the Pitman-Yor process, and can be estimated through the distribution of customers on a suffix tree by Gamma and Beta posterior distributions, respectively. For details, see [9].

n	HPYLM	VPYLM	Nodes(H)	Nodes(V)
3	113.60	113.74	1,417K	**1,344K**
5	101.08	101.69	12,699K	**7,466K**
7	N/A	**100.68**	*27,193K*	**10,182K**
8	N/A	**100.58**	*34,459K*	**10,434K**
∞	—	**100.36**	—	**10,629K**

Table 2: Perplexity Results of VPYLM and HPYLM on the NAB corpus with the number of nodes in each model. N/A means a memory overflow caused by the expected number of nodes shown in italic.

hierarchical Pitman-Yor language model (HPYLM)

variable order hier archical Pitman-Yor language model (VPYLM)

Selected References

Document Clustering in Large German Corpora Using Natural Language Processing
Richard Forster (2006)
University of Zurich

Latent Dirichlet Allocation
Blei, Ng, and Jordan
Journal of Machine Learning Research 3 (2003) 993-1022

The Infinite Markov Model
Daichi Mochihashi, Eiichiro Sumita
NIPS, 2007

LobeLink.com