Lexical and Hierarchical Topic Regression

Viet-An Nguyen[1], Jordan Boyd-Graber[1,2,4] and Philip Resnik[1,3,4]

[1]Department of Computer Science, [2]iSchool, [3]Department of Linguistics, [4]UMIACS
University of Maryland, College Park

Overview

Inspired by a two-level theory from political science that unifies:

- Agenda setting: which **issues** are salient
- Ideological framing: which **aspects** of the discussed issues are salient

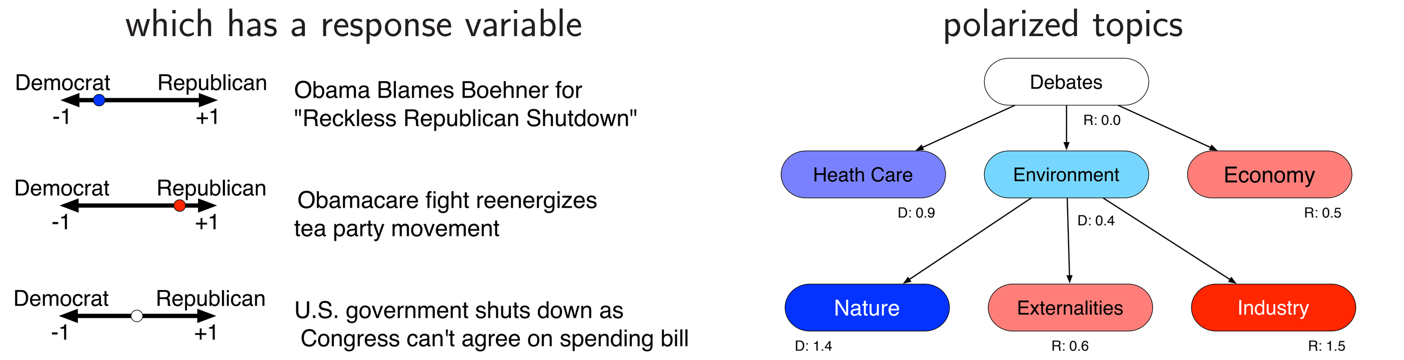
we propose **supervised hierarchical latent Dirichlet allocation** (SHLDA), which jointly captures documents' multi-level topic structure and their polar response variables.

SHLDA's key modeling contributions:

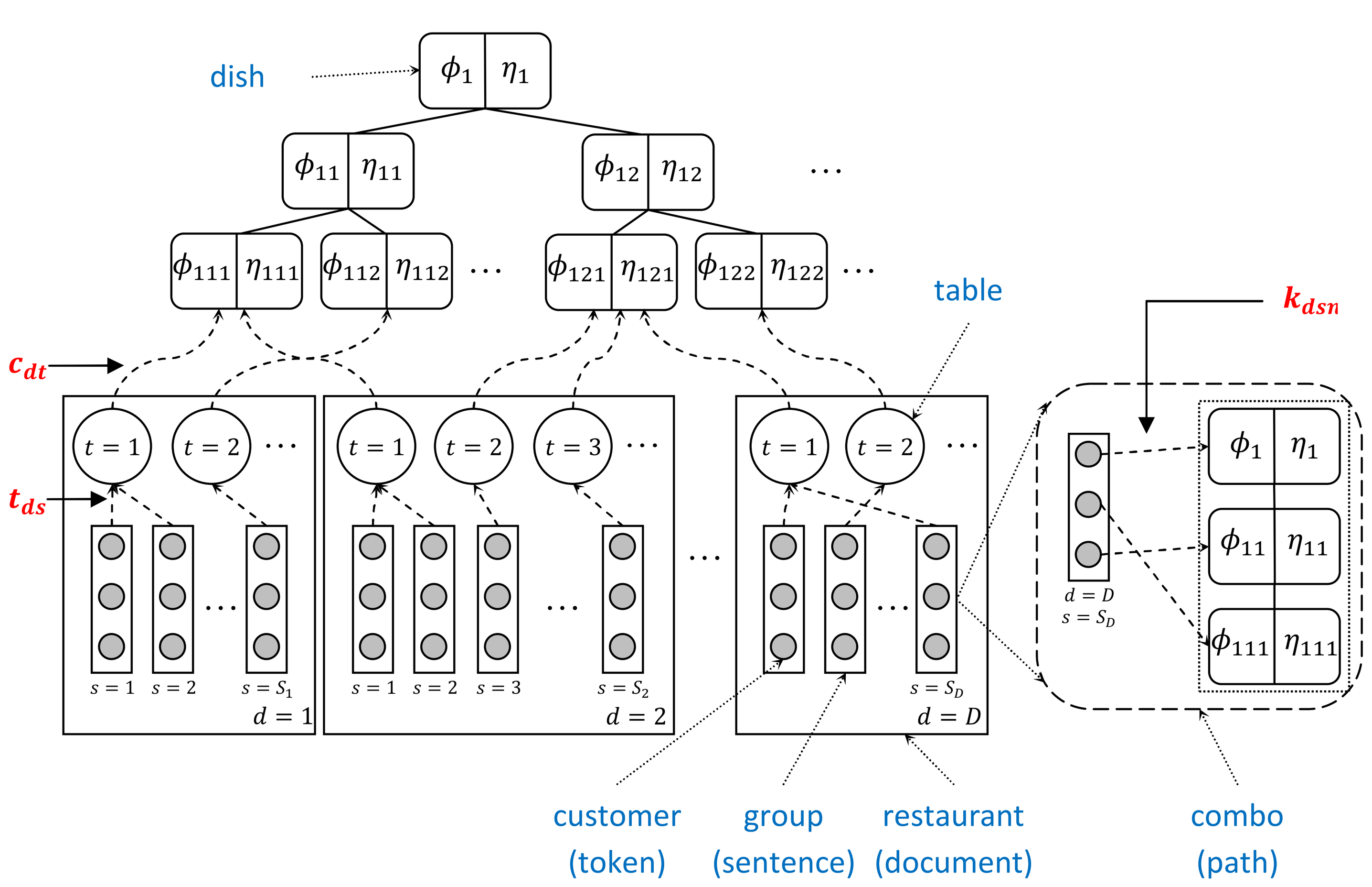
- SHLDA relaxes HLDA's restriction on one-path-per-document by assigning each sentence to a path.
- The response variables are modeled using both hierarchical topic and lexical regressions.

Input: A collection of documents, each of which has a response variable

Output: A tree-structured hierarchy of polarized topics



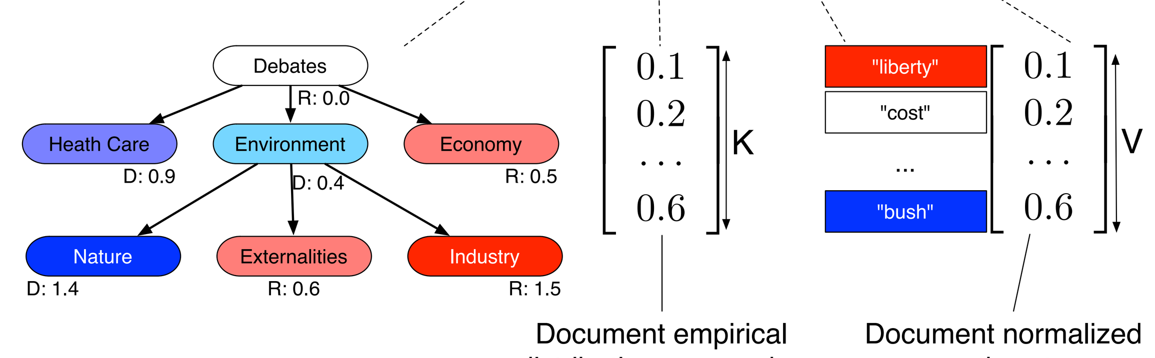
Hierarchical topic structure



- Each document is a bag of exchangeable sentences.
- Each sentence is a bag of exchangeable tokens.
- Sentences in a document are clustered together using per-document CRPs.
- Each CRP's table is assigned to a tree path using nested CRP prior.
- Given the path assigned to a sentence, tokens are assigned to a node using per-document truncated stick breaking process.

Combining lexical and hierarchical topic regression

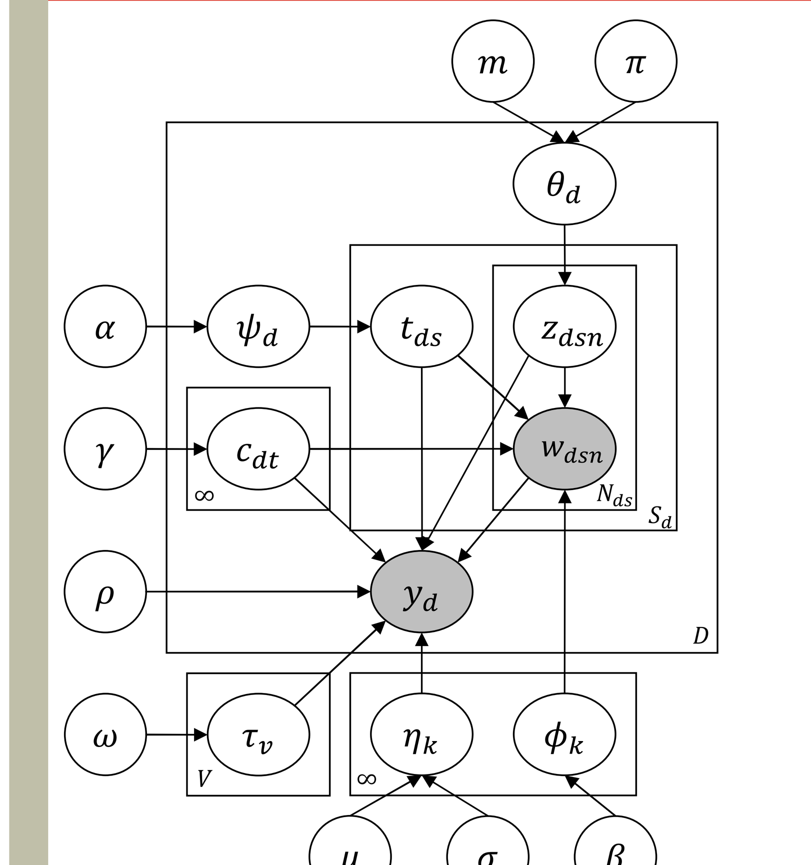
$$y_d \sim \mathcal{N}(\eta^T \bar{z}_d + \tau^T \bar{w}_d, \rho)$$



Response variables are modeled using both

- Hierarchical topics: each tree node has a regression parameter η_k.
  - To capture **context-specific** polarized words, e.g., "unpredictable" is positive for books but negative for car steering
- Lexical items: each word type has a regression parameter τ_v.
  - To capture **constant** polarized words. e.g., "wonderful", "awesome" are almost always positive; while "horrible", "awful" are almost always negative.

Supervised Hierarchical Latent Dirichlet Allocation



1. For each node $k \in [1, \infty)$ in the tree
   (a) Draw topic $\phi_k \sim \text{Dir}(\beta_k)$
   (b) Draw regression parameter $\eta_k \sim \mathcal{N}(\mu, \sigma)$
2. For each word $v \in [1, V]$, draw $\tau_v \sim \text{Laplace}(0, \omega)$
3. For each document $d \in [1, D]$
   (a) Draw level distribution $\theta_d \sim \text{GEM}(m, \pi)$
   (b) Draw table distribution $\psi_d \sim \text{GEM}(\alpha)$
   (c) For each table $t \in [1, \infty)$, draw a path $c_{d,t} \sim \text{nCRP}(\gamma)$
   (d) For each sentence $s \in [1, S_d]$, draw a table indicator $t_{d,s} \sim \text{Mult}(\psi_d)$
      i. For each token $n \in [1, N_{d,s}]$
         A. Draw level $z_{d,s,n} \sim \text{Mult}(\theta_d)$
         B. Draw word $w_{d,s,n} \sim \text{Mult}(\phi_{c_{d,t_{d,s}}, z_{d,s,n}})$
   (e) Draw response $y_d \sim \mathcal{N}(\boldsymbol{\eta}^T \bar{\boldsymbol{z}}_d + \boldsymbol{\tau}^T \bar{\boldsymbol{w}}_d, \rho)$:
      i. $\bar{z}_{d,k} = \frac{1}{N_{d,\cdot}} \sum_{s=1}^{S_d} \sum_{n=1}^{N_{d,s}} \mathbb{I}[k_{d,s,n} = k]$
      ii. $\bar{w}_{d,v} = \frac{1}{N_{d,\cdot}} \sum_{s=1}^{S_d} \sum_{n=1}^{N_{d,s}} \mathbb{I}[w_{d,s,n} = v]$

Inference

We approximate SHLDA's posterior using stochastic EM, alternating between Gibbs sampling and optimization.

Gibbs sampling:

- **Sampling t–table assignments for sentences**:

$$P(t_{d,s} = t \mid \text{rest}) \propto \begin{cases} S_{d,t}^{-d,s} \cdot f_{c_{d,t}}^{-d,s}(w_{d,s}) \cdot g_{c_{d,t}}^{-d,s}(y_d), & \text{for existing table } t; \\ \alpha \cdot \sum_{c \in \mathcal{C}^+} P(c_{d,t^{new}} = c \mid c^{-d,s}) \cdot f_c^{-d,s}(w_{d,s}) \cdot g_c^{-d,s}(y_d), & \text{for new table } t^{new}. \end{cases}$$

where the probability of assigning the table $c_{d,t^{new}}$ to a path c is

$$P(c_{d,t^{new}} = c \mid c^{-d,s}) \propto \begin{cases} \prod_{l=2}^{L} \frac{M_{c,l}^{-d,s}}{M_{c,l-1}^{-d,s} + \gamma_{l-1}}, & \text{for an existing path } c; \\ \frac{\gamma_{l^*}}{M_{c^{new},l^*}^{-d,s} + \gamma_{l^*}} \prod_{l=2}^{l^*} \frac{M_{c^{new},l}^{-d,s}}{M_{c^{new},l-1}^{-d,s} + \gamma_{l-1}}, & \text{for a new path } c^{new} \end{cases}$$

- **Sampling z–level assignments for tokens**:

$$P(z_{d,s,n} = l \mid \text{rest}) \propto \frac{m\pi + N_{d,\cdot,l}^{-d,s,n}}{\pi + N_{d,\cdot,\geq l}^{-d,s,n}} \prod_{j=1}^{l-1} \frac{(1-m)\pi + N_{d,\cdot,>j}^{-d,s,n}}{\pi + N_{d,\cdot,\geq j}^{-d,s,n}} \cdot f_{c_{d,t_{d,s}}}^{-d,s,n}(w_{d,s,n}) \cdot g_{c_{d,t_{d,s}}}^{-d,s,n}(y_d)$$

- **Sampling c–path assignments for tables**:

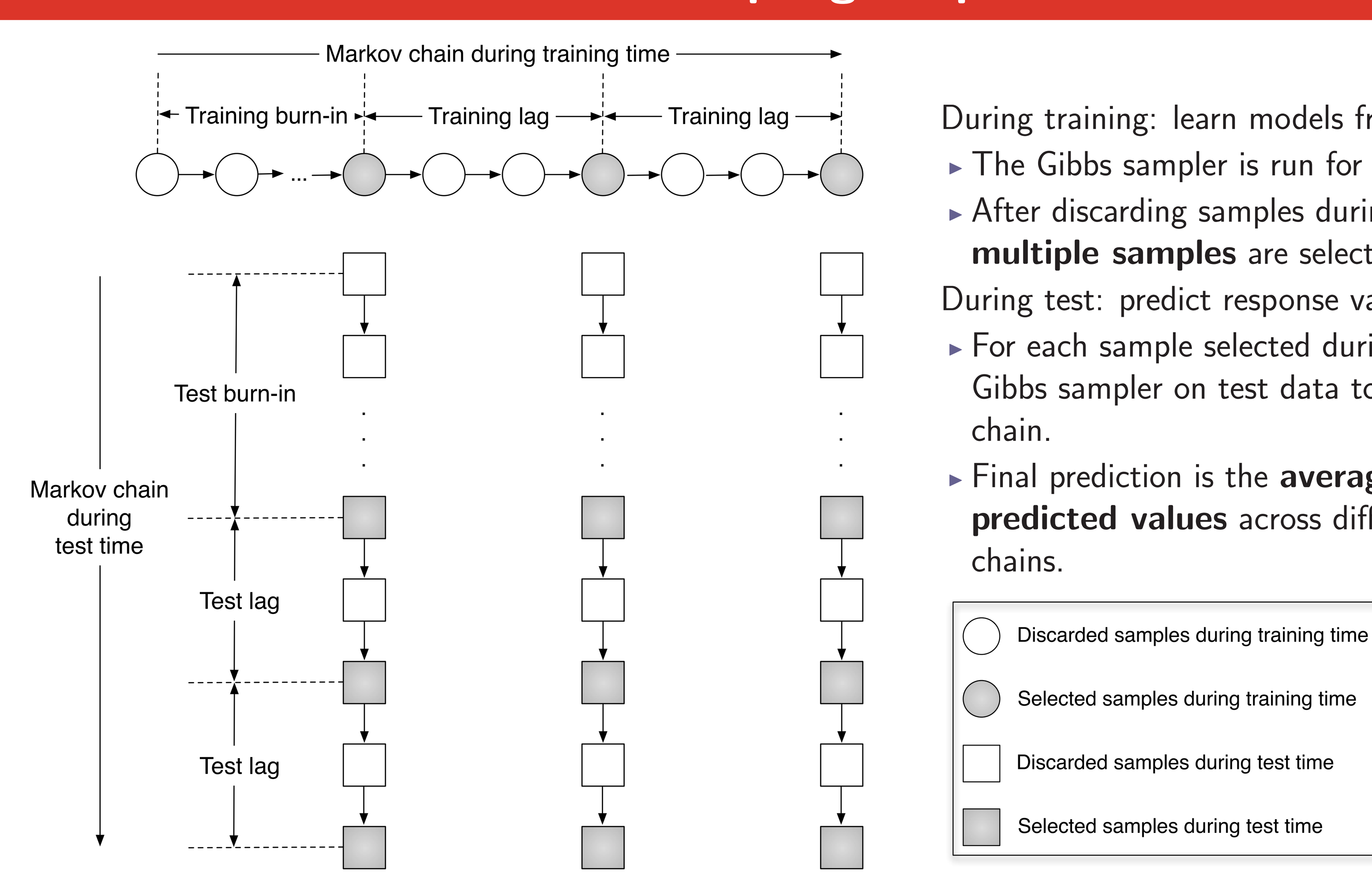
$$P(c_{d,t} = c \mid \text{rest}) \propto P(c_{d,t} = c \mid c^{-d,t}) \cdot f_c^{-d,t}(w_{d,t}) \cdot g_c^{-d,t}(y_d)$$

where $f_c^{-d,x}(v_{d,x})$ and $g_c^{-d,x}(y_d)$ respectively denote the conditional density of $v_{d,x}$ and y_d given that $v_{d,x}$ are assigned to path c.

Optimizing η and τ: We optimize the regression parameters using L-BFGS via the likelihood

$$\mathcal{L}(\eta, \tau) = -\frac{1}{2\rho} \sum_{d=1}^{D} (y_d - \eta^T \bar{z}_d - \tau^T \bar{w}_d)^2 - \frac{1}{2\sigma} \sum_{k=1}^{K^+} (\eta_k - \mu)^2 - \frac{1}{\omega} \sum_{v=1}^{V} |\tau_v|$$

Gibbs sampling for prediction



During training: learn models from training data

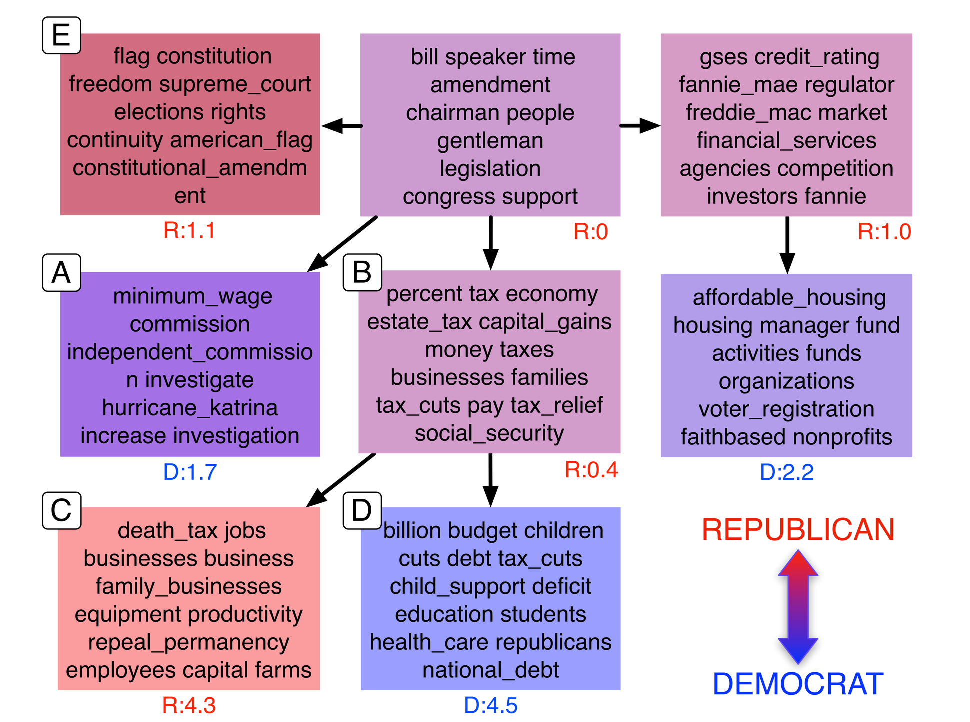
- The Gibbs sampler is run for a number of iterations.
- After discarding samples during the burn-in period, **multiple samples** are selected.

During test: predict response variable for unseen data

- For each sample selected during training time, run a Gibbs sampler on test data to obtain a Markov chain.
- Final prediction is the **average of multiple predicted values** across different test Markov chains.

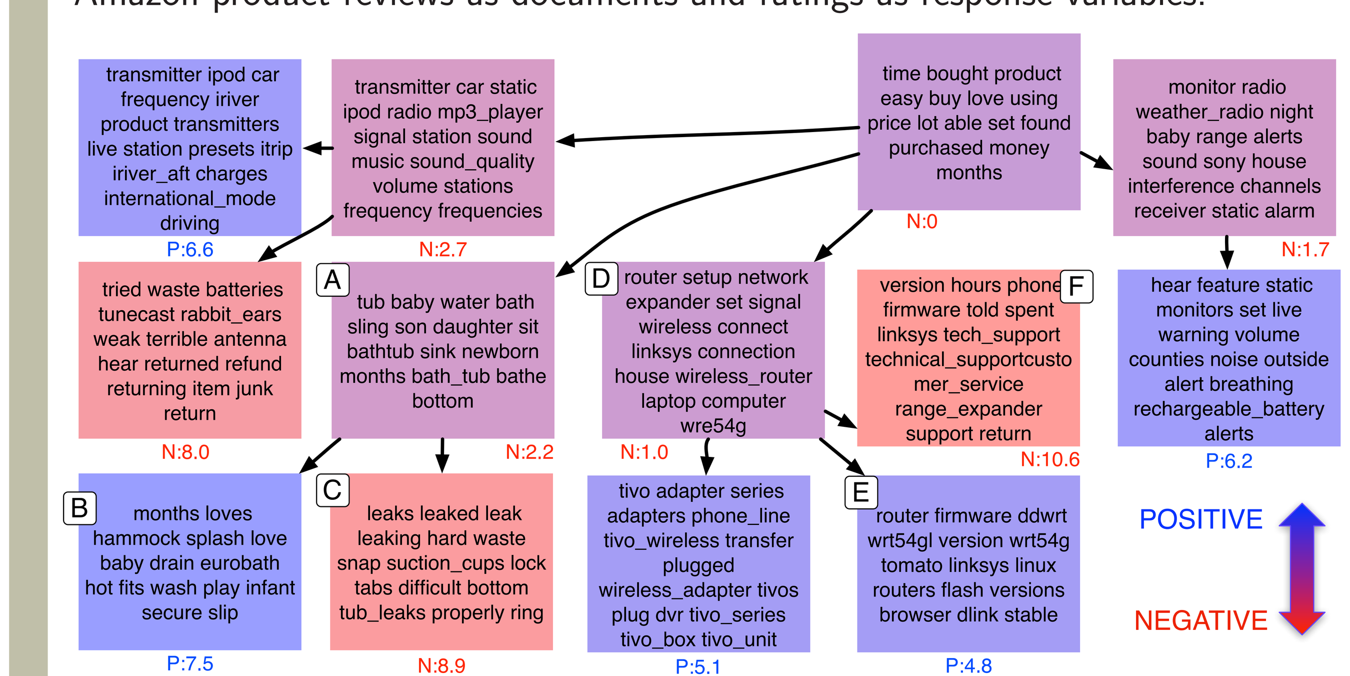
Example hierarchy: Congressional floor debates

Congressional debate turns as documents and speakers' ideological scores as response variables.



Example hierarchy: Amazon reviews

Amazon product reviews as documents and ratings as response variables.



Predicting response variables

Datasets:

- U.S. Congressional floor debates: 5,201 debate turns in the House and 3060 debate turns in the Senate of the 109th U.S. Congress.
- Amazon product reviews: 37191 reviews on manufactured products such as computers, MP3 players, GPS devices etc
- Movie reviews: 5006 movie reviews

Baselines:

- Support vector regression (SVM)
- Multiple linear regression (MLR)
- Supervised latent Dirichlet allocation (SLDA)

Evaluation metrics:

- Pearson's correlation coefficient (PCC, higher is better ↑)
- Mean squared error (MSE, lower is better ↓)

Models	Floor Debates				Amazon Reviews		Movie Reviews	
	House-Senate		Senate-House					
	PCC ↑	MSE ↓	PCC ↑	MSE ↓	PCC ↑	MSE ↓	PCC ↑	MSE ↓
SVM-LDA$_{10}$	0.173	0.861	0.08	1.247	0.157	1.241	0.327	0.970
SVM-LDA$_{30}$	0.172	0.840	0.155	1.183	0.277	1.091	0.365	0.938
SVM-LDA$_{50}$	0.169	0.832	0.215	1.135	0.245	1.130	0.395	0.906
SVM-VOC	0.336	1.549	0.131	1.467	0.373	0.972	0.584	0.681
SVM-LDA-VOC	0.256	0.784	0.246	1.101	0.371	0.965	0.585	0.678
MLR-LDA$_{10}$	0.163	0.735	0.068	1.151	0.143	1.034	0.328	0.957
MLR-LDA$_{30}$	0.160	0.737	0.162	1.125	0.258	1.065	0.367	0.936
MLR-LDA$_{50}$	0.150	0.741	0.248	1.081	0.234	1.114	0.389	0.914
MLR-VOC	0.322	0.889	0.191	1.124	0.408	0.869	0.568	0.721
MLR-LDA-VOC	0.319	0.873	0.194	1.120	0.410	**0.860**	0.581	0.702
SLDA$_{10}$	0.154	**0.729**	0.090	1.145	0.270	1.113	0.383	0.953
SLDA$_{30}$	0.174	0.793	0.128	1.188	0.357	1.146	0.433	0.852
SLDA$_{50}$	0.254	0.897	0.245	1.184	0.241	1.939	0.503	0.772
SHLDA	**0.356**	0.753	**0.303**	1.076	**0.413**	0.891	**0.597**	**0.673**

Results on Amazon product reviews and movie reviews are averaged over 5 folds. For the debate corpus, documents in the House is used to train and test on documents in the Senate (House-Senate) and vice versa (Senate-House).

Email: vietan@cs.umd.edu, jbg@umiacs.umd.edu, resnik@umd.edu