Modeling Influence in Text Corpora

Sean Gerrish and David Blei

11 December 2009

Identifying influential documents

Identifying *influential* documents is a pervasive challenge for many researchers

- Historiography
- Academic research
- Much of Bibliometrics

There are many specific application areas

- News articles
- Legal opinions
- Scientific impact
- Transcriptions of radio content and orations

Predicting and understanding citations

Often citations provide useful information.

Existing research often aims to predict citation counts using a discriminative classifier and specific features, e.g.:

- Document length
- Citations to first author
- Citations to last author
- Key words
- Journal

Or some analyses focus on the citation level:

- Using topics to predict citation influence [3]
- Understanding the influence of blogs [4]
- Relational topic Models [2]

Our goal

1. Our *goal* is to predict the influence of documents without additional information such as citations, i.e. using *just their words*.
2. Our *intuition* is that influential documents change the language of their fields (i.e., their topics).
3. We define influence to be item 2: influential documents change the language of their topics.

Modeling documents with changing topics

There are a number of approaches to modeling changing topics

- Topics over time [6]
- Dynamic Topic Models [1]
- Dynamic Mixture Models [7]

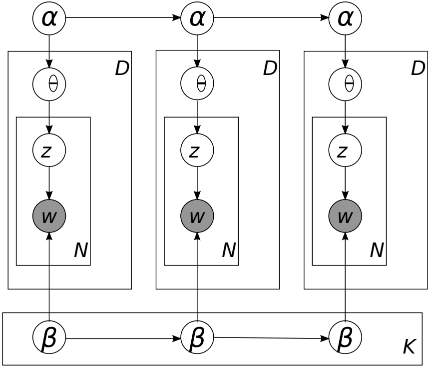
We chose to extend the Dynamic Topic Model.

The Dynamic Topic Model

Assumes topics drift in a Markov chain:

$\beta_t \sim \mathcal{N}(\beta_{t-1}, \sigma^2)$

$D_t \sim \mathrm{LDA}(\alpha_t, \beta_t)$

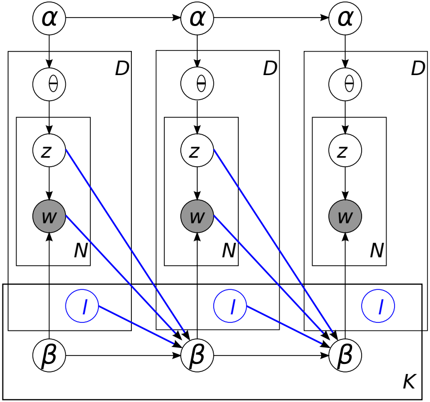


The Document Influence Model

Assumes each document has a weight which affects topic drift...

$l_{d,k} \sim \mathcal{N}(0, \sigma_l^2)$

$\beta_t \sim \mathcal{N}(\beta_{t-1} + \mathrm{Inf}(l_{t-1}, z_{t-1}, w_{t-1}), \sigma^2)$

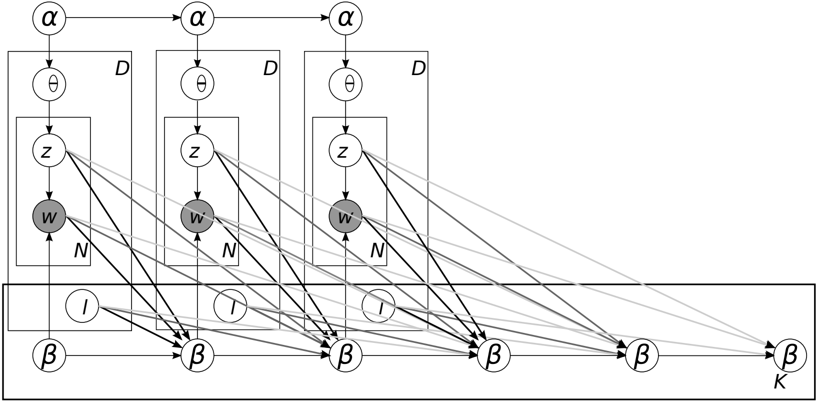


The Document Influence Model

... with documents having potential influence in the distant future.

$l_{d,k} \sim \mathcal{N}(0, \sigma_l^2)$

$\beta_t \sim \mathcal{N}(\beta_{t-1} + \mathrm{Inf}(l_{s<t}, z_{s<t}, w_{s<t}), \sigma^2)$

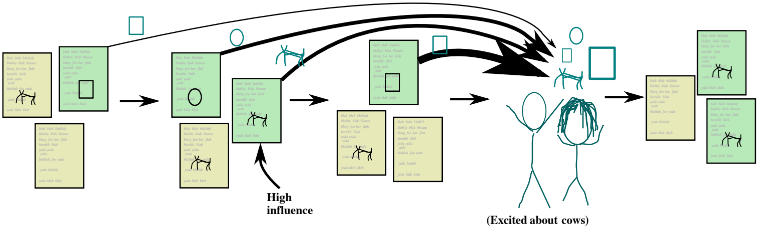


The DIM influence function

Markov step: $\beta_{t,k} \sim \mathcal{N}(\beta_{t-1,k} + Inf(t,k), \sigma^2 I)$,

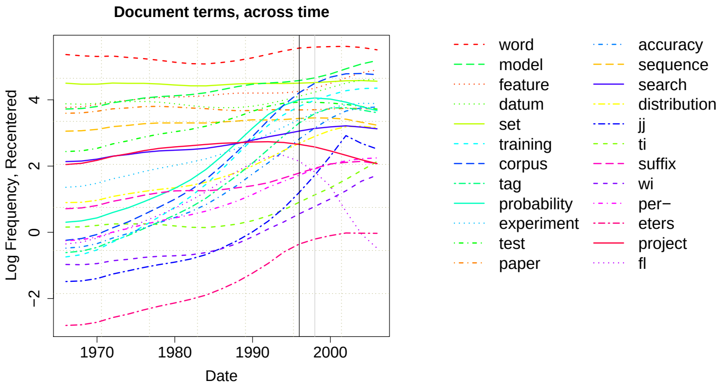
$Inf(s,k) := \exp(-\beta_{s-1,k}) \circ \sum_{i=0}^{s-1} r(s-1-i)([z_i]_k \circ \mathbf{W}_i) l_{i,k}$,

- $r(j)$ is the fraction of a document's influence after j years (called the influence envelope), and
- $[z]_k$ is the indicator describing whether term z is in topic k.
- $\exp(-\beta_{s-1,k})$ is a correction term to make sure our units are correct for log-space drift.



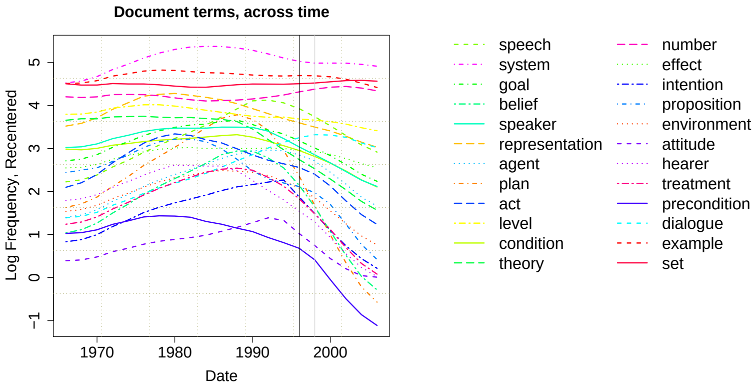
A Closer Look

A Maximum Entropy Model for Part-of-Speech Tagging



A Closer Look

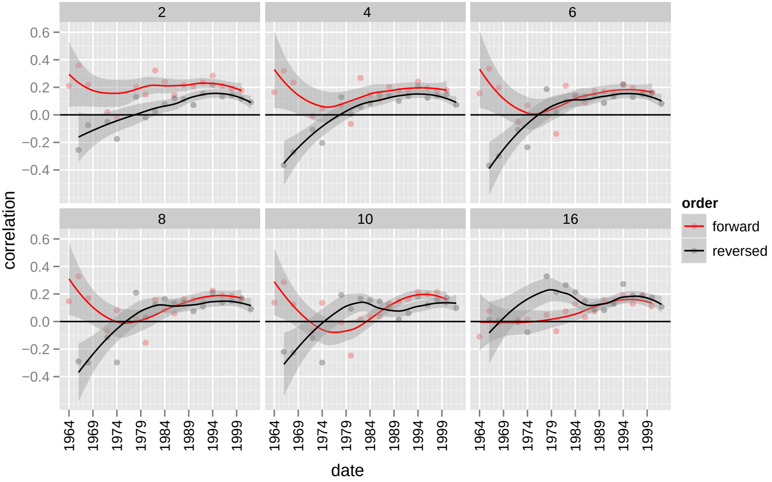
An Ascription-Based Approach to Speech Acts



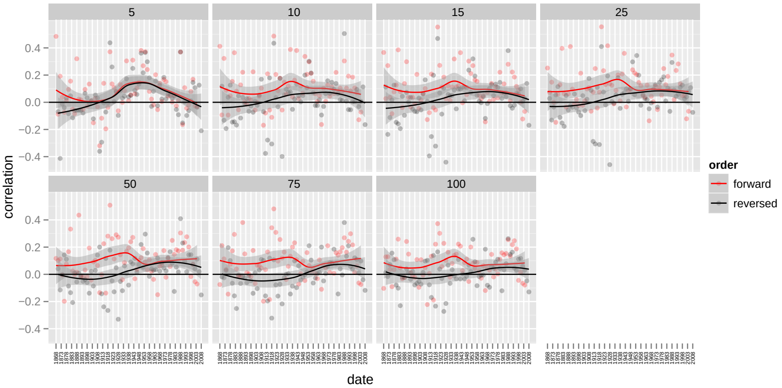
Experiments

- Procedure
  - Derive influence values from corpus
  - Compute correlation with citation counts
- Corpora
  - The *ACL Anthology*
  - *Nature*
- Evaluation with citations
  - ACL Anthology Network [5]
  - Google Scholar

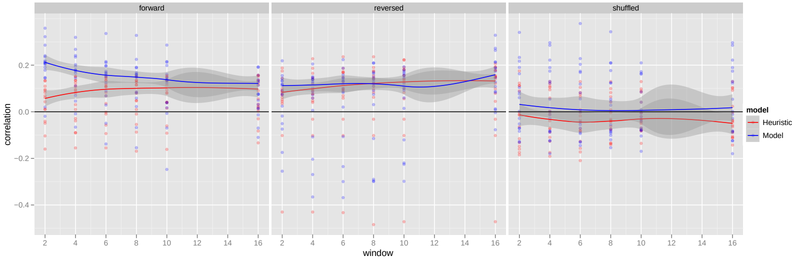
Experiments - ACL



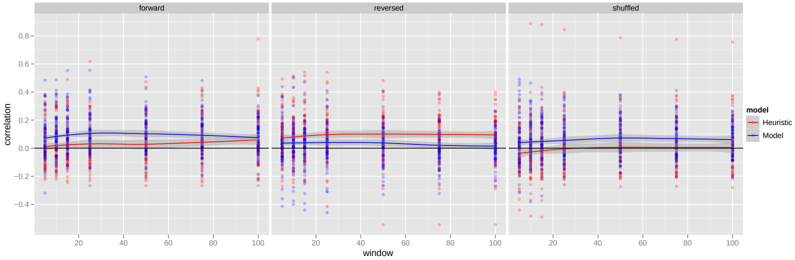
Experiments - Nature



Experiments - ACL validation



Experiments - Nature validation



Summary

- Document Influence Model
- Inference
- Evaluation
- Significance of baseline

Bibliography I

D. Blei and J. Lafferty.
Dynamic topic models.
Proc. of the 23rd ICML, 2006.

J. Chang and D. M. Blei.
Relational topic models for document networks.
Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AIStats) 2009, 5, 2009.

L. Dietz, S. Bickel, and T. Scheffer.
Unsupervised prediction of citation influences.
In *ICML*, 2007.

R. Nallapati and W. Cohen.
Link-plsa-lda: A new unsupervised model for topics and influence of blogs.
International Conference for Weblogs and Social Media, 2008.

D. R. Radev, M. T. Joseph, B. Gibson, and P. Muthukrishnan.
A Bibliometric and Network Analysis of the field of Computational Linguistics.
Journal of the American Society for Information Science and Technology, 2009.

X. Wang and A. McCallum.
Topics over time: A non-markov continuous-time model of topical trends.
Conference on Knowledge Discovery and Data Mining (KDD), 2006.

X. Wei, J. Sun, and X. Wang.
Dynamic mixture models for multiple time series.
IJCAI, 2007.

Motivation for $\exp(-\beta)$ coefficient in $Inf(t, l, z, w)$

$$
\begin{aligned}
& \exp(\beta_t) = \exp(\beta_{t-1}) + Inf_t \\
\iff \quad & 1 = \exp(\beta_{t-1} - \beta_t) + \exp(-\beta_t) Inf_t \\
\iff \quad & 1 - \exp(-\beta_t) Inf_t = \exp(\beta_{t-1} - \beta_t) \\
\iff \quad & \log(1 - \exp(-\beta_t) Inf_t) = \beta_{t-1} - \beta_t \\
\iff \quad & \beta_t = \beta_{t-1} - \log(1 - \exp(-\beta_t) Inf_t)
\end{aligned}
\tag{1}
$$

Note that when $\exp(-\beta_t) Inf_t$ is small, we have $\beta_t \approx \beta_{t-1} + \exp(-\beta_t) Inf_t$.

Regularized linear regression for $\tilde{l}$ updates

$$g(s,q) := \Lambda_{\exp(-\tilde{m}_{q,k}+\tilde{V}_{q,k}/2)}(\mathbf{W}_{s,k} \circ \phi_{s,k}) \tag{2}$$

$$h(s,q) := ((\mathbf{W}_{s,k} \circ \phi_{s,k})^T \Lambda_{\exp(-2\tilde{m}_q+2\tilde{V}_q)+\exp(-2\tilde{m}_q+\tilde{V}_q)}(\mathbf{W}_{s,k} \circ \phi_{s,k}) \tag{3}$$

$$+\Lambda_{(\mathbf{W}_{s,k}\circ\mathbf{W}_{s,k}\circ(\phi_{s,k}-\phi_{s,k}\circ\phi_{s,k}))^T(\exp(-2\tilde{m}_q+2\tilde{V}_q)+\exp(-2\tilde{m}_q+\tilde{V}_q))}\big) \tag{4}$$

$$\tilde{l}_{t,k} \leftarrow \left(\frac{\sigma^2}{\sigma_d^2} I + \left(\sum_{i=t}^{T-1} r(i-t)^2 h(t,i)\right)\right)^{-1}$$
$$\left(\sum_{i=t}^{T-1} r(i-t) g(t,i)^T (\tilde{m}_{i+1,k} - \tilde{m}_{i,k} + \tilde{V}_{i,k} - \sum_{j=0\ldots i, j\neq t} r(i-j) g(j,i) \tilde{l}_{j,k})\right) \tag{5}$$

Inference

We use structured variational inference, as in the DTM.

- Nonconjugacy makes sampling methods more difficult.
- Document-level variational parameters as in LDA.
- Models $\hat{\beta}$ variational parameters as observations of a Markov chain. By the symmetry of the Gaussian, we can use backward-forward Kalman updates for these parameters.
- Update for $\hat{l}_{d,k}$ is regularized linear regression.

The DIM generative model

For time $t = 1, \ldots, T$:

- For topic $k = 1, \ldots, K$:
  Draw natural parameters
  $\beta_{t,k} | \beta_{t-1,k}, \mathbf{z}_{s<t}, l_{s<t} \sim \mathcal{N}(\beta_{t-1,k} + Inf(t,k), \sigma^2 I)$
- For each document d_t:
  - Generate document d_t using traditional LDA with parameters α_t and β_t.
  - For topic $k = 1, \ldots, K$, draw document weight $l_{d,k} \sim \mathcal{N}(\mathbf{0}, \sigma_d^2 I)$;