Modeling Influence in Text Corpora

Sean Gerrish
Department of Computer Science
Princeton University
sgerrish@cs.princeton.edu

David M. Blei
Department of Computer Science
Princeton University
blei@cs.princeton.edu

Abstract

Identifying the most influential documents in a corpus is an important problem in a wide range of fields, ranging from information science and historiography to text summarization and news aggregation. We propose using changes in the linguistic content of these documents over time to predict the importance of individual documents within the collection and describe a dynamic topic model for both quantifying and qualifying the impact of each document in the corpus.

Introduction In many fields, identifying the most influential documents is an important challenge; researchers in information science, historiography, text summarization, and news aggregation, for example, are all interested in identifying influential documents in their respective fields. Researchers in fields like these often use traditional methods of assessing the impact of an article such as analyzing the citations to it: the impact factor of a journal, for example, is based largely on academic citation analysis; and Google's successful PageRank algorithm is based on hyperlink citations between Webpages [1].

Often, however, citation information is not present: certain legal documents, news stories, blog posts, and email, for example, all might lack such citation metadata, while there is a clear notion of influence among articles in these collections. The goal of this work is to develop an unsupervised method for determining the influence of a document in the *absense* of citations. Our intuition is that language changes over time, and that influential documents contribute to this change.

We formalize this intuition and present an algorithm which takes a sequence of documents as input and computes a vector of "influence" for each document. This vector characterizes the document's influence in terms of themes discovered using a topic model we have developed. We validate this method by measuring how well the computed impact predicts citations and demonstrate that this method provides a citation-free measure of bibliometric impact.

The Document Influence Model We base our model, the Document Influence Model (DIM), on the the Dynamic Topic Model (DTM) [2]. The DTM models corpora by assuming that documents are mixtures of themes; it models these themes, in turn, by allowing them to change over time. The concept of a theme in the DTM is formalized as a topic using Latent Dirichlet Allocation (LDA) [3].

We extend the DTM by associating each document d_t with a vector of topic weights $\vec{l}_{d_t,k}$ (see Figure 1). These weights express how much the language used in document d_t affects the drift of these topics over a period of time. The more influential a document is in a topic k, the larger its topic weights should be, making it more likely that future documents about topic k will use the same language.

The influence of a document on each topic k works as follows. We assume that document d_t at time t may have some influence on the language used within each topic. The more influential d_t is on topic k (i.e., the higher its weight for this topic)–and the more its words are "about" topic k in the first place–the more it "nudges" this topic's natural parameters $\beta_{t,k}$ in log space. In LDA, and hence the DIM, the topic assignment for word n is given by the random variable $\mathbf{z}_n$; its role is clarified below the generative model.

The full generative model at time t is then:

1. For topic $k = 1,\ldots,K$:

   Draw natural parameters $\beta_{t,k}|\beta_{t-1,k}, \mathbf{z}_{s<t}, l_{s<t} \sim \mathcal{N}\left(\beta_{t-1,k} + \exp(-\beta_{t-1,k}) \circ \left(\sum_{i=0}^{t-1} r(t-1-i)g(t), \sigma^2 I\right)\right)$, with $g(s)$ and $r(s)$ defined below.

2. For each document d_t:
   (a) Generate all documents at time t using LDA with topics β_t.
   (b) For topic $k = 1,\ldots,K$, draw document weight $\vec{l}_{d,k} \sim \mathcal{N}(\mathbf{0}, \sigma_l^2 I)$,

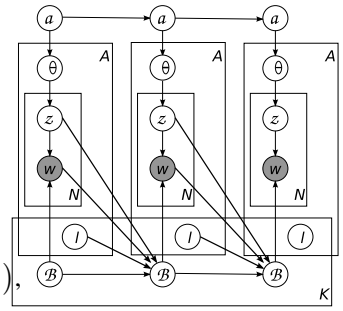

Figure 1: The Document Influence Model.

where above we have $g(s) := ([\mathbf{z}_s]_k \circ \mathbf{W}_s) l_{s,k}$, with $\circ$ denoting the element-wise Hadamard product, $r(j)$ the fraction of a document's influence after j years, and $[z]_k$ the indicator describing whether term z is in topic k. Specific parameters are described in more detail in the following section.

In general, we do not want a document about the Earth's magnetic fields to influence the use of words like "magnet" in the field of medicine. On the other hand, a document about fMRI should certainly be allowed to influence the use of "magnet" in the field of medicine. In our model, this intuition is captured by the indicator $[\mathbf{z}]_k$ in $g(s)$: documents not about medicine will tend to have fewer words from the medicine topic, hence they will affect medicine less.

In the model, $r(t)$, the *influence envelope*, can be interpreted as the fraction of a document's influence on the language of some topic, t epochs after it appears, with $r(i) > 0$ and $\sum r(i) = 1$. We assume here that $r(t)$ is fixed for all documents and all time, although we plan to experiment with a more flexible model of r in the future. For the experiments and results described in this paper, we have fixed $r(0) = 1$, although we are experimenting with different envelope functions.

We believe that it is incorrect to simply add the words in each document to the mean parameters, because this is like comparing apples to oranges. The coefficient $\exp(-\beta_{t-1})$ in the Markov step handles this by mapping the documents, which are in "word-space", to the space of natural parameters. Recalling that β gives the natural parameters of a topic, and $\exp(\beta)$ gives the mean parameters (relative word frequencies) of that topic, we obtain this result by considering the contributions I of documents for time t: $\exp(\beta_t) = \exp(\beta_{t-1}) + I \implies \beta_t = \beta_{t-1} + exp(-\beta_{t-1})\log(1+I)$. When the influence I is small, $\beta_t \approx \beta_{t-1} + \exp(-\beta_{t-1})I$, yielding the Markov step.

We fit this model using variational inference. While inference for this model is similar to that described in [2], some updates are new or significantly different. The updates for the variational parameters behind l in particular are a series of regularized linear regressions.

Results To confirm that this model is finding patterns of influence, we have validated it on 11115 articles in the ACL Anthology Network corpus [4], a collection of Computational Linguistics articles with annotated internal citations. We also fit this model on a large (20%) subset of the Nature journal.

Figure 2 depicts how the mean parameters of some words from two selected documents vary over time. The article *A Maximum Entropy Model For Part-Of-Speech Tagging* was the highest-influence document published in 1996, with 215 citations in the ACL Anthology. We can also ask which terms in each document have had the most influence; one metric of this is the decrease in squared error for this term caused by introduction of this document; we call this metric "Residual Improvement". The most significant words by this metric in *A Maximum Entropy Model*, for example, are "word", "model", "feature", and "datum".

Our principal experiment aimed to determine whether the influence score is correlated with citation counts. We fit a single-topic model and a 20-topic model over the ACL Anthology. Controlling for date, documents' posterior influence scores under this model are significantly correlated with log of citation (plus one): the single-topic model achieved an adjusted R^2 of 0.07 (i.e., it captured 7% of the variance of citations), and the 20-topic model achieved an adjusted R^2 of 0.18, both found using permutation tests.

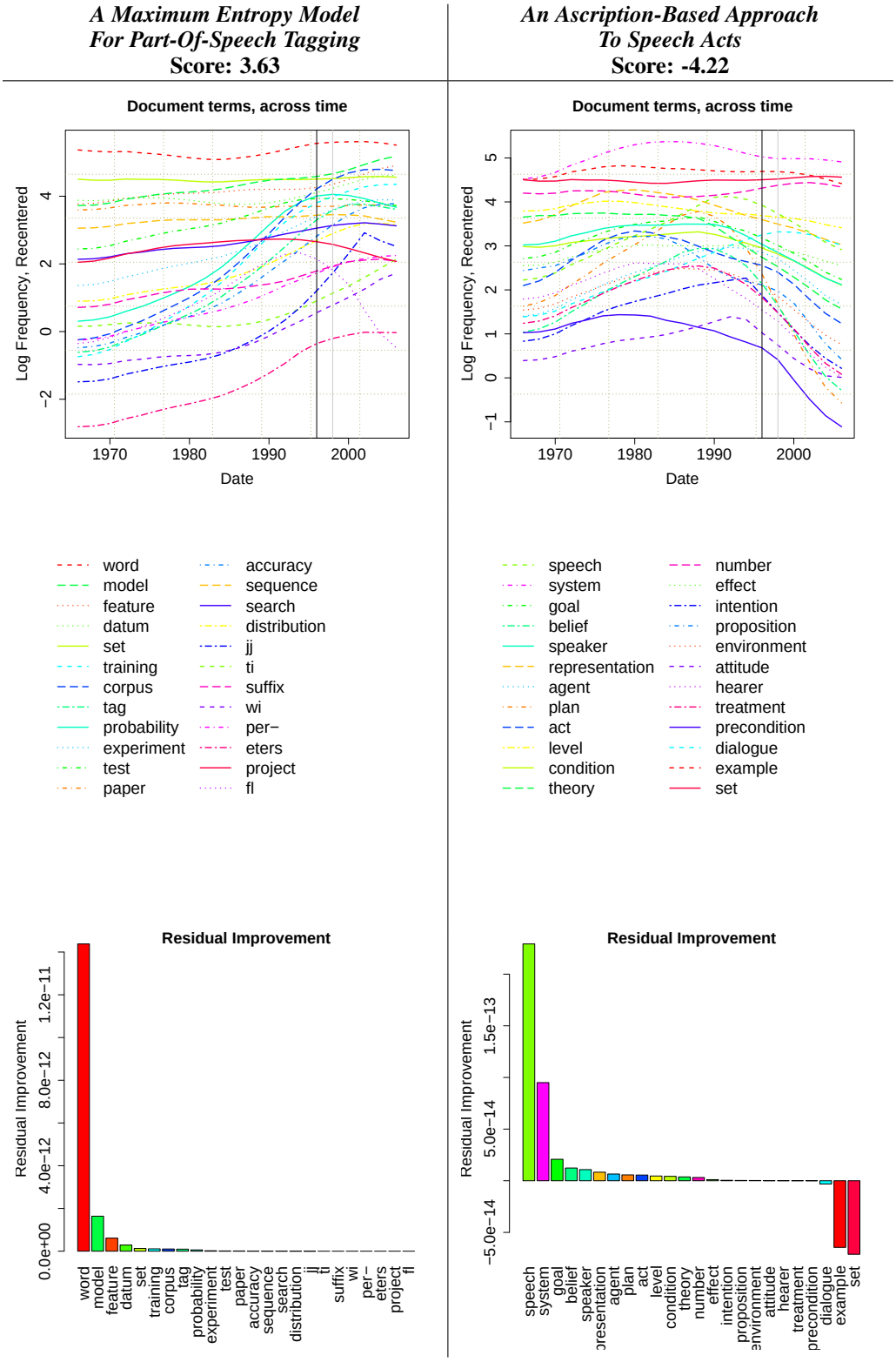

Figure 2: Selected documents from the ACL Anthology having highest and lowest scores in 1996. Note that the scale for Residual Improvement is extremely small due to the $\exp(-\beta)$ apples-to-apples coefficient in the model.

We have also found that posterior influence scores under this model are significantly more correlated with citations than when the corpus's dates are shuffled randomly and the model refit on these shuffled dates. Further, among the top documents in each year of the corpus, 75% are cited more than the median; and 41% are in the upper quartile.

Further Work While we are excited about our current results, the Document Influence Model is a work in progress. We are currently collecting citation counts for Nature articles for additional model validation and fitting the model over documents from the Proceedings of the National Academy of Sciences.

In future work, we plan to experiment with several variants of this model. We are currently working with models in which the influence of a document is spread out over a nontrivial period of time, instead of being applied instantaneously – which we believe is a more realistic model of the influence of a document. In addition, we hope to explore a richer set of models of the flow of ideas, including models of topics as birth and death processes (also noted in [2]) and tracking which documents may influence other documents.

We also hope to better understand this model in the context of traditional metrics of influence, such as academic citations. Understanding when this model and citation counts differ, for example, will help to understand how our metric may fall short when applied to corpora in the wild; in the same vein, it may also provide insights into limitations of established metrics.

References

[1] Brin, S., L. Page. "The anatomy of a large-scale hypertextual web search engine." In "Computer Networks and ISDN Systems," 107–117. 1998.

[2] Blei, D., J. Lafferty. "Dynamic topic models." *Proc. of the 23rd ICML*, 2006.

[3] Blei, D. M., A. Y. Ng, M. I. Jordan. "Latent dirichlet allocation." *JMLR*, 2003.

[4] Joseph, M. T., D. R. Radev. "Citation analysis, centrality, and the acl anthology." Tech. Rep. CSE-TR-535-07, University of Michigan. Department of Electrical Engineering and Computer Science, 2007.