Overview Understanding the flow of ideas over time is an important challenge for scholars, from historiographers to students new to a field. Understanding this flow of ideas can be helped by understanding which documents are influential in the progression of these ideas. Document influence is often determined by looking at networks of citations, such as academic citations or Web Hyperlinks. Citations have proven to be a powerful metric, providing the foundation for such tools as the Impact Factor and Google's Pagerank.

Web hyperlinks and citations are often not available: many collections of text simply do not have them, or this information is hard to collect. Therefore the specific goal of this work is to identify documents which are important from within a corpus – and to understand how they are important – based on changes in the language used over time.

Although this work aims to address cases for which there are no citations, we can still build, test, and fit models on a corpus which has citations, without using them. Then these citations (or other metrics) can be used as a ground truth to determine whether the model is finding patterns that really exist. It is a nongoal of this project to predict citations: work already exists in this area, and were are interested more in studying the flow of ideas than finding features for citation counts.

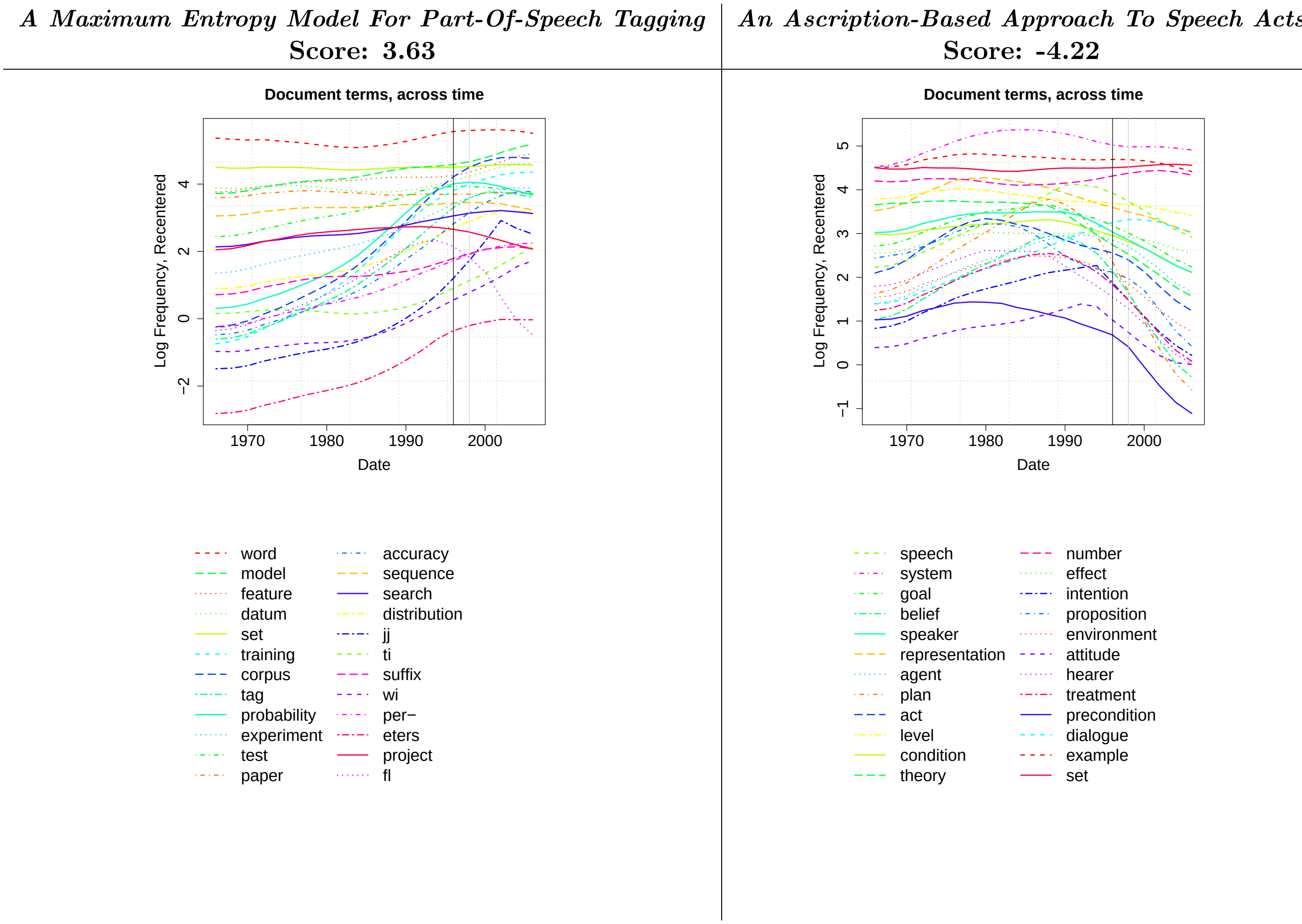
Implementation Details To do this, I will build off of a *dynamic topic model.* Topic models–such as LSI, pLSI, and Latent Dirichlet Allocation (LDA)–are unsupervised ways to find themes in collections of text; using these will allow us to identify themes, and then to identify documents which may be changing these themes. A *dynamic* topic model is a topic model using LDA which tracks the changes in these topics over time.

The current extention of dynamic topic models is simple: we assume each topic drifts over time. Documents – which fall into one or more topics – have an effect on the drift of these topics, pushing the language used in each topic in the direction of words in the document. In the current model we have looked at so far, we assume that the documents at each specific point in time describe the instantaneous "velocity" of words found in a topic. Figure 1 shows examples of two documents found using this method.

A large part of the research focus here lies in the probabilistic models used to represent this data, along with the mathematical "inference" algorithms used to fit these models. I will likely use a form of inference called *variational inference*, because it is fast once implemented – but deriving the math and implementing it can take some work (other options include Markov chain Monte Carlo methods for fitting the model; these are often faster to implement but run more slowly once implemented).

Visualization The visualization for this project will hopefully highlight at least two factors in the influence of documents: how influential is a document, and in what ways (e.g., in what topics) is it influential? This visualization will also depict the flow of ideas in a corpus, possibly by highlighting topics which are prominent or changing at quick pace in the corpus.

Open questions Here are a few areas open for discussion. Your feedback is welcome:



A Maximum Entropy Model For Part-Of-Speech Tagging
Score: 3.63

An Ascription-Based Approach To Speech Acts
Score: -4.22

Figure 1: Documents from the Association for Computational Linguistics Anthology having highest score (*A Maximum Entropy Model*) and lowest score (*An Ascription-Based Approach*) in 1996. This used a one-topic model.

What visualizations should we use? The visualizations will ideally be able to serve as a browser of influential documents while giving the user a sense of the flow of ideas in a corpus.

Which corpus / corpora should we start with? Ideally, the collection will have information about citations for each document.

Which model assumptions should we focus on? I will continue working with a specific model built with my advisor but hope to make improvements throughout the summer. Certain assumptions will be more clear as I give more details over the next few weeks.