MAD Style: Multivalent Authorship Detection (MAD) Topic Models

David Dohan, Charles Marsh, Shubhro Saha, Max Simchowitz
Princeton University, Department of Computer Science

Goals

- Classify author writing style in a wide range of media.
- Extract compact representation of stylistic tendency.
- Determine which features are most indicative of writing style.

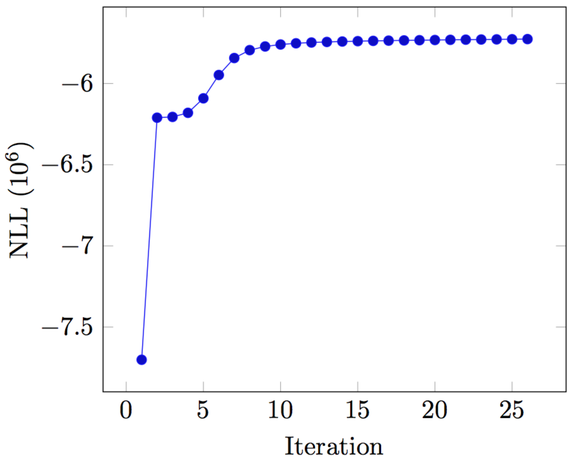
Introduction

In the *authorship detection* problem, one is given:

- A set of documents labeled (by author) on which to train.
- A set of anonymized documents to classify.

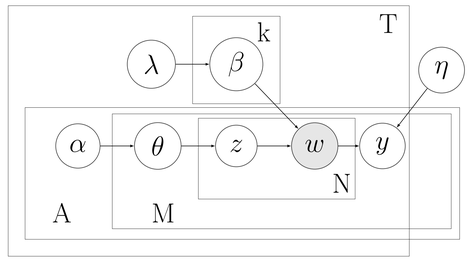
Methods for authorship detection traditionally depended on careful feature extraction and rather black-box methods. Hence, they rely on extensive domain specific knowledge, and can be difficult to decipher.

We present the *MAD Topic Model*, which uses syntactic and stylometric n-gram features (e.g., part-of-speech tags, meter). MAD fits separate topic models to each of these n-gram vocabularies and combines the models through a multi-class logistic regression classifier. After fitting the topic model parameters, new documents can be classified using the multi-class component. As a by-product, MAD also breaks stylistic features into vocabularies over topics, creating a compact representation of stylistic tendency.



MAD's increasing Negative Log Likelihood (NLL) over time.

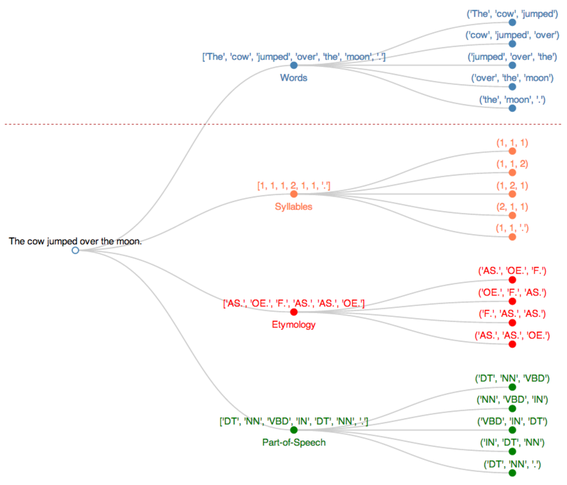
Model



Graphical Model for the MAD Topic Model

The MAD topic model combines the SLDA algorithm presented in [1] with an Author Topic Model, and extends both to account for multiple word types. The model is based on variational inference, following the coordinate ascent updates in [1]. Stochastic variational inference was also tested, but proved impractical for these rather small data sets.

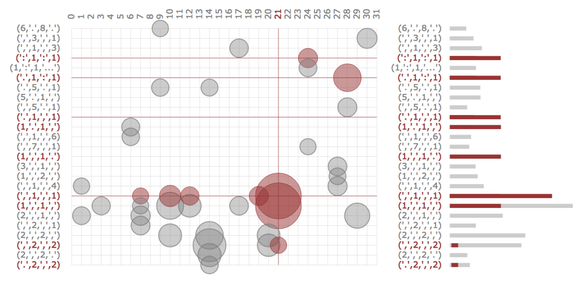
Features



Feature extraction for the MAD Topic Model. Word and syllable counts (between punctuation) were also included.

Visualization

The MAD Topic Model generates topics over n-grams of stylistic features, which in turn highlights writing's implicit structure. We used the Termite visualization tool to generate graphical representations of our topic models.



A topic model for word count (between punctuation) n-grams. Topic 21 represents short, staccato sentences.

Summary

The Multivalent Authorship Detection (MAD) Topic Model extends Latent Dirichlet Allocation to identify authorship in documents with many separate types ("multivalent") of count features. MAD is "doubly supervised": it includes a multi-class logistic regression and also fits per-author Dirichlet distributions for each feature type. We test the MAD Topic Model on several real world corpora using a variety of n-gram features, including part-of-speech, syllable stress, and sequences of word lengths.

Data

To collect data for training and testing, we wrote Python scrapers for Project Gutenberg, Nassau Weekly, and Quora.

Datasets collected for training and testing

Source	Authors	Docs/Author
Project Gutenberg	5	50
Nassau Weekly	550	200
Quora	1600	100

Project Gutenberg contains excerpts from fictional books. Nassau Weekly features narrative & editorial articles from the campus publication. Quora captures responses from top users on the question-answer site. The diversity in topic, language, and length challenges our model to detect consistent fea-

Results

Unfortunately, preliminary results show that which MAD fares far worse as using the same features with another classification scheme. This is consistent with recent findings that suggest that a Pitman-Yor process better captures power law frequencies in language use than Dirichlet methods. Nevertheless, MAD's topic models over the n-gram stylistic features can be used to extract compact representations of stylistic tendency and discern which features are most indicative of individual writing style.

Conclusion

Our (short) conclusion.

References

[1] Chong Wang, David Blei, and Fei-Fei Li. Simultaneous Image Classification and Annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1903–1910. IEEE, 2009.

PRINCETON UNIVERSITY