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 have traditionally depended on lexical analysis of the text, making them relatively context-dependent.

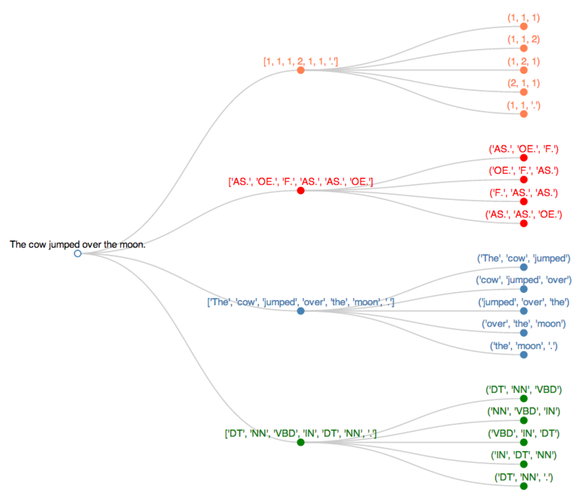
Instead, the *MAD Topic Model* depends solely on syntactic and stylometric features (e.g., part-of-speech tags, meter), which are less context-dependent. MAD treats these features as vocabularies over which topic models can be determined, performing a Supervised Latent Dirichlet Allocation (SLDA) algorithm over n-gram stylistic features to determine authorship of anonymized text.

Preliminary results show significant improvement over more naive techniques (such as Logistic MLE) using the same features. As a by-product, 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 writing style.

Data

To collect data for training and testing, we wrote Python scrapers for Project Gutenberg, Nassau Weekly, and Quora. We selected these three data sources for their diversity in topic, language, and length. Our Project Gutenberg dataset features fictional story excerpts from five, public-domain authors. The Nassau Weekly dataset features over 550 articles from about 200 authors. On the higher end, the Quora dataset features over 1600 comments from roughly 100 popular Quora users. The challenge posed for our authorship models is to detect consistent features in such a variety of contexts.

Features



Visualization

Replace this with some visualization.

Conclusion

Our (short) conclusion.

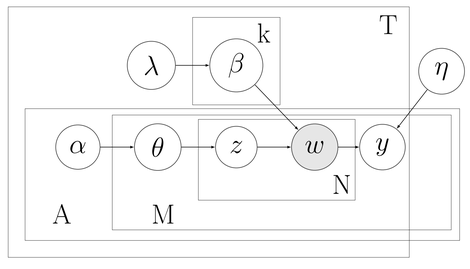
References

PRINCETON UNIVERSITY

Summary

The Multivalence 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 as in [**?**]–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.

Model



Graphical Model for the MAD Topic Model

Results

Our results.