224N Project Proposal

Team

Varun Ramesh (vramesh2), Jean-Luc Watson (jlwatson)

Problem Description

We want to apply RNNs (LSTMs and GRUs) to the problem of stylometry for authorship attribution of disputed works. We plan to look at three problems of increasing scope - the first is the attribution of anonymous works that scholars believe were written by Shakespeare. The second is the identification of Satoshi Nakamoto based off stylometric analysis of forum posts. The third is training a network to embed an author's style into a vector - this network can then be used for authorship analysis on new authors without necessarily having those authors in the RNN training set.

Data

For Shakespeare attribution, we plan to download the undisputed works of Shakespeare and other contemporary playwrights. For the Nakamoto analysis, we will scrape the forum and blog posts of figures who are suspected to be Satoshi Nakamoto. For the general stylometric embeddings, we can acquire as much additional data as needed from the large collection of authors' works hosted online by Project Gutenberg.

Methodology / Algorithm

We plan to split our corpus' into paragraphs. Each paragraph will be fed into an RNN. The final hidden state will then be fed into a softmax classifier with cross-entropy loss, with the labels varying based on the experiment. For Shakespeare, we will perform binary classification to determine authorship, and the Nakamoto trial will be *n*-way classification on potential authors.

For the general embedding, will just use an RNN and interpret the hidden state as an embedding, where euclidean distance represents author similarity. The loss will be formulated such that documents from the same author are pushed together, while documents from different authors are pushed apart.

Related Work

- http://www.jstor.org/stable/30204514 - This paper is the first stylometric analysis that uses neural networks. However, it uses shallow, fully connected neural nets, and predates the wide usage of RNNs.
- https://arxiv.org/abs/1602.05292 - This paper uses neural network language models for authorship attribution. They do not use an RNN.
- https://web.stanford.edu/class/cs224n/reports/2760185.pdf - This prior CS224n project evaluates several stylometric analysis models on multiple datasets. Both LSTM and GRU models are tested.

Evaluation Plan

As described above, we propose evaluating our stylometric methods in three experiments:

- *William Shakespeare / Satoshi Nakamoto*
  For both classification tasks, we will evaluate the model on a hold-out set of undisputed posts/works. We will then run the model on disputed works and qualitatively evaluate the results - in the Shakespeare case, we can compare our results to the current academic consensus. We can visualize our results using a confusion matrix, activation maps, and 2D hidden state embeddings.
- *General Author Embedding*
  We first train a model on a wide variety of written texts. Second, we take a moderate number of authors (e.g. 30), feed paragraphs of their work to generate embeddings, and average the results into a point to represent the author. Finally, we generate embeddings for a hold-out test set of documents from the second set of authors and perform *k*-nearest neighbors on the author point cloud, where mean average precision is used to evaluate the results. We expect to visually display the generated author embeddings to explore the data for meaningful clustering behavior and examine correlation between author style or genre.

Minimal Requirements

- ✓ Dataset with at least 10,000 labeled examples: Shakespeare and his contemporaries have written enough works for analysis. Some Bitcoin developers have written more than a thousand forum / blog posts, and we have public access to self-identified Nakamoto writing in the same format. Finally, Project Gutenberg hosts hundreds of publicly-available raw-text publications that can be easily parsed.
- ✓ Dataset can be completely collected by the project milestone: Scrapers already exist for some forums, and we can download published works in raw text format. We will run an existing tokenizer.
- ✓ Task is feasible: We think the classification problems are feasible - the general embedding may not be. We are open to suggestions regarding the project scope.
- ✓ Identified automatic evaluation metric: We will hold out some undisputed works for evaluation.
- ✓ Using NLP is required to get good performance on the task: For both classification problems, there are other channels of information like post timings and historical records. However, stylometry can serve as one of many data points in authorship analysis.