CIS 545 - Final Project

Peter Kong

December 10, 2018

1 Abstract

We attack an existing authorship identification task, focusing on textual feature extraction paired with exploration in feedforward neural network classifiers.

2 Problem Background

Authorship identification is the task of identifying the authorship of a document, usually by textual analysis. This task has significant real world applications: it can be used to attribute authorship to historical documents whose provenance is disputed.

Many authors, with good or bad intentions, attempt to frustrate identification via various methods, including style modification [Brennan]. Helping hide or reveal authorship of text can save a life or solve a crime, depending on the context.

Authorship identification and its analog tasks are studied heavily in academia. PAN, a digital forensics conference, holds an authorship identification-related task every year.

3 Problem Description

The Spooky Author Identification task is hosted by Kaggle. It is a supervised learning problem. Training data is given as tuples of (text, label), where 'text' is a one-sentence excerpt from an author's corpus, and 'label' is one of Edgar Allen Poe, H. P. Lovecraft, and Mary Wollstonecraft Shelley.

The official task tracks logless as its only metric:

$$\text{logloss} = -\frac{1}{N}\Sigma_{i=1}^{N}\Sigma_{j=1}^{M}y_{ij}log(p_{ij})$$

where M are the authorship classes and y_{ij} is 1 if the prediction is true and 0 if it is not.

We decided to focus instead on matching-class accuracy:

$$\text{accuracy} = \frac{1}{N}\Sigma_{i=1}^{N}\Sigma_{j=1}^{M}\mathbf{1}y_{ij}$$

Our motivation was 2-fold: 1) in a real-world setting, probabilistic classification may be less useful than binary, 2) accuracy was a metric that all of our models could generate.

It should be noted that academic research tends to use mean-averaged precision as the primary metric. Example: [Bagnall].

4 Data Acquisition & Preparation

This is a managed task, so data acquisition was straightforward. We downloaded the training data in csv format. It is comprised of 19479 rows, including gold label.

According to convention, we split the training data into three sets: 15664 rows of training data 1958 rows of validation data 1857 rows of holdout/test data

The only data wrangling necessary was a check for null values across all columns, which returned negative. We computed simple statistics across each column to check for oddities like empty sentences or out-of-vocabulary labels.

5 Exploration

Poe January 19, 1809 October 7, 1849), boston Shelley 30 August 1797 1 February 1851) London Lovecraft (August 20, 1890 March 15, 1937) rhode island

Here are three examples of sentences from the three authors:

mws: 'How lovely is spring As we looked from Windsor Terrace on the sixteen fertile counties spread beneath, speckled by happy cottages and wealthier towns, all looked as in former years, heart cheering and fair.'

'It never once occurred to me that the fumbling might be a mere mistake.'

'This process, however, afforded me no means of ascertaining the dimensions of my dungeon; as I might make its circuit, and return to the point whence I set out, without being aware of the fact; so perfectly uniform seemed the wall.'

Can you attribute each passage? You probably cannot, unless you are intimately familiar with the books these sentences are drawn from. We needed to look at much more data to find meaninful patterns.

We first truncated the corpus slightly so that all three authors were equally weighted with 5635 rows each. This avoided data skew without meaningfully descreasing corpus size.

We wrote a simple tokenizer for this process, which splits on white space and standard punctuation. We ensured that our tokenizer imitated the behavior of sklearn's internal tokenizer, knowing we would probably use `sklearn.TFIDFVectorizer` later.

Vector Space Model	KMeans	Spectral	Agglomerative	Birch
500vec	0.3662	0.2876	0.3748	0.3649
1000vec	0.3661	0.2739	0.3727	0.3652

Table 1: Paired F-Score on the dev set by different vector space models and clustering algorithms.

6 Feature Engineering

7 Implementation

8 Analysis

8.1 Error Analysis

9 Future Work

sequential, language models, LSTM

10 Appendix

asdf

1 https://ieeexplore.ieee.org/abstract/document/1512002 Abbasi

Bagnall https://www.uni-weimar.de/medien/webis/events/pan-16/pan16-papers-final/pan16-author-identification/bagnall16-notebook.pdf

Brennan: https://www.aaai.org/ocs/index.php/IAAI/IAAI09/paper/viewFile/257/1017

11 Dense Vector Representations

Dense Model	KMeans	Spectral	Agglomerative	Birch
SVD	0.3298	0.3165	0.3569	0.3076

Table 2: Paired F-Score on the dev set by different dense vector space models and clustering algorithms.