Project Proposal: Author Attribution

Remy LeWinter
Joe Song
CS4120: Natural Language Processing Fall 2020
Northeastern University
lewinter.r@husky.neu.edu
song.jo@husky.neu.edu

Introduction

Author attribution is a subfield of natural language processing (NLP) which aims to determine the authors of documents. It is best known for its application to the 12 articles of the Federalist Papers which both Alexander Hamilton and James Madison claimed to have authored [1].

In the digital age, the volume of unattributed text has greatly increased, in part due to social networking platforms and other websites allowing users to create multiple anonymous or pseudonymous accounts. As a result, the applications and exigence for authorship attribution have increased. Modern applications include:

- Detecting plagiarism and intellectual property infringement [1]
- Attributing authorship to anonymously published works of interest (such as the Federalist Papers) [1]
- Linking messages in criminal law, such as verifying suicide notes, or identifying writers of harassment or terroristic proclamations [1]
- Determining code authorship within a code base [1]
- Author profiling [1]
- Identifying spam [2]

Due to developments in consumer technology, subjects of analysis in author attribution tasks have largely changed from articles and full books to smaller documents, such as tweets, emails, and blog posts. This resulted in changes in algorithm development and study design. This project will investigate the effectiveness of previously proven algorithms on modern data.

Related Work

According to Luyckx and Daelemans, "authorship attribution accuracy deteriorates as the number of candidate authors increases and size of training data decreases." [3, p. 35]. As a result, author attribution algorithms have changed to prioritize performance on small documents with large author sets. For example, Koppel and Winter present the impostor method, which operates by generating imposter documents along with the test set, comparing pairs of documents, and scoring them by similarity [4]. Luyckx and Daelemans note that some older algorithms, such as the character-level n-gram model Keselj et al. implemented [2], are robust to changes in author set size and data size [3].

Datasets

We will use the Blog Authorship Corpus dataset consisting of 681,288 posts (over 140 million words) by 19,320 individual bloggers between the ages of 13 and 47 [5]. The dataset is distributed as a set of XML files such that each file contains the complete profile of an individual blogger. The mean author profile contains 35 posts and 7250 words [5]. We will filter the dataset to remove authors with fewer than 3 posts.

Our dataset consists of posts that are generally considerably longer than Twitter posts, yet much shorter than an entire novel, and contains text from a large number of authors. This sits nicely in the middle of the datasets commonly used until around the turn of the century, with few authors and large documents, and the microblogging datasets currently used as a model organism, with small documents and many authors.

Methodology

We are beginning with the intention of reproducing the supervised algorithm used by Keselj et al. in 2003 and originally proposed by William Bennett in 1976 [2]. The algorithm uses character-level n-gram models to build author profiles and assign new input text to the most likely author. Though we currently plan to follow Keselj et al.'s methodology, we will continue to explore publications following the 2003 paper (such as Koppel and Winter [4]) for potential improvements that would be relevant and feasible with respect to our course work.

Evaluation

A selection of blog posts from different authors will be pulled out as the testing set. The effectiveness of the character-level n-gram model will be scored by whether or not the model assigns the highest probability to the actual author of the blog post. We will continue to update our evaluation techniques as we learn more in class and explore further publications.

References

[1] E. Stamatatos. 2009. A survey of modern authorship attribution methods. *JASIST* 60:538-556.

[2] V. Keselj, F. Peng, N. Cercone, and C. Thomas. 2003. N-gram-based author profiles for authorship attribution. In *PACLING'03* 255-264 Halifax, Canada.
[3] K. Luyckx and W. Daelemans. 2011. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing* 26(1):35-55.
[4] M. Koppel and Y. Winter. 2014. Determining if two documents are by the same author. *JASIST* 65(1):178–87.
[5] J. Schler, M. Koppel, S. Argamon and J. Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.*