Authorship Identification: Classification of the Speeches of US Presidents

Final Project Proposal
W266 Natural Language Processing

Thong Bui, Jason Vantomme

Overview

Nirkhi and Dharaska define authorship identification (also known as authorship attribution) as a process that “determines the likelihood of a piece of writing to be produced by a particular author by examining other writings by that author.”[3]

What is this an important problem? Attribution of authorship has been a topic of inquiry since the late 1800’s[8] and, relevant to our context, has been applied to the decisions of the US Supreme Court~~5~~ as well as the pre-presidential radio addresses of Ronald Reagan[1]. A common focus area of authorship identification in a present-day research is its use in criminal forensics (e.g., tracing of anonymous online criminal activities). Furthermore, it is easy to understand in today’s era of “fake news” how the verification of statements, policies or quotes emerges as a critical activity.

Implementation

Bagnall[2] details an approach to the authorship attribution task at the CLEF 15 conference that, while simple in notion, outperformed all other approaches in 3 of 4 languages tested. (Bagnall[3] continues with this approach in its application to author clustering at the CLEF 16 conference.)

In this project, we will be considering authorship identification as synonymous with speaker identification where the speakers being identified are US Presidents. We will be implementing Bagnall's approach and applying it to the speeches of Presidents of the United States with the goal of implementing a classifier that will be able to identify a test speech as one of the presidents represented in the training set. These speeches exhibit a range of variance in style and characteristics in their choice of words and though often on differing topics over time, they are easily considered of the same "genre". These speeches are also generally of a standard format and are readily available.

Pre-processing methods noted in this method are not substantive and the author notes that "The character mappings were settled before training started and no attempts were made to test their efficacy" ; as a result, preprocessing methods to improve the performance of the approach will also be explored such as that noted in Stamatatos [4].

A key challenge in this project will be determining and adjusting for the impact of *topic* on classification probabilities in what is (in effect) ultimately a *style* classification problem.

Data Sources

Texts used in this process will be limited to prepared speeches and will not include largely unscripted communications such as the Q&A session of a news conference or a political debate. Limiting texts in the way is an important as presidents generally have a consistent, distinct "voice" in prepared speeches even if when they are written for them.

Primary sources for speech text from Presidents Hoover through Trump will come from The American Presidency Project @ UC Santa Barbara (https://goo.gl/rCJ8Gv). This source is not only easy to parse, but is also comprehensive and kept up-to-date which will be important to ensure we capture as many speeches from President Trump as possible before the project end. Additional sources will also be used to capture speech transcripts for President Trump in order to ensure a volume of text similar to other presidents.

References

1. Airoldi, E., S. Fienberg, and K. Skinner. “Whose Ideas? Whose Words? Authorship of Ronald Reagan’s Radio Addresses.” PS: Political Science & Politics 40, no. 3 (n.d.): 501–6. doi:10.1017/S1049096507070874.
2. Bagnall, Douglas. “Author Identification Using Multi-Headed Recurrent Neural Network.” In CLEF 2015 Labs and Workshops, Notebook Papers; CEUR Workshop Proceedings. September 8-11, 2015, 2015. http://ceur-ws.org/Vol-1391/150-CR.pdf.
3. ———. “Authorship Clustering Using Multi-Headed Recurrent Neural Networks.” In CLEF 2016 Labs and Workshops, Notebook Papers; CEUR Workshop Proceedings, 791–804. Évora, Portugal, 2016. http://ceur-ws.org/Vol-1609/16090791.pdf.
4. Nirkhi, Smita, and R. V. Dharaskar. “Comparative Study of Authorship Identification Techniques for Cyber Forensics Analysis.” arXiv:1401.6118 [cs], December 24, 2013. http://arxiv.org/abs/1401.6118.
5. Rosenthal, Jeffrey, and Albert H Yoon. “Detecting Multiple Authorship Of United States Supreme Court Legal Decisions Using Function Words.” The Annals of Applied Statistics 5, no. 1 (2011): 283–308. doi:10.1214/10-AOAS378.
6. Stamatatos, Efstathios. “Authorship Attribution Using Text Distortion.” In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 1: Long Papers:1138–49. Valencia, Spain: Association for Computational Linguistics, 2017. https://www.aclweb.org/anthology/E/E17/E17-1107.pdf.
7. Stamatatos, Efstathios, Walter Daelemans, Ben Verhoeven, Patrick Juola, Aurelio López-López, Martin Potthast, and Benno Stein. “Overview of the Author Identification Task at PAN 2015.” In CLEF 2015 Labs and Workshops, Notebook Papers; CEUR Workshop Proceedings. Toulouse, France, 2015. http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-papers-final/pan15-authorship-verification/stamatatos15-overview.pdf.
8. “Stylometry.” Wikipedia, April 27, 2017. https://en.wikipedia.org/w/index.php?title=Stylometry&oldid=777553870.