## ***Speaker Identification: Classifying the Speeches of US Presidents from Hoover to Trump***

## **Final Project Proposal W266 Natural Language Processing**

Thong Bui, Jason Vantomme

**Overview**

What is it? Why is it important? Why is it challenging?

Fraud, forensics. Reference something here from the CLEF overview.

**What is our application?**

Methodology. Implementing published approach with new dataset. Bagnall.(1); interesting approach, very successful at the CLEF 15 task; has been applied to clustering (2).

We have chosen the speeches of United States Presidents for our speaker classification problem. These speakers exhibit a range of variance in style and characteristics in their choice of words. <…> These texts have the advantage of being in the same topic space, of the same general format and are readily available.

Challenges, how measure success. (confusion of topic and style…source size differences between subjects). Accuracy, but also interested in confusion.

**Data Sources**

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.

**Implementation**

<Bagnall> (1)

Possibility of additional methods to apply – in preprocessing for example; (3)

**References**

1. 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.
2. ———. “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.
3. 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.
4. 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.
5. 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.