Written style analysis using Part-of-speech (POS) tags

Jingze Zhang, Gordon Erlebacher
Department of Scientific Computing, Florida State University

Abstract

Stylometry studies written language style. Generally, writing style is only determined by the choice of words, sentence and paragraph structure, regardless of semantic meaning. Traditional methods operate on the frequency of function words to represent the style. However, function words, such as articles and pronouns, comprise only a small component of style. Part-of-speech (POS) vectors, which classify each word into one of 43 categories, define an alternate vocabulary, which will allow an elementary characterization of style. Starting from a collection of 13,000 emails, we first construct a frequency table of POS elements, to which we apply an SVD. Each text is represented by a vector, where each entry represents the POS frequency. We apply POS analysis to translate each entity (word, punctuation) into a POS element. We then construct a POS frequency table across the entire mail corpus and display the coefficients associated with the three dominant singular values for each email. We find that different people use POS elements differently, thus have different styles. In future work, we plan to measure written deception through stylistic change.

Background

Established in 1998, Tallahassee Community Redevelopment Agency (CRA) aims to redevelop and enhance selected areas in the central urban district in Tallahassee. In 2017, the FBI initiated an investigation into CRA due to suspicion of corruption. To identify evidence in emails, we analyze the entire collection of 13,000 email communications between 2012 and 2017 among the involved people and try to unveil the truth with the help of stylometry.

Deception detection

Deception is defined as the act to intentionally cause to have a false belief that is known or believed to be false, while corruption is the abuse of power by a public official for private gain. Although corruption and deception are two different concepts, it's believed that corruption typically involves deception[1]. So the findings of deception detection can also work for corruption detection. Detecting deception relies on several different observable cues, such as sweating and stammer. When working with computer-mediated communication (CMC), such as emails, the writing style serves to identify deceptive practices[2]. We hypothesize that the deception detection can help with corruption detection, since corruption typically involves deceiving either a person or the government.

Stylometry

Stylometry studies written language style. Generally, writing style is only determined by the choice of words, sentence and paragraph structure, regardless of semantic meaning. Traditional methods operate on the frequency of function words[3], such as articles and pronouns, to represent the style. However, the function words comprise only a small component of the word choice. The choice of sentence and paragraph structure that also lead to a different writing style are not considered.

Part of Speech (POS)

Part of Speech categorizes words by their grammatical properties. A well-acknowledged approach to categorize English POS are as follows: noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection and article/determiner. Some computer software have more detailed subcategories. The natural Language Toolkit (NLTK)[4], the Python package we are using, has 43 POS, including punctuations. The different frequencies of POS in a sentence is an important representation of the sentence/paragraph level style information.

Methods and visualizations

We apply a bag of word model, which discards any effect of word ordering. We first process all emails through the NLTK POS tagger and convert each of them to a list of POS elements. Frequency tables are built based on the lists. Now each email can be represented by a vector of length 43. All vectors can construct the frequency matrix that has a shape of 43 by 13,000, to which we apply a principal component analysis (PCA). The singular values of the PCA are plotted in Fig. 1.

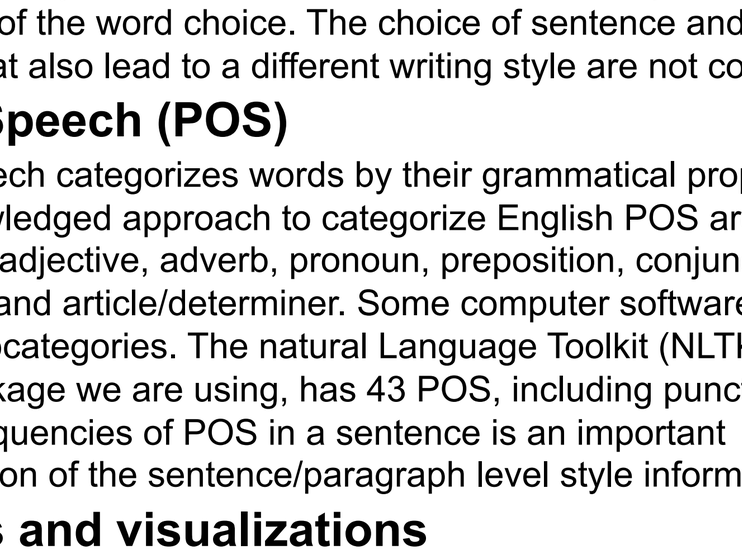

Fig. 1. The curve have a turning point at 5. The first 5 components cover the most information

To visualize the POS vectors, we project them onto the three dominant directions. Emails from four different individuals are selected and scattered into the vector space. Fig. 2 indicates that emails from the same author tend to cluster together. The same color points can also be represented by a confidence ellipsoid where the center is the mean and the size is determined by the covariance, as shown in Fig. 3.

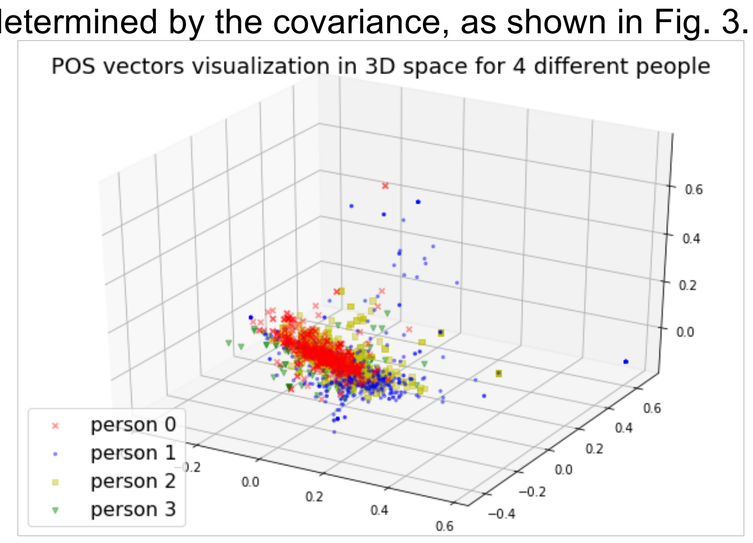

Fig. 2. The high dimensional POS vectors are projected into 3D space. Each point corresponds to a single email, while color denotes the email author.

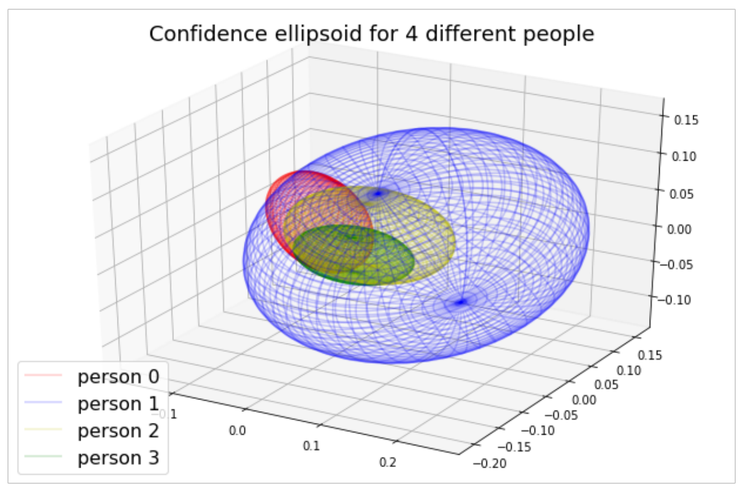

Fig. 3. Confidence ellipsoids define a one-sigma region about the mean for a given author. A larger ellipsoid indicates that the author changes writing styles frequently. The different locations of ellipsoid suggest that different people have different writing styles.

Conclusion/Future work

The POS vectors clearly indicates that different authors use POS elements differently, which corresponds to different writing styles. In the future research, the word order will be used. The POS is a good identifier for the sentence structure style. It is not yet clear how to capture style characteristics related to word choice and paragraph structure. The target of our research is to develop an algorithm that combines all three styles of information. Some modern machine learning methods, such as neural networks, might prove useful.

Reference

[1] Nodelman, U., Allen, C., & Anderson, R. L. (1995). Stanford encyclopedia of philosophy.

[2] Afroz, S., Brennan, M., & Greenstadt, R. (2012, May). Detecting hoaxes, frauds, and deception in writing style online. In 2012 IEEE Symposium on Security and Privacy (pp. 461-475). IEEE.

[3] Feng, S., Banerjee, R., & Choi, Y. (2012, July). Syntactic stylometry for deception detection. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2 (pp. 171-175). Association for Computational Linguistics.

[4] Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. arXiv preprint cs/0205028.