

Weekly Report 2

Jeremy Lachowicz

1/29/2022

1 Accomplished this week

This week I learned more about Stylometry with machine learning. I re-read the articles from last week and read two new articles. At this point, I have a solid understanding of the process of using machine learning with Stylometry and the common problems faced when doing so. I have also set up a meeting with a professor in the religion department and have prepared questions for her to make sure my information is correct. Finally, I learned some language processing, and found a processing library in Python that supports the Ancient Greek language. I tested this library on sample data, and learned about the available processing tools.

2 Accomplish next week

I will have my meeting to make sure I can proceed with using the sources I have collected. After that meeting, I plan to continue learning language processing techniques and applying that to my data to be used for Stylometry.

References

1. Ramyaa, C.H., Rasheed, K., He, C.: Using machine learning techniques for stylometry (2004)

The problem discussed in this paper was primarily features used to train machine learning models for Stylometry. To solve this, they presented 21 features used to determine authorship. Some of these include mean sentence length, number of quotation marks, mean word length, etc. They also presented the idea of using winnowing to find the most important features, and get rid of the unhelpful ones. They found a neural network was the most accurate overall. For future work, they suggest using a meta-learner to train decision trees and neural networks as well as finding a way to generalize features.

2. Talati, A., EDU, U., Narayanan, R.: Deep learning based authorship identification (2017)

The problem discussed in this paper was using deep learning to improve authorship identification. To solve this, they used two deep learning models and found the Article-level GRU model performed best on authorship identification. For future work, they suggest changing features and using other models such as MvRNN or CNN, or use an ensemble model.