Weekly Report 1

Jeremy Lachowicz

1/23/2022

1 Accomplished this week

This week, I focused on learning about machine learning applications with Stylometry, especially for authorship attribution. I read papers on other examples of authorship attribution using these methods, one was about source code plagiarism detection and the other was about text plagiarism detection. These papers helped me understand the workflow of creating a machine learning model with large amounts of text as data. An interesting take away from these papers was that having a large data set is important for authorship attribution, but most prior applications have used an unrealistically large amount of text. In most real world applications, there is not that much text. So it was good to hear it is possible to use smaller data sets and still have significant results, which is important for my project as some of the candidates have limited data.

2 Accomplish next week

I will read a couple more papers about machine learning with Stylometry (referenced below), and I will start data collection and preparation.

References

1. Bogomolov, E., Kovalenko, V., Rebryk, Y., Bacchelli, A., Bryksin, T.: Authorship attribution of source code: A language-agnostic approach and applicability in software engineering. Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (2021). https://doi.org/10.1145/3468264.3468606

   The problem discussed in this paper was plagiarism with source code. Their first step was to develop a language agnostic authorship attribution solution for which they used neural networks and random forests. They random forest models worked better for smaller data sets, while the neural network performed better on larger data sets. Another problem was finding code data for their model so they created a way to use GitHub to pull code from. This paper did not have any directions for future work by the authors.

2. Luyckx, K., Daelemans, W.: Authorship attribution and verification with many authors and limited data. Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08 (2008). https://doi.org/10.3115/1599081.1599146

The problem discussed in this paper was how to use machine learning with Stylometry with limited data and more than two authors. To solve this, they used an implementation of SVM and an implementation of Maximum Entropy Learning. Overall, they found memory-based learning to be the best when dealing with limited data. For future research they recommend trying different machine learning algorithms and a deeper investigation into feature selection and other optimization issues.

3. Ramnial, H., Panchoo, S., Pudaruth, S.: Authorship attribution using stylometry and machine learning techniques. Advances in Intelligent Systems and Computing p. 113–125 (2015). https://doi.org/10.1007/978-3-319-23036-8$_1$0

4. The problem discussed in this paper was plagiarism with text. To solve this they used k-NN and SMO algorithms to do authorship attribution. They also created features that have not been used before in stylomety machine learning. Overall, they found writing styles of authors to be sufficiently dissimilar, and their algorithms were able to detect these differences with a high accuracy. For future work the authors suggest using a smaller amount of words per data point and with more authors.

Ramyaa, C.H., Rasheed, K., He, C.: Using machine learning techniques for stylometry (2004)
Talati, A., EDU, U., Narayanan, R.: Deep learning based authorship identification (2017)