Weekly Report 4

Jeremy Lachowicz

2/11/2022

1 Accomplished this week

This week I finished my proposal and figured out which hyperparamters and feature sets I will use for my Word2Vec and Doc2Vec models. The papers I read this week gave me a good idea of the process of using Word2Vec and Doc2Vec for authorship attribution as well as what hyperparamters and feature sets worked for them. It was interesting to note there is conflicting results for hyperparameters and features among different papers. It seems to be largely language-dependent, so I will be sure to use their results as a starting point, while also making sure I incorporate other types to see what works the best for Ancient Greek. Finally, I downloaded the required Word2Vec and Doc2Vec packages for python.

2 Accomplish next week

Start coding the Word2Vec process for the first set of Greek sources.

References

1. Gómez-Adorno, H., Posadas-Durán, J.P., Sidorov, G., Pinto, D.: Document embeddings learned on various types of n-grams for cross-topic authorship attribution. Computing **100**(7), 741–756 (Jul 2018). https://doi.org/10.1007/s00607-018-0587-8, http://link.springer.com/10.1007/s00607-018-0587-8
   In this paper, they wanted to figure out which Doc2Vec hyperparamters and feature types are best for Doc2Vec authorship attribution. They had a corpus consisting of about 30 articles per author from 1999 to 2009 from different topics in the The Guardian newspaper. The average document contained about 1,000 words. For feature sets they used n-grams (ranging from 1-5), POS tags, and words, with a logistic regression classifier. They found using a combination of n-grams as a feature set was the best.
2. Rahgouy, M., Babaei Giglou, H., Rahgooy, T., Karami, M., Mohammadzadeh, E.: Cross-domain authorship attribution: Author identification using a multi-aspect ensemble approach notebook for pan at clef 2019 (07 2020)
   In this paper, they wanted to compare different models for authorship attribution and use an ensemble to see if that creates better results. They used a corpus of English, Italian, Spanish, and French texts, but they did not state the length or number of documents. They tested Word2Vec and Doc2Vec and found Word2Vec was better overall, as well using a logisitc regressor for their ensemble classifier. Then, the feature sets they used were N-grams, Word2Vec, and TF-IDF. They found Word2Vec to perform the best out of the three, but when combined in an ensemble, it sometimes performed even better.