Mid Term Report

Magdaleno, Alejandro
amagale@student.kennesaw.edu

Lyles-Woods, Quin'darius
qlyleswo@student.kennesaw.edu

October 17, 2021

Progress

The goal of this research is to build a classification model to determine where bodies of text originate from using a CNN [1] deep learning model to take advantage of non-linearity. The progress that we have had so far is multifold. For the data, we have been able to write a script that allows us to get 40-60 articles at a time and strip them of only their stories and outputting that data into a folder properly labeled. So the current article count is around 120 and the token count is around $124,998$ and will grow at the pace new articles are produced. We have a pipeline for creating a vocabulary and frequency count for the current data set. With this pipeline that directly helps us Vectorize our data and then use that information for our sequential model. Our loss values with this model are .58 and the Accuracy that the model is achieving now with this limited data is .72.

Methods

Results

Challenges

So far the biggest challenge has been the dataset. Collection and cleansing of the data has been a difficult thing to do. The main issue with this is that there were no pre-existing data set that had the data in the format that we needed to build such a classification model and with our want to have relevant and up to date articles. We still need some more labels when pulling the data to reach some of the other auxiliary goals that were mentioned in the proposal. So getting the author name the title of the article and such would allow use to gather more information from our data set. Creating a working model to our standards has not posed a great challenge, currently we are getting 72% with very limited data and we expect this to increase with more data. Going forward we can see the finish line and we will simply polish the processes if we have more time left over.

[1] Convolutional Neural Network