April 7, 2022

Pdf Processing and Optical Character Recognition (OCR)

Caleb Smotherman

The problem that I chose to address was related to OCR and image processing for PDF. The files that I wanted to focus on are related to finance and accounting in the military, so the real-world application of such a process would be very helpful in automating workflows connected with these documents. The files that I was provided were stripped of personal information, and I focus solely on the table data of the document.

Based on the processing techniques I chose to use (eroding, diluting, threshing, and resizing), I was able to consistently locate the table data that I need within the document and further break the table into a data frame like object of bounding boxes. Those boxes were then iterated over and the text extracted from them was added to a Pandas data frame.

As for the modeling, I worked with off-the-shelf, pre-trained models and found that pyocr and easyocr were fantastic. They each have their strengths and weakness; pyocr being the faster of the two, while easyocr was very, very accurate with little pre-processing required. In future versions of these projects, I will continue to focus more on the pre-processing and document preparation for the OCR models, because that is where the majority of the performance of the model is impacted.

Consideration for future processing; I'd like to create a model that checks the structure of a document to classify what type of form it is. I will have a lot of different finance documents that will each have their own structures, so knowing how to verify the document type would help me process them more accurately.