PDF Classifier

By: Matthew Wong

Motivation:

It requires a lot of time to sort through hundreds of PDFs and file them accordingly. This costs businesses both time and money. What if we could create a model using machine learning to help classify these PDFs, so employees could spend less time on sorting these PDFs and more time on other tasks integral to the business?

Text Extraction:

- One of the biggest issues was getting the text from these PDFs so that they could be analyzed and run through various NLP models.
- Used two different methods for text extraction:
  - PDFminer
  - PyOCR (python wrapper for Tesseract)

Differences:

PDFminer could not extract text from PDFs that were scans or pictures, so I had to use an OCR (optical character recognition) to get text from the remaining documents.

The OCR did a good job at getting text from all the other documents, but it did take a lot longer to process.

Different Models:

Naive Bayes	Cosine Similarity - Averages	Cosine Similarity - kNN
My first model assumes words or features are independent of one another, and classifies documents depending on the probability of those words or features appearing in specific categories of documents to determine the classification.	My second model compared the cosine similarity of a document to the average cosine similarity of all the documents in each category to determine its classification.	My third model compares the cosine similarity of a document to each other known document and takes the majority vote of the most similar n documents to determine its classification.

Validation:

Each category varied in the number of documents that were available, so it was important to stratify the data when splitting into the training and test sets.

Naive Bayes Accuracy: ~0.62

Average Cosine Similarity Accuracy: ~0.75

kNN Cosine Similarity Accuracy: ~0.82

Future Work:

- Extracting specific text strings like name, address, and values so employees would not have to go through these documents for them.
- Continue working on model accuracy for document types it has the most trouble with.

Thank You

Questions?