PDF Classifier

By: Matthew Wong

Motivation

- Businesses get hundreds of PDFs
- PDFs are sorted by employees
- Takes a lot of time and costs money
- Goal: Classify PDFs

Data

- 14 different classifications
  - Appraisal
  - Income
  - Insurance
- ~700 documents and uneven classes
- PDFs in each classification vary
  - Forms
  - Scans
  - Faxes

Text Extraction

- PDFminer
  - Works well for PDFs with text layers
  - Much quicker than PyOCR
- PyOCR (python wrapper for Tesseract)
  - Works well for most PDFs
  - Takes a long time

Different Models

Naive Bayes

Simpler NLP model

Worked well for some categories and not so great with others

TF-IDF

Average Cosine Similarity

Tried to capture signal from variation in documents of the same category

TF-IDF

kNN Cosine Similarity

Only looked at the n most similar documents

Cross Validated Accuracy

- Used Stratified Cross Validation due to uneven classes

Naive Bayes	Average Cosine Similarity	kNN Cosine Similarity
PDFminer:	PDFminer:	PDFminer:
~ 0.58	~ 0.67	~ 0.73
PDFminer and PyOCR:	PDFminer and PyOCR:	PDFminer and PyOCR:
~0.62	~0.75	~0.84
PyOCR:	PyOCR:	PyOCR:
~0.48	~0.67	~0.83

Future Work

- Renaming/Filing documents automatically
- Extracting specific text strings
  - Name
  - Address
  - $$$ Values

Thank You
Questions?

Matthew Wong
mjwong1991@gmail.com
https://www.linkedin.com/in/matthew-j-wong/
https://github.com/Unrelenting