OCR Project for Hackathon

Manjit Trehan (mtrehan@iu.edu)

Problem statement: Some legal firms have combined PDF documents that contain multiple sub-documents, e.g. Appearance, Complaint, Summons, Notice, Affidavit, etc. So, the first three pages of a PDF might be the Complaint; fourth page might be the Appearance; fifth page might be the summons, and so on. When these documents are submitted to the courts, they need to be separate PDFs. i.e. Complaint.pdf and Summons.pdf, Appearance.pdf, etc. Some large Law Firms have hundreds or maybe thousands of combined PDFs that need to be separated into individual PDFs. So, with OCR, we can simplify the process. The first page of each sub-document indicates the type of sub-document. That is, page 1 would say COMPLAINT, and pages 2 and 3 would have no identifying text; page 4 would say SUMMONS, etc. The combined PDF would be named with the defendant ID, e.g. 123456.pdf and as each sub-document is separated, 123456_Complaint.pdf, 123456_Summons.pdf, 123456_Appearance.pdf could be extracted.

Project: Develop a standalone executable tool that uses a configuration file to determine how to split the combined PDFs.

Sample.cfg

```
[Files]
INPUT=D:\Firm1\Cases\
OUTPUT=D:\Firm1\Out\

[OCR]
Text1=”COMPLAINT”
Loc1=200,100,500,400
Text2=”SUMMONS”
Loc2=200,100,500,400
Text3=”APPEARANCE”
Loc3=200,100,500,400
```

The OCR section of the settings file shows each string to look for in the rectangular area of a page, where Location is 4 numbers set:

top-left-x, top-left-y, bottom-right-x, bottom-right-y

If the text COMPLAINT is inside the rectangle 1 coordinates, then it begins the complaint until Text2, or Text3 appears on a subsequent page.

So, a PDF named 123456.pdf with pages:

Complaint Page 1
Complaint Page 2
Complaint Page 3
Summons Page 1
Appearance Page 1

Would result in 3 separate PDFs:

123456_Complaint.PDF

- Complaint Page 1
- Complaint Page 2
- Complaint Page 3

123456_Summons.pdf

- Summons Page 1

123456_Appearance.pdf

- Appearance Page 1

Please try to build the foundation of this project as a hackathon project, and there may be an opportunity to work further on it via an internship.