Capstone Final Report:

Pdf Processing for Optical Character Recognition (OCR)

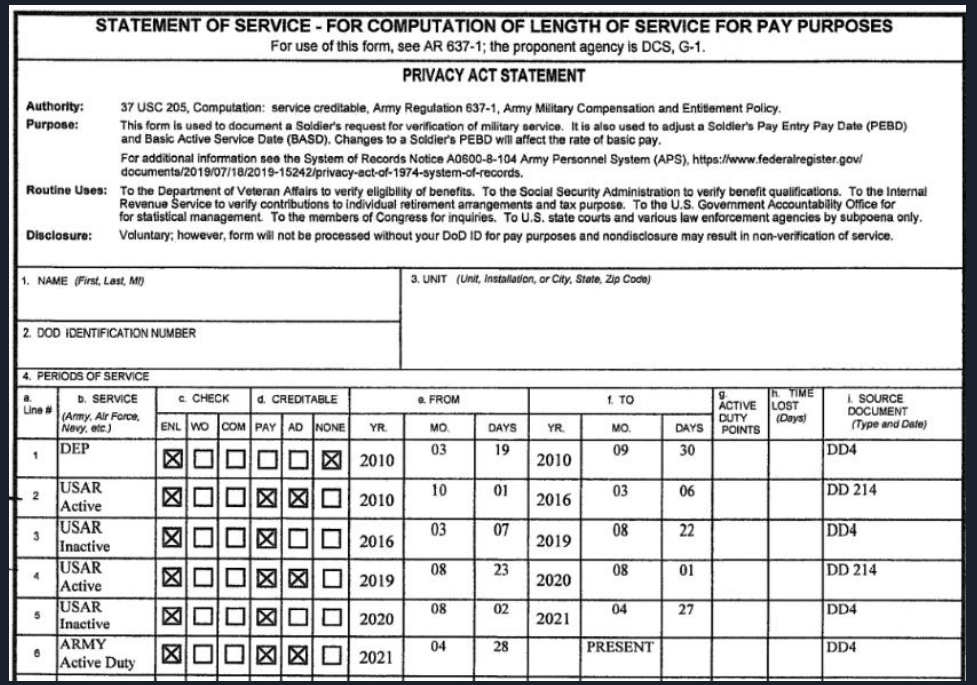
Caleb Smotherman

Contents

Problem Statement

I would like to process finance related Pdfs to extract essential data for further CRUD operations and “math-checks.” The form will be practicing with is a DA 1506. The most important part of the document is the Block 4.

Once I can process the document and extract the information, the potential for the utility of this model is limitless and the pre-processing techniques can scale to many other forms.



STATEMENT OF SERVICE - FOR COMPUTATION OF LENGTH OF SERVICE FOR PAY PURPOSES

For use of this form, see AR 637-1; the proponent agency is DCS, G-1.

PRIVACY ACT STATEMENT

Authority: 37 USC 205, Computation: service creditable, Army Regulation 637-1, Army Military Compensation and Entitlement Policy.

Purpose: This form is used to document a Soldier's request for verification of military service. It is also used to adjust a Soldier's Pay Entry Pay Date (PEBD) and Basic Active Service Date (BASD). Changes to a Soldier's PEBD will affect the rate of basic pay.

For additional information see the System of Records Notice A0600-8-104 Army Personnel System (APS), https://www.federalregister.gov/documents/2019/07/18/2019-15242/privacy-act-of-1974-system-of-records.

Routine Uses: To the Department of Veteran Affairs to verify eligibility of benefits. To the Social Security Administration to verify benefit qualifications. To the Internal Revenue Service to verify contributions to individual retirement arrangements and tax purpose. To the U.S. Government Accountability Office for for statistical management. To the members of Congress for inquiries. To U.S. state courts and various law enforcement agencies by subpoena only.

Disclosure: Voluntary; however, form will not be processed without your DoD ID for pay purposes and nondisclosure may result in non-verification of service.

1. NAME *(First, Last, MI)*

2. DOD IDENTIFICATION NUMBER

3. UNIT *(Unit, Installation, or City, State, Zip Code)*

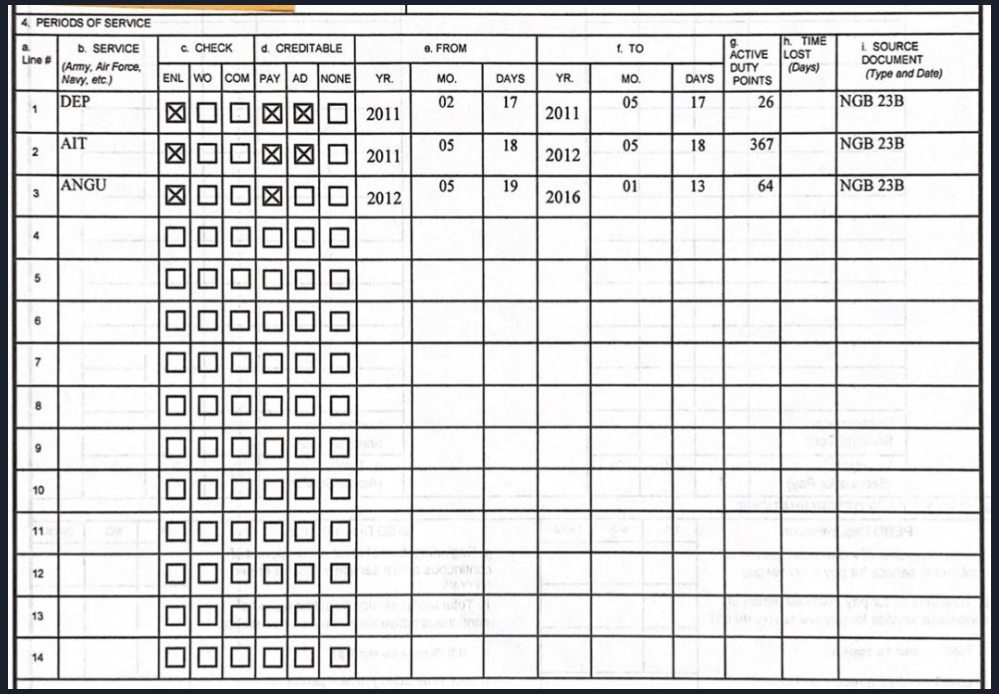
4. PERIODS OF SERVICE

a. Line #	b. SERVICE (Army, Air Force, Navy, etc.)	c. CHECK ENL	c. CHECK WO	c. CHECK COM	d. CREDITABLE PAY	d. CREDITABLE AD	d. CREDITABLE NONE	e. FROM YR.	e. FROM MO.	e. FROM DAYS	f. TO YR.	f. TO MO.	f. TO DAYS	g. ACTIVE DUTY POINTS	h. TIME LOST (Days)	i. SOURCE DOCUMENT (Type and Date)
1	DEP	☒	☐	☐	☐	☐	☒	2010	03	19	2010	09	30			DD4
2	USAR Active	☒	☐	☐	☒	☒	☐	2010	10	01	2016	03	06			DD 214
3	USAR Inactive	☒	☐	☐	☒	☐	☐	2016	03	07	2019	08	22			DD4
4	USAR Active	☒	☐	☐	☒	☒	☐	2019	08	23	2020	08	01			DD 214
5	USAR Inactive	☒	☐	☐	☒	☐	☐	2020	08	02	2021	04	27			DD4
6	ARMY Active Duty	☒	☐	☐	☒	☒	☐	2021	04	28		PRESENT				DD4

Data

The “data” for this project are Pdf files. I do not have a lot of working copies of these files, but the two files I do have should be enough to get my testing started.

This information must be extracted from the document with 100% accuracy, else the option to manually update the information after extraction will be built into the user interface.



4. PERIODS OF SERVICE

a. Line #	b. SERVICE (Army, Air Force, Navy, etc.)	c. CHECK			d. CREDITABLE			e. FROM			f. TO			g. ACTIVE DUTY POINTS	h. TIME LOST (Days)	i. SOURCE DOCUMENT (Type and Date)
		ENL	WO	COM	PAY	AD	NONE	YR.	MO.	DAYS	YR.	MO.	DAYS			
1	DEP	☒	☐	☐	☒	☒	☐	2011	02	17	2011	05	17	26		NGB 23B
2	AIT	☒	☐	☐	☒	☒	☐	2011	05	18	2012	05	18	367		NGB 23B
3	ANGU	☒	☐	☐	☒	☐	☐	2012	05	19	2016	01	13	64		NGB 23B
4		☐	☐	☐	☐	☐	☐									
5		☐	☐	☐	☐	☐	☐									
6		☐	☐	☐	☐	☐	☐									
7		☐	☐	☐	☐	☐	☐									
8		☐	☐	☐	☐	☐	☐									
9		☐	☐	☐	☐	☐	☐									
10		☐	☐	☐	☐	☐	☐									
11		☐	☐	☐	☐	☐	☐									
12		☐	☐	☐	☐	☐	☐									
13		☐	☐	☐	☐	☐	☐									
14		☐	☐	☐	☐	☐	☐									

Analysis

Being that the data was in Pdf, I decided to use fitz for creating .png files out of the Pdf pages. Once the images were created, it was a matter of processing the images into a more readable format for the OCR models. This included threshing, resizing, diluting, eroding, and creating high contrast and brightness for each section of data.

Additionally, I needed to extract the information as a table. I found a great resource on GitHub to do so, allowing me to handle each cell of the table as its own image and push the extracted text into a Pandas dataframe.

Individual cells

Modeling

The off-the-shelf models I chose to test are pytesseract, pyocr, and easyocr. Each had its strengths, but pyocr takes the crown. At first, I was most happy with the easyocr model, as it was able to accurately detect the text with little pre-processing on my end. This was largely impart to its organic pre-processing and text identification (CRAFT). However, it more than 5 times slower than the pyocr model. Once I had pre-processed my text appropriately, the pyocr model has been wildly successful, accomplishing complete extraction in as little as 5 seconds.

CRAFT: Character-Region Awareness For Text detection

Official Pytorch implementation of CRAFT text detector | Paper | Pretrained Model | Supplementary

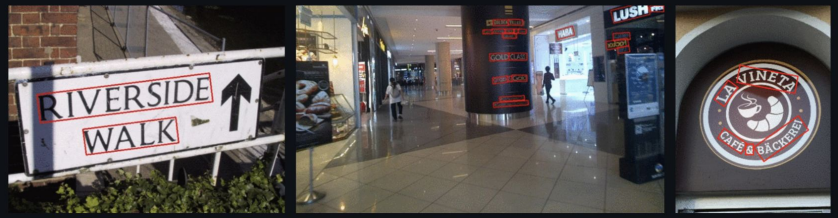
Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, Hwalsuk Lee.

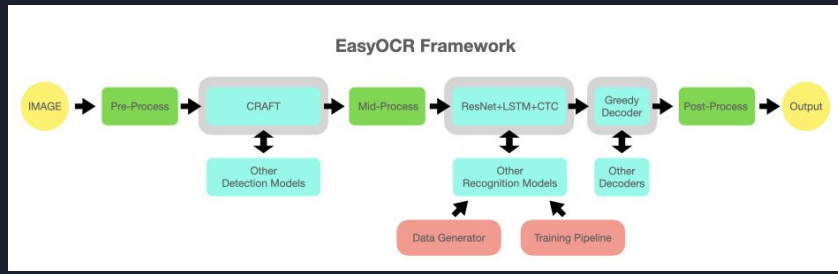
Clova AI Research, NAVER Corp.

Sample Results

Overview

PyTorch implementation for CRAFT text detector that effectively detect text area by exploring each character region and affinity between characters. The bounding box of texts are obtained by simply finding minimum bounding rectangles on binary map after thresholding character region and affinity scores.





Conclusion

- The image pre-processing skills that I learned in this project are so valuable and will working beautifully with my appreciation for text recognition and text classification.
- The models that I learned about helped me to appreciate the Off-the-shelf, pretrained models for what they are; easy to use, accurate enough, and free.
- I will be working on future documents with pyocr and easyocr, so knowing that these are accessible and well in Streamlit is nice.