Computer Vision for Scraping non-searchable PDFs

Zack Boyce
October 6, 2018

Introduction

boyce.zack@gmail.com
www.linkedin.com/in/zack-boyce

Employment:

Markel Corporation

- Data Scientist / Jan. ‘17 - Present
- Business Intelligence Analyst / Oct. ‘13 - Dec. ‘17

Education:

- Masters in Decision Analytics / ‘16 - ‘18
- Bachelors in Economics and Business / ‘08 - ‘12

Unstructured information might account for more than 70%–80% of all data in organizations.

How is it defined?

Information that either does not have a pre-defined data model or is not organized in a pre-defined manner.

What does it look like?

- Pictures
- Emails
- PDFs
- Videos
- etc.

Why it's important at Markel?

30k submissions = 650k pages

650k pages = minimal data captured

Minimal data captured = tough to build models for underwriters.

So how do we get the data from forms like this???

AGENCY CUSTOMER ID: _____

ACORD®

COMMERCIAL GENERAL LIABILITY SECTION

DATE (MM/DD/YYYY)

AGENCY	CARRIER	NAIC CODE
Brown and Brown	Markel Corporation	
POLICY NUMBER / EFFECTIVE DATE	APPLICANT / FIRST NAMED INSURED	
	Middleton Auto	

IMPORTANT - If CLAIMS MADE is checked in the COVERAGE / LIMITS section below, this is an application for a claims-made policy. Read all provisions of the policy carefully.

COVERAGES

- [X] COMMERCIAL GENERAL LIABILITY
  - [ ] CLAIMS MADE [X] OCCURRENCE
  - [ ] OWNER'S & CONTRACTOR'S PROTECTIVE

DEDUCTIBLES
- [ ] PROPERTY DAMAGE $
- [ ] BODILY INJURY $ [ ] PER CLAIM [ ] PER OCCURRENCE
- $

LIMITS

		PREMIUMS
GENERAL AGGREGATE	$	
LIMIT APPLIES PER: [ ] POLICY [ ] LOCATION [ ] PROJECT [ ] OTHER:		PREMISES/OPERATIONS: 27,500
PRODUCTS & COMPLETED OPERATIONS AGGREGATE	$ 2,000,000	PRODUCTS:
PERSONAL & ADVERTISING INJURY	$ 2,000,000	OTHER:
EACH OCCURRENCE	$ 1,000,000	TOTAL: 27,500
DAMAGE TO RENTED PREMISES (each occurrence)	$ 1,000,000	
MEDICAL EXPENSE (Any one person)	$ 100,000	
EMPLOYEE BENEFITS	$ 5,000	
	$	

OTHER COVERAGES, RESTRICTIONS AND/OR ENDORSEMENTS (For hired/non-owned auto coverages attach the applicable state Business Auto Section, ACORD 137)

APPLICABLE ONLY IN WISCONSIN: IF NON-OWNED ONLY AUTO COVERAGE IS TO BE PROVIDED UNDER THE POLICY:
1. UM / UIM COVERAGE [ ] IS [ ] IS NOT AVAILABLE. 2. MEDICAL PAYMENTS COVERAGE [ ] IS [ ] IS NOT AVAILABLE.

SCHEDULE OF HAZARDS

LOC #	HAZ #	CLASSIFICATION	CLASS CODE	PREMIUM BASIS	EXPOSURE	TERR	RATE PREM/OPS	RATE PRODUCTS	PREMIUM PREM/OPS	PREMIUM PRODUCTS
	1	Body Shop	58942	S	3,000,000	001				
	2	Park and Go	54278	S	100,000	001				
	3	Gas Sales	51397	S	1,500,000	001				

RATING AND PREMIUM BASIS
(S) GROSS SALES - PER $1,000/SALES
(P) PAYROLL - PER $1,000/PAY
(A) AREA - PER 1,000/SQ FT
(C) TOTAL COST - PER $1,000/COST
(M) ADMISSIONS - PER 1,000/ADM
(U) UNIT - PER UNIT
(T) OTHER

CLAIMS MADE (Explain all "Yes" responses)

EXPLAIN ALL "YES" RESPONSES	Y / N
1. PROPOSED RETROACTIVE DATE:	
2. ENTRY DATE INTO UNINTERRUPTED CLAIMS MADE COVERAGE:	
3. HAS ANY PRODUCT, WORK, ACCIDENT, OR LOCATION BEEN EXCLUDED, UNINSURED OR SELF-INSURED FROM ANY PREVIOUS COVERAGE?	
4. WAS TAIL COVERAGE PURCHASED UNDER ANY PREVIOUS POLICY?	

EMPLOYEE BENEFITS LIABILITY

1. DEDUCTIBLE PER CLAIM: $	3. NUMBER OF EMPLOYEES COVERED BY EMPLOYEE BENEFITS PLANS:
2. NUMBER OF EMPLOYEES:	4. RETROACTIVE DATE:

Computer vision is an interdisciplinary field that deals with how computers can be made to gain high-level understanding from digital images or videos.

Or in other words…

“This is one of the most complex processes we’ve ever attempted to comprehend – let alone recreate. Inventing a machine that sees like we do is a deceptively difficult task, not just because it’s hard to make computers do it, but because we’re not entirely sure how we do it in the first place.

What actually happens is roughly this: the image of the ball passes through your eye and strikes your retina, which does some elementary analysis and sends it along to the brain, where the visual cortex more thoroughly analyzes the image. It then sends it out to the rest of the cortex, which compares it to everything it already knows, classifies the objects and dimensions, and finally decides on something to do: raise your hand and catch the ball (having predicted its path). This takes place in a tiny fraction of a second, with almost no conscious effort, and almost never fails. So recreating human vision isn’t just a hard problem, it’s a set of them, each of which relies on the other.”

-Devin Coldewey@techcrunch / 2 years ago

But for now, let’s let our friends at Stanford and MIT develop cutting edge algorithms… we’re in the application game.

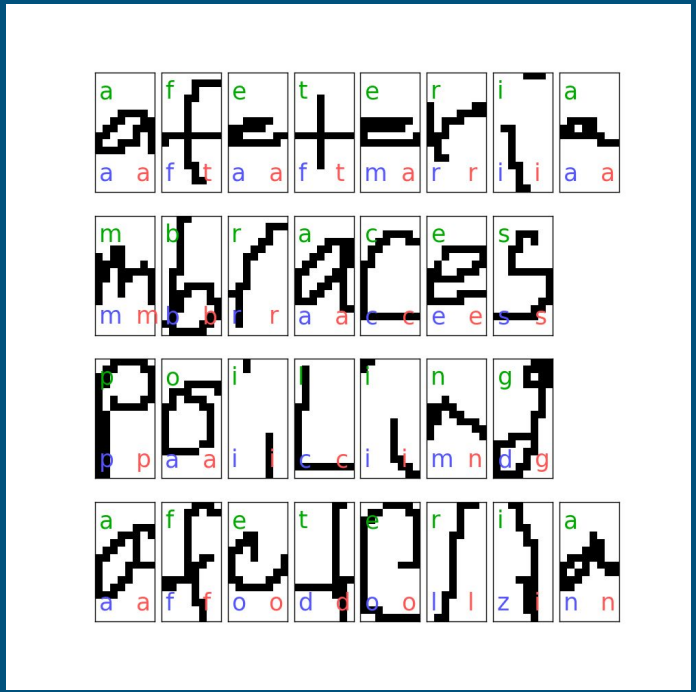
Computers can help us get value from unstructured data through process like Optical Character Recognition (OCR)

OCR is a form of image **classification**.

Common classification algorithms are used (... or used to be used) to help classify images.

- Naive Bayes
- SVM
- KNN

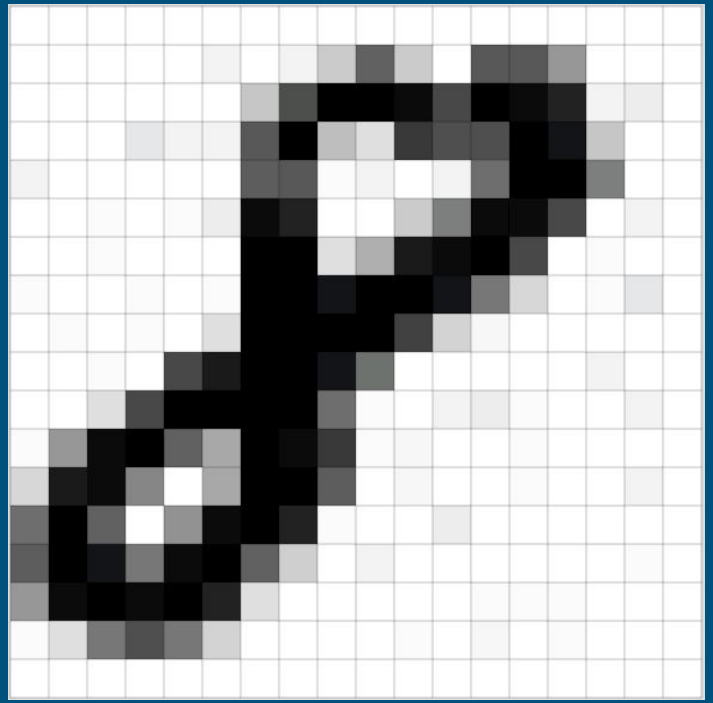
But now.... Convolutional Neural Networks (CNN)



https://pystruct.github.io/auto_examples/plot_letters.html

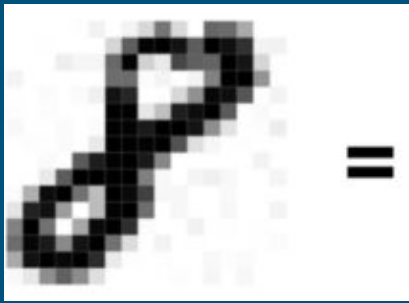
But... what exactly are you classifying? I mean thought you needed data for classification models?

- A computer sees an image as an array of values.
  - Grayscale images take the form of a 2d array with values ranging from 0 - 255.
  - Color images take the form of a 3d array with r, g, b values ranging from 0 - 255.



https://medium.com/@tifa2up/image-classification-using-deep-neural-networks-a-beginner-friendly-approach-using-tensorflow-94b0a090ccd4

And voilà, flatten those values and now you have a single observation and n-features.



=

```
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 12, 0, 11, 39, 137, 37, 0, 152, 147, 84, 0, 0, 0, 0,
0, 1, 0, 0, 0, 41, 160, 250, 255, 235, 162, 255, 238, 206, 11, 13, 0, 0, 0, 0, 16, 9, 9, 150, 251, 45, 21, 184, 159, 154, 2
55, 233, 40, 0, 0, 10, 0, 0, 0, 0, 0, 145, 146, 3, 10, 0, 11, 124, 253, 255, 107, 0, 0, 0, 0, 3, 0, 4, 15, 236, 216, 0, 0,
38, 109, 247, 240, 169, 0, 11, 0, 1, 0, 2, 0, 0, 0, 253, 253, 23, 62, 224, 241, 255, 164, 0, 5, 0, 0, 6, 0, 0, 4, 0, 3, 252
, 250, 228, 255, 255, 234, 112, 28, 0, 2, 17, 0, 0, 2, 1, 4, 0, 21, 255, 253, 251, 255, 172, 31, 8, 0, 1, 0, 0, 0, 0, 0, 4,
 0, 163, 225, 251, 255, 229, 120, 0, 0, 0, 0, 0, 11, 0, 0, 0, 0, 21, 162, 255, 255, 254, 255, 126, 6, 0, 10, 14, 6, 0, 0, 9
, 0, 3, 79, 242, 255, 141, 66, 255, 245, 189, 7, 8, 0, 0, 5, 0, 0, 0, 0, 26, 221, 237, 98, 0, 67, 251, 255, 144, 0, 8, 0, 0
, 7, 0, 0, 11, 0, 125, 255, 141, 0, 87, 244, 255, 208, 3, 0, 0, 13, 0, 1, 0, 1, 0, 0, 145, 248, 228, 116, 235, 255, 141, 34
, 0, 11, 0, 1, 0, 0, 0, 1, 3, 0, 85, 237, 253, 246, 255, 210, 21, 1, 0, 1, 0, 0, 6, 2, 4, 0, 0, 0, 6, 23, 112, 157, 114, 32
, 0, 0, 0, 0, 2, 0, 8, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

https://medium.com/@tifa2up/image-classification-using-deep-neural-networks-a-beginner-friendly-approach-using-tensorflow-94b0a090ccd4

While understanding the technology is important, you likely will never need to build your own OCR engine.

Google Cloud Platform

ABBYY®

Resources

Modules/Utilities:

- **XpdfReader**
  - **Pdftopng**
- **OpenCV**
- **Numpy**
- **Keras (for deep learning)**

Resources:

- http://cs231n.stanford.edu/
- https://becominghuman.ai/
- https://hackernoon.com/

So let’s get to it