Eviction Addresses Pipeline

Brancen Gregory

7/9/2021

Contents

1 Project Overview

A growing group of stakeholders in Tulsa is interested in using data to respond to evictions. Open Justice Oklahoma collects and stores data from district court dockets, including party names, events, and dispositions, that are available in HTML text on case pages. Other key information, however, is not available in HTML text, but in PDF documents. This project outlines three different processes to collect address information from PDFs through a combination of technological tools and manual review.

2 Proposal

Given the high-priority and critical nature of the project, Open Justice Oklahoma suggests using either the Full Manual or Human-in-the-Loop strategy. A fully automated solution simply won't give a level of address coverage that is sufficient for the needs of stakeholders.

Selection between the two remaining strategies is a matter for BEST to weigh, whether BEST has sufficient ability to contract/hire someone part-time and onboard them to perform these functions; or whether a remote, platform-based workforce is required.

In our estimation the Human-in-the Loop strategy is the way to go. There is ample room to increase the confidence threshold required of an automated address transcription, such that coverage approaches the level of a Full Manual approach. In addition the manual labor portion of the process would be a 'code-only' interaction. No human resources would be involved, beyond that of the developer.

If OJO were to move forward with the Human-in-the-Loop strategy, I would suggest a 2-4 week trial run of pricing address transcription via a distributed workforce system such as Amazon Mechanical Turk. This would require minimal development and would allow OJO to get a more concrete sense of the cost associated with the manual labor.

3 Definitions

OJO	Open Justice Oklahoma
HIT	Unit task assigned to a worker e.g. transcribe these 10 addresses

4 Strategies

4.1 Full automation

- Cheapest solution
- Stable cost over time
- Least coverage
- Extensive development time to improve model

4.2 Full manual

- Most expensive solution
- Stable cost over time
- Most coverage
- Little development time
- Local or remote labor

4.3 Human-in-the-Loop

- Can modify confidence threshold to increase coverage
- Cost approaches full manual solution with increased confidence
- Local or remote labor, local solution takes more development
- Cost varies with demand for HITs if using remote labor
- Requires most infrastructure

5 Common Infrastructure

- Deployed script to download and store documents for new evictions
- Deployed script for document pre-processing
  - Cropping
  - Contrast boosting
- Deployed script for address validation and standardization

6 Strategy Specific Infrastructure

6.1 Full Automation

- OCR/computer vision processing
- Address extraction (needed if cropping is too irregular)

6.2 Full Manual

- Remote
  - Platform specific script to submit new cases
- Local
  - Simple platform for data entry and delivery back to OJO

6.3 Human-in-the-Loop

- All full automation infrastructure
- Remote or local manual infrastructure
- Remote
  - Script for pricing HITs according to demand

7 Infrastructure Pricing

Pricing the various infrastructure components at this stage is imperfect to say the least, though some are easier than others.

As addressed in the Proposal section, depending on the strategy selected, there would need to be a trial period to estimate total costs. This is largely relevant to the Human-in-the-Loop strategy, which requires dynamic pricing to achieve the kind of coverage of eviction addresses that is desired.

Actual cloud computing costs are minimal in all cases. The bulk of cost comes from manual labor and platform fees.

7.1 Full Automation

Estimating a monthly case load of ~2000 evictions, the cost for the Full Automation strategy would cost less than $100 per month.

100% of this cost is cloud computing, most of which stays within the free-tier of cloud services.

7.2 Full Manual

Again using a case load of 2000/month, adding a labor cost of $15/hr, and assuming a labor speed of 100 addresses transcribed per hour, I estimate the total cost to be around $500/month. We could be more precise by testing how many addresses could reasonably be typed. To do this we would need to approximate the system which workers would be using to perform the task.

7.3 Human-in-the-Loop

Pricing this strategy in full is unfortunately not possible without some kind of trial. However, we can use the infrastructure cost from the Full Automation strategy, together with additional infrastructure costs, to arrive at an estimate of less than $150/month in cloud computing resources.

We can assume that the total human labor cost of this strategy would not exceed that of the Full Manual strategy, and so can include the $400/month labor cost, arriving at an upper estimate of less than $600/month.

8 Development

Regardless of the chosen strategy there is significant labor involved to set up a seamless workflow. However, the development labor required in each strategy varies greatly and is not a direct trade-off with data-entry manual labor.

Some work has already been done such that deploying the infrastructure common to all strategies could be completed rapidly. Other strategy specific tasks will require research and optimization on the fly, and so the time estimates will be less solid.

In addition, we provide two time estimates for each task. The first will represent the actual time spent on the project, whereas the second will represent the number of weeks necessary for me to deliver that task given my current workload.

8.1 Common Infrastructure

Task	Labor Time (hrs)	Time to Delivery (weeks)
Document Acquisition	8	1
Document Pre-processing	12	1
Address Validation	10	1
Total	30	3

8.2 Full Automation

Task	Labor Time (hrs)	Time to Delivery (weeks)
Common	30	3
OCR optimization	20	1-2
Address extraction	12	1
Total	62	5-6

8.3 Full Manual

Task	Labor Time (hrs)	Time to Delivery (weeks)
Common	30	3
Case submission and delivery	20	1-2
Total	50	4-5

8.4 Human-in-the-Loop

Task	Labor Time (hrs)	Time to Delivery (weeks)
Common	30	3
Automation	32	2-3
Manual	20	1-2
Automated pricing	12	1
Total	94	7-9

9 Total Yearly Cost

Using the infrastructure cost estimates above, the one-time development estimates, and a maintenance cost of 5% of development, the total cost for each strategy is given below. If we seek a contractor out for this work, the totals will depend on the hourly or project rate agreed upon.

Strategy	Total
Full Automation	$9,638
Full Manual	$14,063
Human-in-the-Loop	$21,338