Regular Expressions for Address Parsing in Unstructured Data

Colin Riker, Sara Haycox, Anastasiia Noguier, Stefan Larson

Introduction

Unstructured data is a growing occurrence in our modern data-driven society. The prevalence of unstructured data presents issues with traditional programmatic methods for automation and manipulation. An area of concern in unstructured data is personally identifiable information (PII) which is an area of interest both for individuals, corporations, and governments.

As such, the project chose to focus on an intersection of PII and unstructured data: street addresses. Addresses present several fundamental challenges when attempting to identify and parse them accurately. This comes from the complexity of postal systems and the imperfections of data given by real people. However, there are many methods to working with unstructured data from machine learning, pattern matching, and manual annotation. The project elected to explore how far a pattern-matching approach could be taken. The project's goal was to utilize a data-driven approach to building regular expressions for unstructured data, i.e. addresses.

What are Regular Expressions?

- Regular expressions are a language for describing patterns in natural language
- An example expression for Canadian postal boxes would be:

```
(?P<postal_box>\b(?:
    # English - PO Box 123
    (?:[Pp][\ \.]{0,2}[Oo][\ \.]{0,2}[Bb][Oo][Xx]\ \d{1,6})
    |
    # French - B.P. 123
    (?:[Bb][\ \.]{0,2}[Pp][\ \.\n]{0,2}\d{1,6})
    |
    # C.P. 123 CP123 CP 123 C.P123 C.P.123 etc
    (?:[Cc][\ \.]{0,2}[Pp][\ \.\n]{0,2}\d{1,6})
    |
    # Case postale 123
    (?:[Cc]ase\ [Pp][Oo][Ss][Tt][Aa][Ll][Ee][\ \n]\d{1,6})
))
```

Methods

- Project builds off prior work,the open-source project called the python address parser (PYAP)[1]
- Using a custom evaluation script developed as a part of the project, issues and edge cases were identified in the PYAP expressions.
- Utilize key sub-patterns or markers present in the data to enhance pattern matching (See Key Markers for Identification)

Qualifying Success

- Matches can either perfectly match or be a partial or subset of the annotated address (the answer).
- Allowing subset matching helps to allow matching of novel cases that still have underlying common patterns shared with more standard cases.
- The metrics of improvement were recall and precision. Where recall represents the total percentage of all possible results returned. Precision represents the percentage of results returned that are correct.
- A composite metric of precision and recall, called F-Measure, was employed to track overall improvements between the two metrics.

Key Markers for Identification

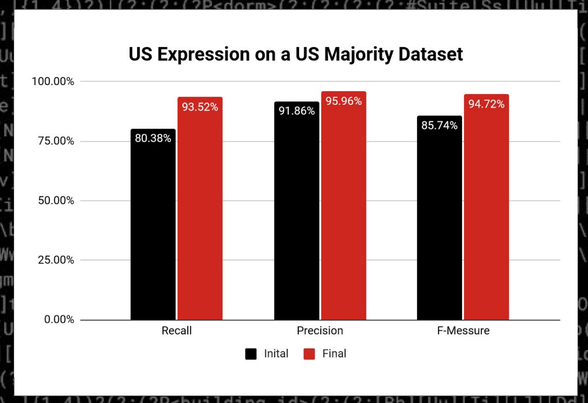
500 S. State Street, Ann Arbor, MI 48109

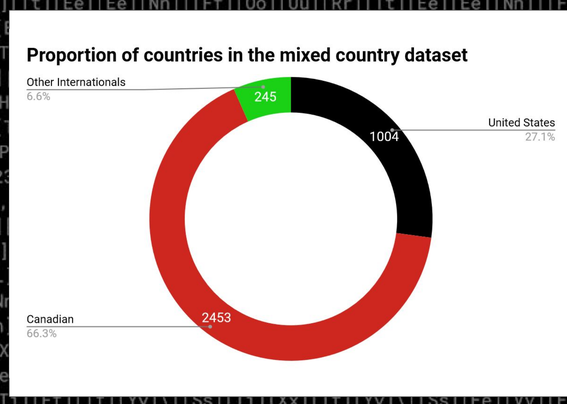
Street Number	Street Type	State & Zipcode

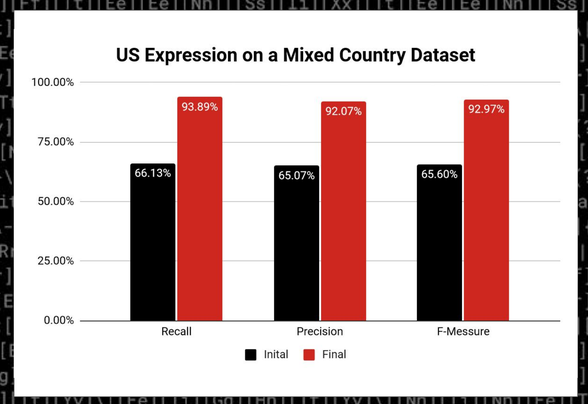
- Key Markers are elements in the data that have well defined characteristics or patterns
- Poor markers like street name can be any character thus lacking any good pattern to match
- Markers can have even better quality with strong internal patterns (i.e. Canadian postcodes).
- Full addresses are defined as a combination of at least 3 key markers.

Results

- US Expression F-measure improvement of 8.98% in the US majority dataset and 27.37% in the mixed country dataset
- Canadian Expression F-measure improvement of 13.5% in the mixed country dataset
- Built US majority dataset consisting of resumes, directories, and other categories, containing over 1700 addresses across 2000 documents
- Built a mixed country dataset with 2698 International addresses and 1004 US addresses







Example Matches and Errors

Covered Patterns:

123A Wadsworth Hall, 1700 Townsend Dr, Houghton, 49931
- Expression also identifies a style of dorm room addressing

P.O. Box 100, Lexington, PA 38351
- P.O Boxes are a substitute for street number & type

100 N. 10th St. Apt. 1C, Ann Arbor, MI 49053
- Supports many forms of multi-occupancy addresses

123456 OH-34, Archbold, OH 43502
- Handles rural route

123 Bloor Street East, 4th floor, Toronto ON, M4W 2A2
- Muli-floor urban address support

25 Promenade du Portage, Montréal, Québec, J2T2B7
- Allows for the use of french canadian styled addressing

Typical Errors:

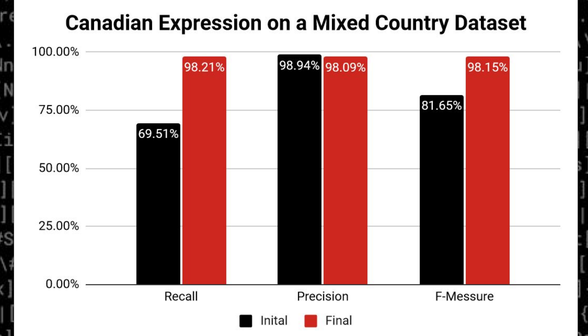
1234 Route South, Bldg D6 Raritan Office Campus CA 8869
- Typos or incomplete structures are the biggest source of False negatives

20th Ave and W. 30th Ave. Heading North on
- Common Province and State abbreviations often are also words and can create false positives

100-100 10 Ave W Vancouver ___ ___
- Lacks sufficient markers to qualify as a match

5 Outdoor Field Trip Guide - Michigan
- Field is a valid street type even though the context is not an address

12345 Alaska Rd 12345 Alaska (Rd Fort St. John V1J1B2)
- The US expression falsely matches due to the duplicated street portion of this canadian address that contains a valid US state and five digit street number.



Conclusion

- Addresses can be accurately found in unstructured data using a pattern matching approach like regular expressions.
- Data driven pattern development allowed substantial gains over initial expression performance to happen.
- Composite patterns can lead to better quality matching if sufficient sub pattern exist in the data.
- Future applications in PII identification in unstructured data might also be solved through pattern matching

References

1. Goncharov, V., & Bird, T. "Pyap: Python address parser" https://github.com/vladimarius/pyap

dryvIQ

UROP COMMUNITY COLLEGE SUMMER FELLOWSHIP PROGRAM