

# Geo::Parser::Text

Geo parsing and Geo coding locations from text.

Ervin Ruci - YAPC::NA 2016

## About this talk

- ① I gave the first version of this talk at FOSDEM 2016 (Geospatial Devroom)
- ② Some members of the Audience expressed surprise at me choice of Perl for such an application (?!)
- ③ So, here is an expanded version aimed at a Perl friendly audience
- ④ This talk (and various test data) are hosted at <https://github.com/eruci/openaddresses/tree/master/test>
- ⑤ The main application is available as a server image on AWS and <https://metacpan.org/pod/Geo::Parser::Text>

## Text => Geo Locations in Text

This and That and the Other street in Porters Lake Nova Scotia

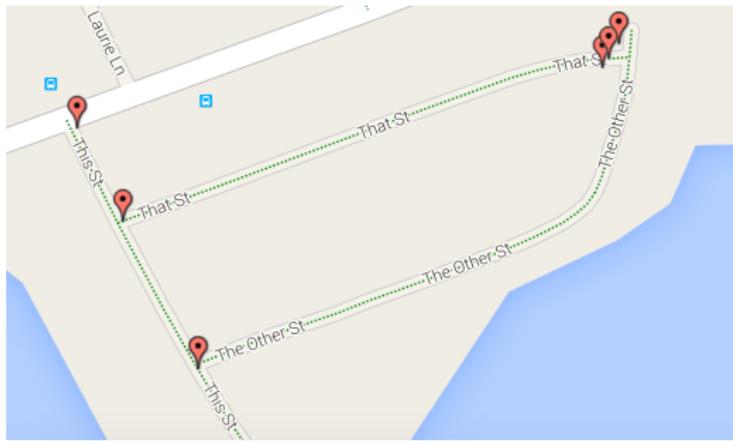
Q. How many locations are in this text?

# The Problem: Text => Geo Locations in Text

This and That and the Other street in Porters Lake Nova Scotia

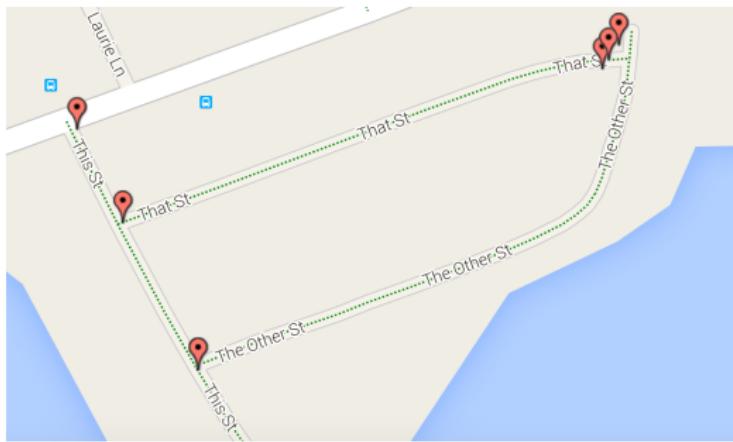
Q. How many locations are in this text?

A. 6. <http://geocoder.ca>



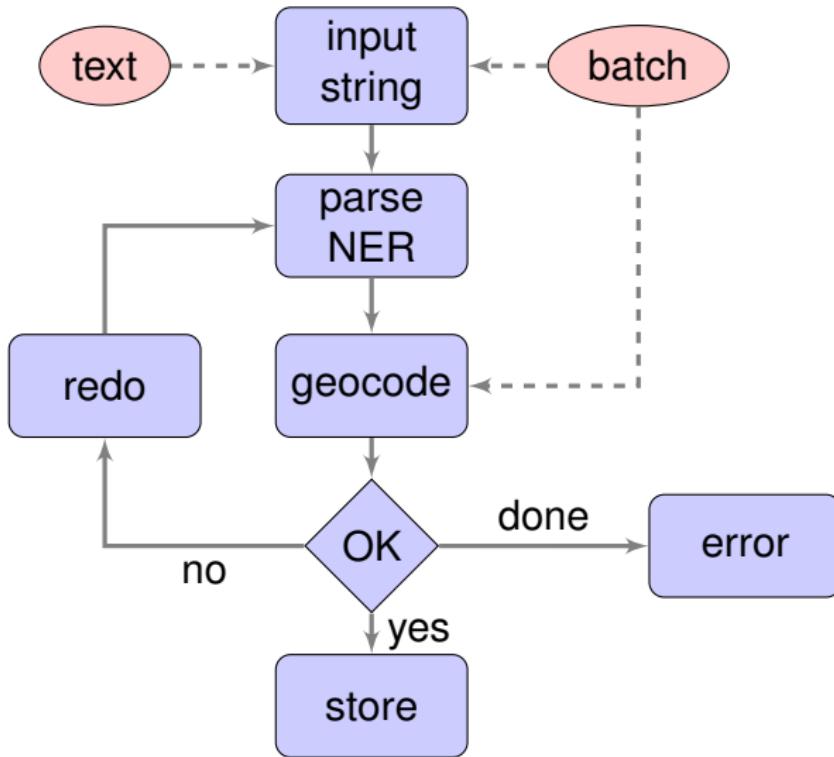
# The Problem: Text => Geo Locations in Text

- 1 THIS ST AND THAT ST, PORTERS LAKE, NS
- 2 THIS ST, PORTERS LAKE, NS
- 3 THAT ST, PORTERS LAKE, NS
- 4 THE OTHER ST AND THAT ST, PORTERS LAKE, NS
- 5 THE OTHER ST AND THIS ST, PORTERS LAKE, NS
- 6 THE OTHER ST, PORTERS LAKE, NS



Reqs: Identify addresses, intersections, city names, province/state.

# The Solution: Geoparsing and Geocoding



# Geocoding Spans Many Fields

- 1 linguistics (matching/translating across different languages)
  - 2 data processing (normalization, standardization and input)
  - 2.1 data structures (R-trees, KD-trees,...)
  - 3 natural language processing (parsing, named entity recognition)
  - 4 computational geometry (point in polygon)
  - 5 pattern recognition (fuzzy match)
  - 6 geography (dealing with projections)
  - 7 Ai (learning, hidden markov models)
  - ... and a few others (tokenization, data cleanup, UI..)
- AND** Testing, testing and more testing

# And there are Many Geocoders.

And Many more are being built, Plus a few I've tested:

- ① Google Geocoder (Coverage: 99%, Accurate 93%) (Canada)
- ② HERE.com (Coverage 98%, Accurate 92%) (Canada)
- ③ Nominatim (Coverage 80%, Accurate 57%) (Canada)
- ④ Geocoder.ca (Coverage 99%, Accurate 94%) (Canada)
- ⑤ Geocode.xyz (Coverage 80%, Accurate 58%) (Spain)
- ⑥ Mapzen.com (Coverage 86%, Accurate 80%) (Spain)

Download test data and results here: <https://github.com/eruci/openaddresses/tree/master/test>

## Why create a new Geocoder?

No Geocoder does it all. Google Geocoder (presumably the most complete in the bunch) does not

- ① Geocode parcel data (avail as opendata in Canada and USA)
- ② Extract location data from text
- ③ Do address parsing and standardization (incl postal codes)
- ④ Return all addresses that match a partial address in text
- ⑤ Provide 100% coverage (open problem)
- ⑥ Provide 100% accuracy (open problem too)

## Accuracy and Coverage Differ Because

- a Geocoding/Geoparsing is an imprecise process and various Geocoders fail in various ways.
- b Ambiguities, incomplete data, incorrect data, software bugs, are the main causes

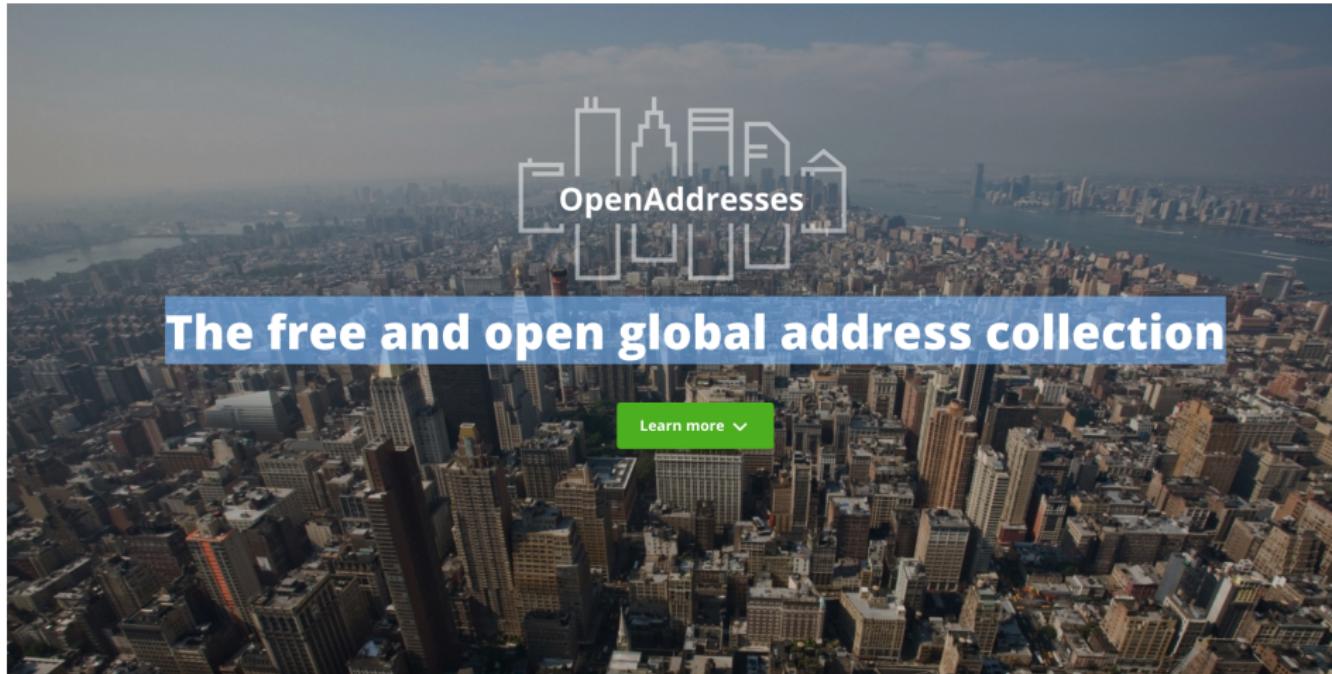
# The key ingredients of the solution

- 1 DATA
- 2 A good parser

# Parsing

- 1 Fuzzy vs Exact (correct spelling errors)
- 2 Partial vs Complete (fill in missing location entities)
- Quick demo: <http://geocoder.ca/textscan>

# Data



over 218,548,165 addresses (was half that only 6 months ago)

# Even Google Maps (presumably the best) Fails!

- ✖ Even in well mapped big cities.

# Even when you Google wrong location you get:



wrong location |



wrong location

wrong location **on google maps**

wrong location **on google**

wrong location **on facebook**

wrong location **tinder**

wrong location **on iphone**

wrong location **on phone**

wrong location **android**

wrong location **on traffic ticket**

wrong location **on parking ticket**



How to write In Deutsch  
Grazes of Country  
Are you registered to  
Some District of Col  
with the mail.

: Wrong Location Google Maps!.

# Carrer de Colomines, 2 Entresuelo 200AA, Barcelona, Catalunya 08003, Spain

≡ Carrer de Colomines, 2 Entresuelo 2<sup>a</sup>, Barcelona, C. X

🔍

📍 Search nearby: hotels · restaurants

Ank

Bruc 19, Entresuelo 2<sup>a</sup>  
08010 Barcelona  
Spain

Directions Save +34 717 12 75 79

Send to device

Ad Book a room

Check-in Thu, Aug 20 Check-out Fri, Aug 21

Booking.com \$104 / night Book

BudgetPlaces.com \$104 / night Book

View 1 more booking option at \$103

Street View...

Mango Outlet  
Grupotel Gran Via 678  
Caspe  
Carrer del Bruc  
Carrer de Llúria  
Carrer d'AUS  
Ank

: my Airbnb in Barcelona: Wrong!.

# <http://Geocode.xyz> - A geocoder for the EU

Fixed with:

**GeoCode.xyz**

API

BARCELONA, ES » 2 COLOMINES, BARCELONA, ES » [41.3855238000,2.1789661000 Directions](#)

2 COLOMINES Calle, BARCELONA, ES ([BARCELONA,ES polygon](#)) [Directions](#) [Reverse Geocode](#)

Confidence Score: 0

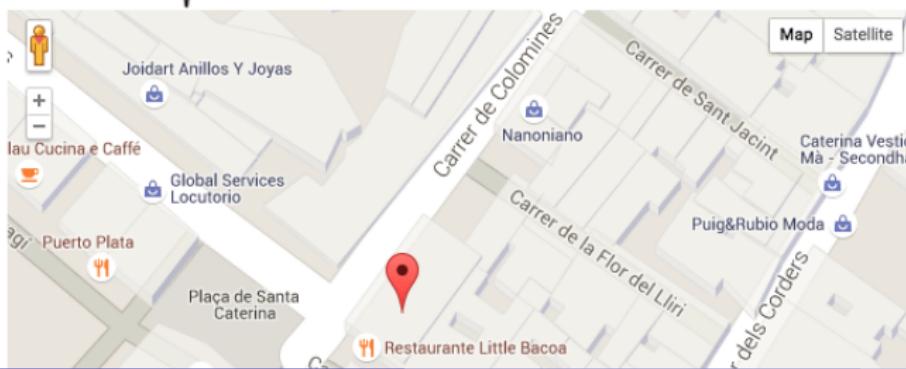
Is the location shown in the Map incorrect?

Then click here to send your corrections

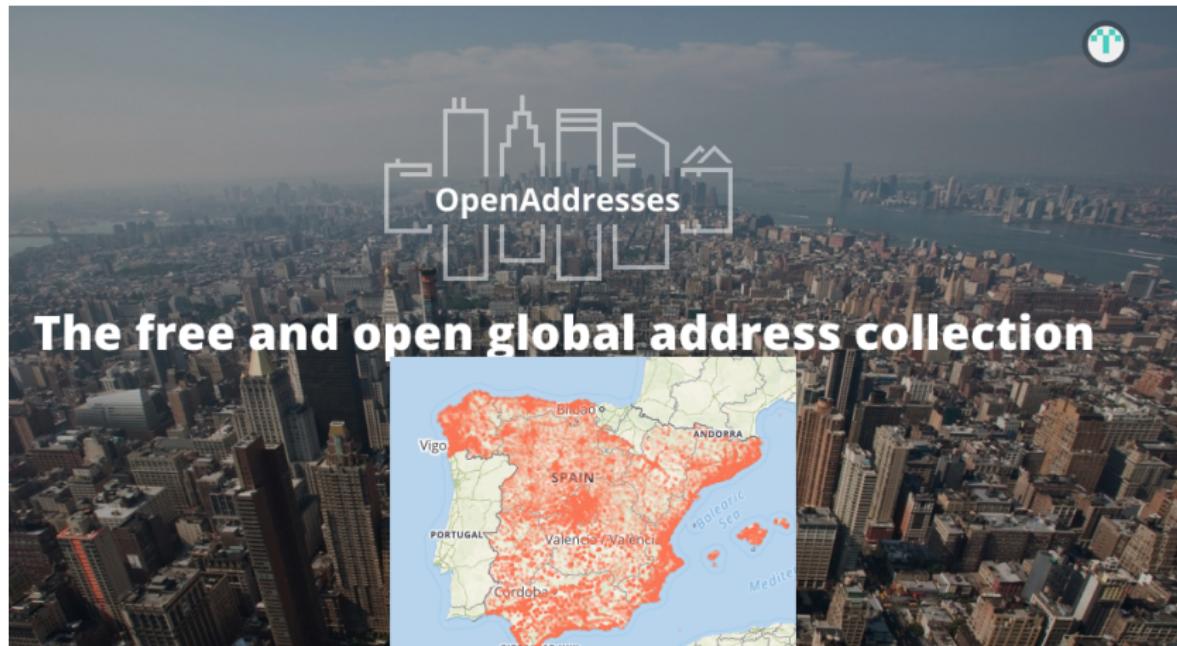
Drag the marker to correct this location

41.3855238000, 2.1789661000

[Geocode this Location on a Map](#)



# Correct, but, Coverage for Spain is limited...



: [Data]

For the latest coverage map see <http://openaddresses.io>

# Geoparsing and Geocoding in unstructured text

Text (from wikipedia entries to microblog posts) => Geocoded Locations.

- 1 extraction
- 2 disambiguation
- 3 geolocation

Demo: <http://geocode.xyz>

# In case internet does not work for the demo



Toronto Police OPS @TPSOoperations · Jan 26

Robbery investigation on @GOtransit. Suspect pulled brake, hopped off train at Birchmount Rd/Raleigh Ave on LakeShore line. ^vk



17

8

•••

: [Twitter Feed Demo geocoder.ca]

40 locations within a 3484.180 km radius, found in this text:

Toronto Police OPS Verified account @TPSOoperations · Jan 26

Robbery investigation on @GOtransit. Suspect pulled brake, hopped off train at Birchmount Rd/Raleigh Ave on LakeShore line. ^vk

Reprocess text and Download results

On a Map



#### Match Location

1	Raleigh Ave And Birchmount Rd, Toronto, On	Confidence Score: 1
2	Raleigh, Nc	Confidence Score: 0.8
3	Toronto, On	Confidence Score: 0.8
4	Raleigh, Ms	Confidence Score: 0.8
5	Raleigh, Fl	Confidence Score: 0.8
6	Toronto, Pe	Confidence Score: 0.8
7	Toronto, Ks	Confidence Score: 0.8
8	Toronto, In	Confidence Score: 0.8
9	Raleigh, Wv	Confidence Score: 0.8
10	Toronto, Tx	Confidence Score: 0.8

# In case internet does not work for the demo

Geocoder.ca Services | Products | Solutions Terms Login Create Account API Contact

Found 5 locations in this line:

6 locations within a 321.011 km radius of Birchmount Rd/Raleigh Ave, ON

Toronto Police OPS Verified a Robbery investigation on @G

Reprocess text and Data

Suspect pulled brake, hopped off train at Birchmount<sup>1</sup> Rd<sup>1</sup>/Raleigh<sup>1</sup> Ave<sup>1</sup> on LakeShore line

Map Satellite

1. RALEIGH AVE AND BIRCHMOUNT RD, TORONTO, ON (Confidence: 1)  
2. LAKESHORE AVE, TORONTO, ON (Confidence: 0.8)  
3. RALEIGH AVE, TORONTO, ON (Confidence: 0.6)  
4. BIRCHMOUNT RD, TORONTO, ON (Confidence: 0.5)  
5. LakeShore, ON (Confidence: 0.5)

1	Raleigh Ave And Birchmount Rd, Toronto, On	Confidence: 1
2	Lakeshore Ave, Toronto, On	Confidence: 0.8
3	Toronto, On	Confidence: 0.6
4	Raleigh Ave, Toronto, On	Confidence: 0.5
5	Birchmount Rd, Toronto, On	Confidence: 0.5

: [Geoparsed]

# In case internet does not work for the demo

**GeoCode.xyz** API

2 locations within a 1.084 km radius, found in this text:

The most important museums of Amsterdam are located on the Museumplein (Museum Square), located at the southwestern side of the Rijksmuseum.

Reprocess text and Download results On a Map

Match	Location	Confidence Score
1	Amsterdam, NL	0.7
2	Museumplein, Amsterdam, NL	0.3

Click a marker for more information.

: [Demo geocode.xyz]

# In case internet does not work for the demo

GeoCode.xyz

Found 2 locations in this line:

The most important museums of Amsterdam<sup>2</sup> are located on the Museumplein<sup>2</sup> (Museum Square), located at the southwestern side of the Rijksmuseum

1. [Amsterdam, NL](#) (Confidence: 0.7)  
2. [MUSEUMPLEIN, AMSTERDAM, NL](#) (Confidence: 0.3)

2 locations within a 1.084 km radius:  
The most important museum in Amsterdam is the Rijksmuseum.

Reprocess text and Data

A map of Amsterdam with the city center highlighted in red. The 'Museumplein' is marked with a blue dot in the southern part of the city. The map also shows the 'GRACHTENGORDEL' (canal belt) and various neighborhoods like 'AMSTERDAM-WEST', 'OVERTOMSE VELD', 'WEESPERBUURT EN PLANTAGE', 'INDISCHE BUURT', and 'DOSTERPARKBUURT'. A legend on the right indicates 'Match' and 'Location' for the two found matches.

Match	Location	Confidence Score
1	Amsterdam, NL	0.7
2	Museumplein, Amsterdam, NL	0.3

Click a marker for more information.

: [Demo geocode.xyz]

# In case internet does not work for the demo

GeoCode.xyz

API

7 locations within a 1.803 km radius, found in this text:

Bruxelles/Brussel - Brussels encompasses many charming and beautiful attractions, with deeply ornate buildings on the Grand Place/Grote Markt, and a fish-and-crustacean overdose of St. Catherine's Square (Place St-Catherine/Sint-Katelijneplein). Stroll along, (and stop in for a drink)

Reprocess text and Download results

On a Map



Match	Location
1	Antoine Dansaertstraat, Brussels, BE Confidence Score: 0.8
2	Brussels, BE Confidence Score: 0.7
3	Antoine Dansaert Rue, Brussels, BE Confidence Score: 0.7
4	Sint-Goriksplein, Brussels, BE Confidence Score: 0.6
5	Sint-Katelijneplein, Brussels, BE Confidence Score: 0.4
6	Grote Markt, Brussels, BE Confidence Score: 0.2
7	Grand Place, Brussels, BE Confidence Score: 0.2

: [Demo geocode.xyz]

# In case internet does not work for the demo

GeoCode.xyz API

Found 7 locations in this line:

7 locations within a 1.803 km

Bruxelles/Brussel - Brussels Markt, and a fish-and-crustacean overdose of St. Catherine's Square (Place St-Catherine/Sint-Katelijneplein), Stroll along, (and stop in for a drink) at one of the many bars on Place St-GÃ©ry/Sint-Goriksplein, or max out your credit card on the trendy Rue Antoine<sup>3</sup> Dansaert<sup>3</sup>/Antoine<sup>3</sup> Dansaertstraat

Reprocess text and D

ANTOINE DANSAERTSTRAAT, BRUSSELS, BE (Confidence: 0.8)

2. Brussels, BE (Confidence: 0.7)

3. ANTOINE DANSAERT RUE, BRUSSELS, BE (Confidence: 0.7)

4. SINT-GORIKSPLEIN, BRUSSELS, BE (Confidence: 0.6)

5. SINT-KATELIJNEPLEIN, BRUSSELS, BE (Confidence: 0.4)

6. GROTE MARKT, BRUSSELS, BE (Confidence: 0.2)

7. GRAND PLACE, BRUSSELS, BE (Confidence: 0.2)

1. ANTOINE DANSAERTSTRAAT, BRUSSELS, BE (Confidence: 0.8)

2. Brussels, BE (Confidence: 0.7)

3. ANTOINE DANSAERT RUE, BRUSSELS, BE (Confidence: 0.7)

4. SINT-GORIKSPLEIN, BRUSSELS, BE (Confidence: 0.6)

5. SINT-KATELIJNEPLEIN, BRUSSELS, BE (Confidence: 0.4)

6. GROTE MARKT, BRUSSELS, BE (Confidence: 0.2)

7. GRAND PLACE, BRUSSELS, BE (Confidence: 0.2)

1. ANTOINE DANSAERTSTRAAT, BRUSSELS, BE (Confidence: 0.8)

2. Brussels, BE (Confidence: 0.7)

3. ANTOINE DANSAERT RUE, BRUSSELS, BE (Confidence: 0.7)

4. SINT-GORIKSPLEIN, BRUSSELS, BE (Confidence: 0.6)

5. SINT-KATELIJNEPLEIN, BRUSSELS, BE (Confidence: 0.4)

6. Grote Markt, Brussels, BE (Confidence: 0.2)

7. Grand Place, Brussels, BE (Confidence: 0.2)

: [Demo geocode.xyz]

# Coding a Geocoder that does this is easy (in theory)

But.. making it recognize over 90% of input at over 90% accuracy requires at least these steps

- ① importing and parsing country specific data from openaddresses.io (suffixes, prefixes, city names, numbering schemes)
- ② cleaning up errors post import.
- ③ test and pick away at errors, one at a time

In theory, there is no difference between theory and practice. But, in practice, there is.

# Perl makes easy problems very easy

.. and hard ones, easy. (in both theory and practice)

- 1 No other language is better at slicing and dicing text.
- 2 CPAN has lots of valuable bits for any NLP puzzle such as this one
- 3 Do I need another reason to use perl?

## Source Code / Data

Source code and Data: <http://geocode.xyz>

Just grab the server image on AWS, it is free for a micro instance

G  
e  
o  
c  
o  
d  
e  
r  
.c  
a

If you need help: e: [eruci@geocoder.ca](mailto:eruci@geocoder.ca) twitter: [@geolytica](https://twitter.com/geolytica)  
*PS. One more thing. The core module is 47355 lines of Perl code.*