

The Problem: Text => Geo Locations in Text

A horizontal strip of light gray paper with irregular, torn edges, giving it a scrap-like appearance. It is centered on the slide.

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

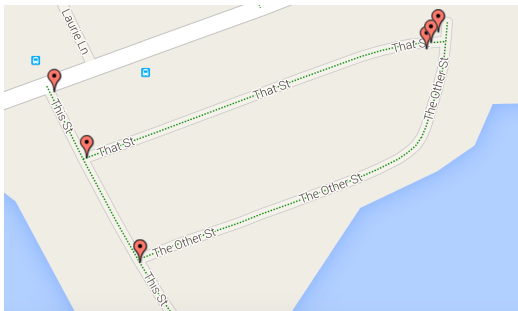
Q. How many locations are in this text?

The Problem: Text => Geo Locations in Text

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

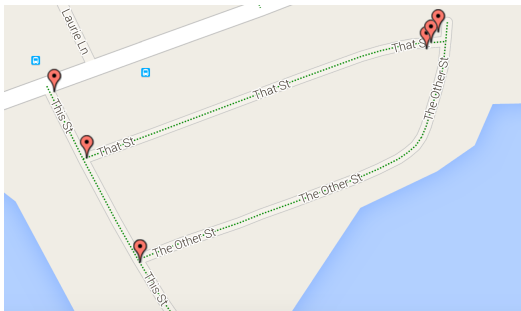
Q. How many locations are in this text?

A. 6.



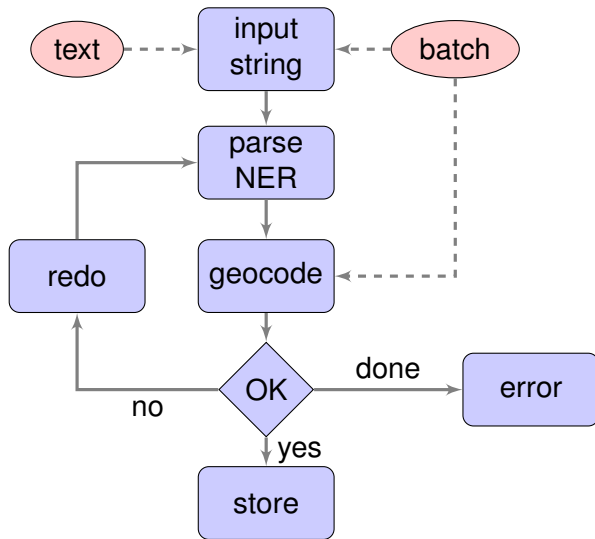
The Problem: Text => Geo Locations in Text

- 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: 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)
- 3 Return all addresses that match a partial address
- 4 Provide 100% coverage (open problem)
- 5 Provide 100% accuracy (open problem too)

Accuracy and Coverage Differ Because

- a Geocoding 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>

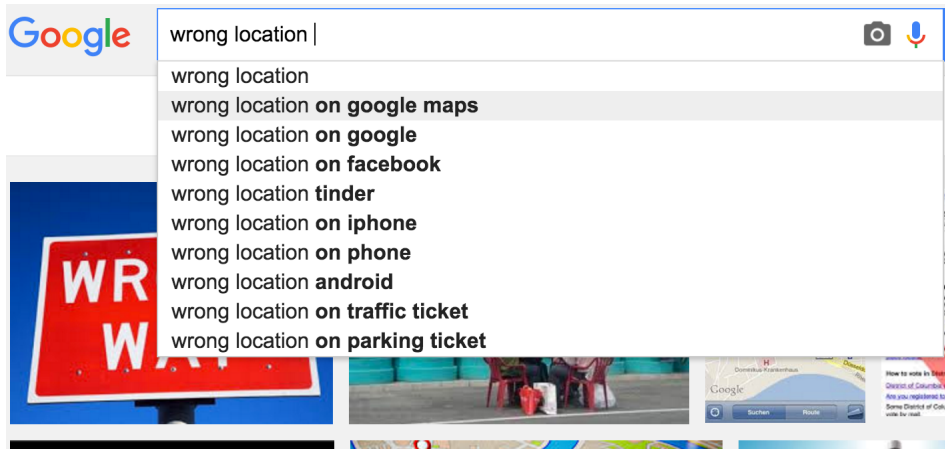


over 218,548,165 addresses (was half that only 6 months ago)

Even Google Maps (presumably the best) Fails!

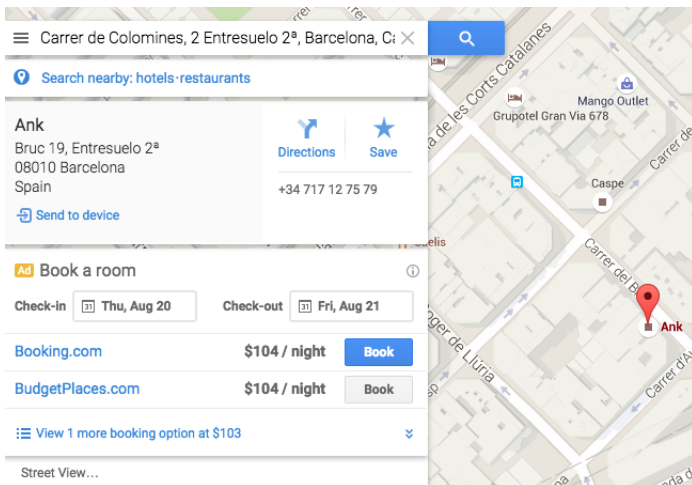
- ✖ Even in well mapped big cities.

Even when you Google wrong location you get:



: Wrong Location Google Maps!.

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



: my Airbnb in Barcelona: Wrong!.

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

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? 

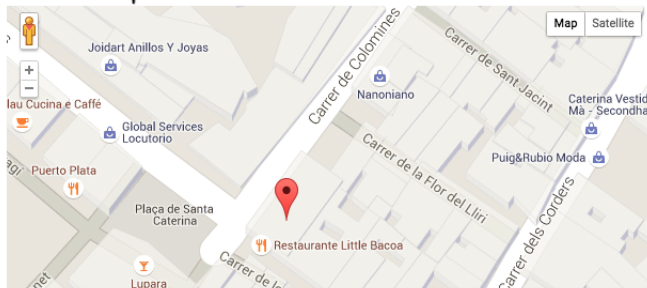
Then click here to send your corrections

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]

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

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

1

Amsterdam, NL

Confidence Score: 0.7

2

Museumplein, Amsterdam, NL

Confidence Score: 0.3

Click a marker  for more information.

: [Demo geocode.xyz]

In case internet does not work for the demo

The screenshot displays the GeoCode.xyz web application. At the top left is the logo "GeoCode.xyz" and at the top right is the text "API". A text input field contains the sentence: "The most important museums of Amsterdam² are located on the Museumplein² (Museum Square), located at the southwestern side of the Rijksmuseum." Below the input, a button labeled "Reprocess text and D" is visible. A white modal box is open, displaying the results:

Found 2 locations in this line:

The most important museums of Amsterdam² are located on the Museumplein² (Museum Square), located at the southwestern side of the Rijksmuseum

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

Below the modal, a map of Amsterdam is shown with a street view. To the right of the map is a table with two columns: "Match" and "Location".

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

At the bottom of the table area, there is a button that says "Click a marker for more information." with a location pin icon.

: [Demo geocode.xyz]

In case internet does not work for the demo

GeoCode.xyz

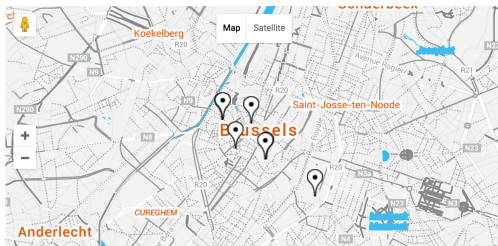
API

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

1

Location

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-Gorikspein, 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

The screenshot shows the GeoCode.xyz website interface. At the top, the logo "GeoCode.xyz" is visible. Below it, a search bar contains the text "Bruxelles/Brussel - Brussels - Brussels". A button labeled "Reprocess text and D" is visible. A map of Brussels is shown in the background. A white pop-up box is overlaid on the map, containing the following text:

Found 7 locations in this line:

Brussels³/Brussels³ - Brussels³ 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) at one of the many bars on Place St-Géry/Sint-Gorikspein, or max out your credit card on the trendy Rue³ Antoine³ Dansaert³/Antoine³ Dansaertstraat

Below the text, a list of 7 locations is shown with their confidence scores:

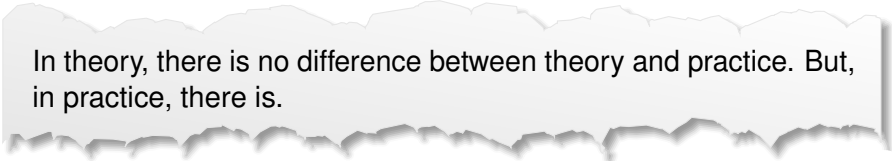
1. ANTOINE DANSARTSTRAAT, BRUSSELS, BE (Confidence: 0.8)
2. Brussels, BE (Confidence: 0.7)
3. ANTOINE DANSART RUE, BRUSSELS, BE (Confidence: 0.7)
4. SINT-GORIKSPEIN, BRUSSELS, BE (Confidence: 0.6)
5. SINT-KATELIJNEPLEIN, BRUSSELS, BE (Confidence: 0.4)
6. GROTE MARKT, BRUSSELS, BE (Confidence: 0.2)
7. 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.

That is where you come in!

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 twitter: [@geolytica](https://twitter.com/geolytica)
PS. One more thing. The core module is 47355 lines of Perl code.