

Mapping World Literature.



Yet Another Book Analysis Project (YABAP)

State of Digitized Literature

In August 2010, Google blogged that there were 129,864,880 books in the world. They were going to scan them all.

A Moonshot that missed the target by a lot

approx 25 million in Google Books, before lawsuits froze 50 or 60 petabytes of words on disk

Authors Guild v. Google. Google it!

Similar Projects

Hathi Trust Digital Library

The Open Content Alliance (OCA)

Project Gutenberg

... and many more - Copyright Infringement looms

Main Goals

Analyze all available books for mentions of locations/place names

Organize books by location

Let anyone add their books to the map & grow the project

Improve geoparsing quality on <https://geocode.xyz>. Avoid getting sued.

Main Motivation

Answer questions such as:

Which books talk about places around X,Y? In what context?

What is the most popular setting for a given type of literature?

What is the connection of literature to the physical world?

Where it's at? Map the Unexplored World of literature

Features

Language Independent.

Accurate.

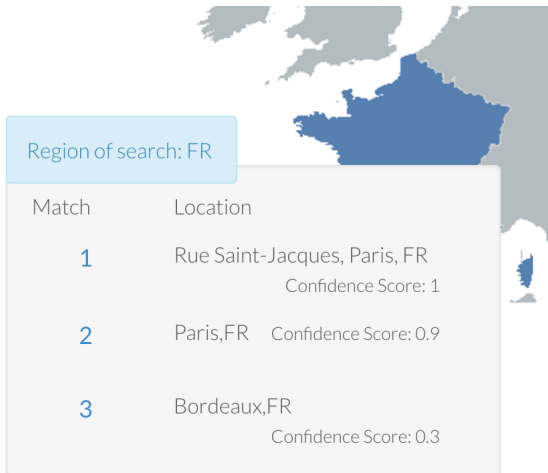
Fast.

Methodology

Books => paragraphs => Geo::Parser::Text

“Why, my dear boy, when a man has been proscribed by the mountaineers, has escaped from Paris in a hay-cart, been hunted over the plains of Bordeaux by Robespierre’s bloodhounds, he becomes accustomed to most things. But go on, what about the club in the Rue Saint-Jacques?” “Why, they induced General Quesnel to go there, and General Quesnel, who quitted his own house at nine o’clock in the evening, was found the next day in the Seine.”

Methodology



redo

Comparison to other NLP Systems

1	'Why, my dear boy, when a man has been proscribed by the mountaineers, has escaped from <u>Paris</u> in a hay-cart, been hunted over the plains of <u>Bordeaux</u> by <u>Robespierre's</u> bloodhounds, he becomes accustomed to most things.
2	But go on, what about the club in the <u>Rue Saint-Jacques</u> ?"
3	'Why, they induced General <u>Quesnel</u> to go there, and General <u>Quesnel</u> , who <u>quitted</u> his own house at <u>nine o'clock in the evening</u> , was found the next <u>day</u> in the Seine.'

All NLP systems have flaws - this is the state of the art Stanford CoreNLP. (Rue Saint-Jacques is an Address, not an Organization)

Methodology

There is also *Rue Saint-Jacques Bordeaux*. But the context is *Paris*

Why, my dear boy, when a man has been proscribed by the mountaineers, has escaped from Paris¹ in a hay-cart, been hunted over the plains of Bordeaux by Robespierre's bloodhounds, he becomes accustomed to most things. But go on, what about the club in the Rue Saint-Jacques¹? "Why, they induced General Quesnel to go there, and General Quesnel, who quitted his own house at nine o'clock in the evening, was found the next day in the Seine.

1. RUE SAINT-JACQUES, PARIS, FR (Confidence: 1.0)
2. Paris,FR (Confidence: 0.3)
3. Bordeaux,FR (Confidence: 0.1)



Compute the highest confidence based on knowledge from prior iterations

The Devil is in the Context

The location context(s) of a book - is crucial to geoparsing

Location is an important detail of a story - it provides context

How the software determines context

Identify every word that contributes to location context. Then every word combination. Their weights. Recalc. and so on

Say if "Paris" and "Seine" are mentioned throughout the book that is an important signal

Context evolves throughout the book - setting may move from one location to another.

Applying Machine Learning

'Learn' by comparing previous iterations of book geoparsing

Say, the word 'This' is a city in France. All English language books share this 'location' (pretty much)

Future iterations will improve the model by lowering the location score of 'This' when there is a weak score for 'France'

The more books we scan the better we become at geoparsing. The more you use `Geo::Parser::Text` the better it becomes.

Re-scanning books to improve geoparsing - Genetic Algorithm Approach

Based on a genetic Algo implementation for the Traveling Salesman Problem I wrote a few years back with a student of mine: <http://www.cpan.org/authors/id/E/ER/ERUCI/tsp.pl>

The fitness function mainly optimizes around finding the minimum number of location overlaps between different books in different contexts.

Also, certain location mentions are more likely together (say "Vlora" and "Katmandu" (unlikely), but "Vlora" and "Tirana" (likely)

OpenStreetMap.org, OpenAddresses.io, geonames.org

Book Data

Project Gutenberg

<https://FictionPad.com> , you, etc...

Software - The ease of building a Geoparser with Perl

Most of it already written on CPAN.

My task consists of putting it all together.

At <https://metacpan.org/pod/Geo::Parser::Text>

Perl is paying my bills for hobbies such as this one

Geocode.xyz is free with throttled access.

Large companies pay to get their own xyz server on the cloud.

Use cases from my client base: Mapping, Search Indexing, Geo-situational awareness, Geospatial Analytics, Summarization, Social Media Monitoring, ...

About the books analyzed so far



49,902 books scanned; 12,936,586 more to go.

See <http://books.geocode.xyz> for the latest count

To Do

API access to all geoparsed book data so that people can build apps

Add other named entity data such as dates, points of interest, people's names, organizations,

Improve the model and run more iterations

Repeat

Next - Song Lyrics

There's a destination a little up the road

From the habitations and the towns we know

A place we saw the lights turn low

Jig-saw jazz and the get-fresh flow

If TIME, then DEMO <https://books.geocode.xyz>

About me and this project.

Ervin Ruci - I'm a developer, I've got problems (I code solutions)

Twitter - <https://twitter.com/geolytica>

API - <https://geocode.xyz>