interested in running a geocoder comparison analysis for the Portland, Oregon region

13 messages

Madeline Steele <madeline.steele@gmail.com> Tue, Jun 7, 2016 at 4:47 PM
To: geodata@geocoder.ca

Hello Mr. Ruci,

The transit organization where I work is currently exploring different options for geocoding. We're hoping to generate accuracy rates for different geocoders that are specific to our area. While researching this topic, I found your presentation from FOSDEM as well as your fourscare-based test data on github. I looked around a bit, but couldn't locate the tool you used to actually run the comparisons (though I may have missed it). I'm wondering it there is any chance you'd be willing to share your tool for generating these scores?

Thanks very much for considering, and congrats on the good outcome of the lawsuit!

Madeline

Ervin Ruci <geodata@geocoder.ca> Wed, Jun 8, 2016 at 5:23 AM
To: Madeline Steele <madeline.steele@gmail.com>

Hi Madeline,

The first step to building a comparison geocoding engine, is to find “ground truth”, geocoded locations that you know are correct. We use foursquare open api and select those locations that have many checkins.

You may download the Oregon data we have obtained from a recent crawl here: http://geocoder.ca/onetimedownload/oregon.sql.gz

Then use a multi-geocoding platform like, http://geocoder.readthedocs.io/ , to test various geocoders. You may use either the command line interface or via a python script. Geocoder.ca usage is described here: http://geocoder.readthedocs.io/providers/Geocoder-ca.html (eg, `geocode '114 Laneda Ave, Manzanita, OR 97130' --provider geolytica`)

Then, store the results, and compare the returned lat,lon with the stored value. There are many libraries implementing distance calculations between two lat,lon points, if you wish I can send you our code.

Regards,
Ervin.

ps. if you run into throttling problems, send me your ip address and I will whitelist it.
[Quoted text hidden]

Madeline Steele <madeline.steele@gmail.com> Wed, Jun 8, 2016 at 10:29 AM
To: "Humphries, Grant" <HumphriG@trimet.org>

Hi Grant,

I've been researching how to properly test geocoders and this guy Ervin gave me some great tips and even a test dataset for Oregon. I could use some help getting this up and running though - hopefully we'll have some time to work on this during our meeting tomorrow.

Thanks,

Madeline
[Quoted text hidden]

Humphries, Grant <HumphriG@trimet.org> Wed, Jun 8, 2016 at 10:57 AM
To: Madeline Steele <madeline.steele@gmail.com>

Sounds good, we can definitely go over this tomorrow.

From: Madeline Steele [mailto:madeline.steele@gmail.com]
Sent: Wednesday, June 08, 2016 10:29 AM
To: Humphries, Grant
Subject: Fwd: interested in running a geocoder comparison analysis for the Portland, Oregon region

[Quoted text hidden]

Madeline Steele <madeline.steele@gmail.com> Wed, Jun 8, 2016 at 11:31 AM
To: Ervin Ruci <geodata@geocoder.ca>

Hello Ervin,

Thank you so much for sending the Oregon data and providing these links. This is extremely helpful and much appreciated. I may have a further question or two for you, but will try to get this going on my own first.

Best regards,

Madeline
[Quoted text hidden]

Ervin Ruci <geodata@geocoder.ca> Fri, Jun 10, 2016 at 8:27 AM
To: Madeline Steele <madeline.steele@gmail.com>

Hi Madeline,

The “Real” test of geocoders however, is when you compare their output from “dirty” or “partial” addresses. With nicely formatted addresses you will get over 90% match from pretty much all of them.

Let me know if you’d like such a sample to test with,

Regards,
Ervin.
[Quoted text hidden]

Madeline Steele <madeline.steele@gmail.com> Fri, Jun 10, 2016 at 9:03 AM
To: Ervin Ruci <geodata@geocoder.ca>

Hi Ervin,

If you have addresses like that for Oregon and it's not too much trouble for you, then yes, I'd very much appreciate getting such a sample!

Thank you,

Madeline
[Quoted text hidden]

Madeline Steele <madeline.steele@gmail.com> Fri, Jun 10, 2016 at 10:32 AM
To: Ervin Ruci <geodata@geocoder.ca>

I'd also be interested in seeing your code for comparing results with known lat/lon if it's not too much trouble (esp. if they're in Python).

Thanks again for your assistance!

Madeline
[Quoted text hidden]

Ervin Ruci <geodata@geocoder.ca> Fri, Jun 10, 2016 at 6:58 PM
To: Madeline Steele <madeline.steele@gmail.com>

I have a script written in perl. Will that help?
[Quoted text hidden]

Madeline Steele <madeline.steele@gmail.com> Mon, Jun 13, 2016 at 8:45 AM
To: Ervin Ruci <geodata@geocoder.ca>

Hi Ervin,

I have my own script going in python so I won't use it directly, but I'm interested in ready your logic for converting distance between true and geocoded lat/lon to scores. I'm not sure if perl is very readable, but I know a perl expert here who can help me make sense of it if it isn't clear to me. So, in short, yes, I think it would be helpful if you don't mind sending.

Thanks much,

Madeline
[Quoted text hidden]

Ervin Ruci <geodata@geocoder.ca> Tue, Jun 14, 2016 at 9:51 AM
To: Madeline Steele <madeline.steele@gmail.com>

Hi Madeline,

There are basically two files. The first runs a geocoding process against two different providers, then saves the results to a file. The second one analyzes that file and prints out some statistics/analysis. (I'm attaching them both, along with a sample run text file and analysis of geocoder.ca vs mapbox for addresses in Oregon)

The first file is called: geogeotestmapbox.pl
The analysis file is : analysegeo.pl
data file: geogeomapbx.txt

Regards,
Ervin.

sample run:
perl analysegeo.pl

Geocoder.ca : 39 out of 45 (accurate within 500m)
MapBox : 38 out of 45 (accurate within 500m)
Geocoder.ca More accurate: 20 times
MapBox More accurate: 20 times
The Same: 5
Coverage:
Geocoder.ca 45 out of 45
MapBox 44 out of 45

> On Jun 13, 2016, at 11:45 AM, Madeline Steele <madeline.steele@gmail.com> wrote:
>
> Hi Ervin,

I have my own script going in python so I won't use it directly, but I'm interested in ready your logic for converting distance between true and geocoded lat/lon to scores. I'm not sure if perl is very readable, but I know a perl expert here who can help me make sense of it if it isn't clear to me. So, in short, yes, I think it would be helpful if you don't mind sending.

Thanks much,

Madeline

On Fri, Jun 10, 2016 at 6:58 PM, Ervin Ruci <geodata@geocoder.ca> wrote:

> I have a script written in perl. Will that help?
>
> > On Jun 10, 2016, at 1:32 PM, Madeline Steele <madeline.steele@gmail.com> wrote:
> >
> > I'd also be interested in seeing your code for comparing results with known lat/lon if it's not too much trouble (esp. if they're in Python).
> >
> > Thanks again for your assistance!
> >
> > Madeline
> >
> > On Fri, Jun 10, 2016 at 9:03 AM, Madeline Steele <madeline.steele@gmail.com> wrote:
> >> Hi Ervin,
> >>
> >> If you have addresses like that for Oregon and it's not too much trouble for you, then yes, I'd very much appreciate getting such a sample!
> >>
> >> Thank you,
> >>
> >> Madeline
> >>
> >> On Fri, Jun 10, 2016 at 8:27 AM, Ervin Ruci <geodata@geocoder.ca> wrote:
> >>> Hi Madeline,
> >>>
> >>> The “Real” test of geocoders however, is when you compare their output from “dirty” or “partial” addresses. With nicely formatted addresses you will get over 90% match from pretty much all of them.
> >>>
> >>> Let me know if you’d like such a sample to test with,
> >>>
> >>> Regards,
> >>> Ervin.
> >>>
> >>>
> >>>> On Jun 8, 2016, at 2:31 PM, Madeline Steele <madeline.steele@gmail.com> wrote:
> >>>>
> >>>> Hello Ervin,
> >>>>
> >>>> Thank you so much for sending the Oregon data and providing these links. This is extremely helpful and much appreciated. I may have a further question or two for you, but will try to get this going on my own first.
> >>>>
> >>>> Best regards,
> >>>>
> >>>> Madeline
> >>>>
> >>>> On Wed, Jun 8, 2016 at 5:23 AM, Ervin Ruci <geodata@geocoder.ca> wrote:
> >>>>> Hi Madeline,

The first step to building a comparison geocoding engine, is to find “ground truth”, geocoded locations that you know are correct. We use foursquare open api and select those locations that have many checkins.

You may download the Oregon data we have obtained from a recent crawl here: http://geocoder.ca/onetimedownload/oregon.sql.gz

Then use a multi-geocoding platform like, http://geocoder.readthedocs.io/ , to test various geocoders. You may use either the command line interface or via a python script. Geocoder.ca usage is described here: http://geocoder.readthedocs.io/providers/Geocoder-ca.html (eg, `geocode '114 Laneda Ave, Manzanita, OR 97130' --provider geolytica )`

Then, store the results, and compare the returned lat,lon with the stored value. There are many libraries implementing distance calculations between two lat,lon points, if you wish I can send you our code.

Regards,
Ervin.

ps. if you run into throttling problems, send me your ip address and I will whitelist it.

> On Jun 7, 2016, at 7:47 PM, Madeline Steele <madeline.steele@gmail.com> wrote:
>
> Hello Mr. Ruci,
>
> The transit organization where I work is currently exploring different options for geocoding. We're hoping to generate accuracy rates for different geocoders that are specific to our area. While researching this topic, I found your presentation from FOSDEM as well as your fourscare-based test data on github. I looked around a bit, but couldn't locate the tool you used to actually run the comparisons (though I may have missed it). I'm wondering it there is any chance you'd be willing to share your tool for generating these scores?
>
> Thanks very much for considering, and congrats on the good outcome of the lawsuit!
>
> Madeline

3 attachments

geogeomapbx.txt
6K

geogeotestmapbox.pl
12K

analysegeo.pl
2K

Madeline Steele <madeline.steele@gmail.com> Tue, Jun 14, 2016 at 5:42 PM
To: Ervin Ruci <geodata@geocoder.ca>

Thank you Ervin,

I'm so appreciative of your generosity with your time, data, and code! This will be a great help to us as we try to identify the best geocoder for our needs

All the best,

Madeline

[Quoted text hidden]

> [Quoted text hidden]
> Regards,
> Ervin.
>
> sample run:
> perl analysegeo.pl
>
> Geocoder.ca : 39 out of 45 (accurate within 500m)
> MapBox : 38 out of 45 (accurate within 500m)
> Geocoder.ca More accurate: 20 times
> MapBox More accurate: 20 times
> The Same: 5
> Coverage:
> Geocoder.ca 45 out of 45
> MapBox 44 out of 45
> [Quoted text hidden]

Madeline Steele <madeline.steele@gmail.com> Tue, Jun 14, 2016 at 5:42 PM
To: steelem@trimet.org

---------- Forwarded message ----------
From: **Madeline Steele** <madeline.steele@gmail.com>
Date: Tue, Jun 14, 2016 at 5:42 PM
Subject: Re: interested in running a geocoder comparison analysis for the Portland, Oregon region
[Quoted text hidden]