Project Topic: OSM, Photos, and Tours

The OpenStreetMap *(https://www.openstreetmap.org/)* project collects community-provided map data that is free to use. The full data dump, known as planet.osm *(https://wiki.openstreetmap.org/wiki/Downloading_data)* , provides all of the data from their maps in a slightly ugly XML format, ready for analysis.

The Idea

The OSM data set has a huge collection of things you might have seen while walking around the city: Canada Place *(https://www.openstreetmap.org/way/223635729)* , The Steam Clock *(https://www.openstreetmap.org/node/703754423)* , a bench *(https://www.openstreetmap.org/node/3789883495)* , etc. Maybe you have walked by these and not even noticed. Wouldn't it be nice if they were pointed out to you?

We have previously worked with GPX data: the file produced by fitness trackers, GPS systems, or anything else that tracks movements with GPS *(https://en.wikipedia.org/wiki/Global_Positioning_System)* signals (or related technology). The problem with GPX is that you have to create it. You are more likely to naturally find geographic information in photographs: Exif data in JPEG images *(https://en.wikipedia.org/wiki/Exif)* can contain latitude and longitude data as well and most phones automatically add it

The challenge: take a collection of geotagged photos representing my walk/tour/vacation, give me a tour of the things I *should* have seen, or try to guess what is in the photos.

Provided Data

I have downloaded the planet.osm data and done some work to turn the monster of an XML file into data that's more usable, `amenities-vancouver.json.gz` in the provided data and code. The OSM data was turned into what I gave you with the code in that archive:

1. Turned the monolithic XML file into a split files with its top-level elements one-per-line so they can sensibly be approached with Spark. (With `dissassemble-osm.py`, producing the data in `/courses/datasets/openstreetmaps` on our cluster's HDFS.)
2. Processed the fragmented XML, keeping only nodes *(https://wiki.openstreetmap.org/wiki/OSM_XML)* that are an “amenity”, and saving as a more reasonable JSON format. (With `osm-amenities.py` as a Spark job on the cluster.)
3. Extracted only data that was roughly within Greater Vancouver. (With `just-vancouver.py` on the cluster.)

If you would like to work with a different subset of the data, you can modify the code and repeat steps 2 or 3. You'd have to be insane to repeat step 1.

The data set is in JSON format, with fields for latitude, longitude, timestamp (when the node was edited), the amenity type (like “restaurant”, “bench”, “pharmacy”, etc), the name (like “White Spot”, often missing), and a dictionary of any other tags *(https://wiki.openstreetmap.org/wiki/Tags)* in the entry.

In Pandas, the tags field will be loaded as a Python dictionary (mapping keys to values). In Spark, it will be loaded as a `MapType` of string to string (see `just-vancouver.py` for a schema).

Other Data

The problem with the OSM data is probably that it's too complete. I don't need to know about every park bench that I walked by. To make this useful, you have to find the *interesting* things I passed.

The result: **the provided data is likely insufficient** to get good results. You might be able to combine with WikiData information (e.g. the Steam Clock has tag “wikidata” with value “Q477663”, referring to its wikidata entry *(https://www.wikidata.org/wiki/Q477663)*). Or with Wikipedia data (e.g. the Steam Clock has a “wikipedia” tag referring to its Wikipedia entry *(https://en.wikipedia.org/wiki/en:Steam%20clock?uselang=en#Gastown_steam_clock)*).

You may be able to do some clever processing on the OSM data to guess what attractions would be interesting to the user. It's possible that you can apply some heuristic to find interesting points (more-complete entries obviously get more attention) or exclude boring ones (like park benches and infrastructure).

Notes

Remember this is supposed to be a larger-scale and more independent project. If your plan is to just re-purpose the idea from the exercises and do something like “find nearby stuff”, that's not much of a project and your mark will probably reflect that. We expect to see more creativity here to attack a more open-ended problem like this.

Other Questions

There are thousands of interesting data science questions that could be posed from the OSM data (and other data sets that could be integrated with it). You are, of course, free to adapt as you like.

A few things I see that you could approach with this data:

- If I was planning a tour of the city (by walking/biking/driving), where should I go? Are there paths that take me past an interesting variety of things?
- I feel like there are some parts of the city with more chain restaurants (e.g. McDonand's or White Spot franchises, not independently-owned places): is that true? Is there some way to find the chain places automatically and visualize their density relative to non-chains?
- If I was going to choose a hotel (or AirBnb), where should it be? What places have good amenities nearby?
- Any of these could be turned into “big data” problems by working with the global version of the data set. You can recreate that data as described above.

Updated Wed April 07 2021, 09:16 by ggbaker.