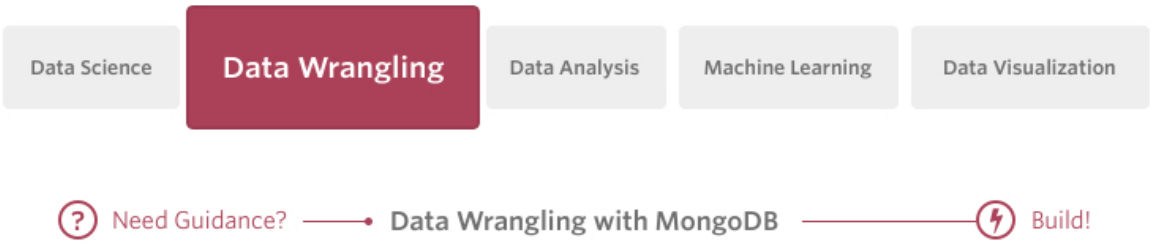
Data Wrangle Openstreetmaps Data



Project Overview

You will choose any area of the world in https://www.openstreetmap.org and use data munging techniques, such as assessing the quality of the data for validity, accuracy, completeness, consistency and uniformity, to clean the OpenStreetMap data for a part of the world that you care about.

Prepare for this project with: Data Wrangling with Mongo DB.

Note

If you have successfully completed the project for the Data Wrangling with MongoDB course in the past (which entails having graduated from the course and having access to your course certificate), simply email us at dataanalyst-project@udacity.com with your passing evaluation and we'll give you credit for this project.

Why this Project?

What's so hard about retrieving data from databases or various files formats? You grab some data from this file and that database, clean it up, merge it, and then feed it into your state of the art, deep learning algorithm … Right?

But the reality is this -- anyone who has worked with data extensively knows it is an absolute nightmare to get data from different data sources to play well with each other.

And this project will teach you all of the skills you need to deal with even the most nightmarish data wrangling scenarios.

What will I learn?

After completing the project, you will be able to:

- Assess the quality of the data for validity, accuracy, completeness, consistency and uniformity
- Parsing and gather data from popular file formats such as json, xml, csv, html
- Process data from many files and very large files that can be cleaned with spreadsheet programs
- Learn how to store, query, and aggregate data using MongoDB

Why is this Important to my Career?

As this New York Times article points out, the less heralded part of doing data science is manually collecting and cleaning data so it can be easily explored and analyzed later. Or otherwise known as "data wrangling" or "data munging" in the data science community.

Though not as glamorous as building cool machine learning models, data wrangling is a task that data scientists can spend up to 50-80% of their time doing according to many practicing data analyst and data scientists.

How do I Complete this Project?

This project is connected to the Data Wrangling with MongoDB course, but depending on your background knowledge of data wrangling, you may not need to take the whole thing to complete this project.

Here's what you should do:

Step One - Finish Lesson 6

Make sure all Lesson 6 programming exercises are solved correctly.

Step Two - Review the Rubric and Sample Project

The Project Rubric will be used to evaluate your project. It will need to Meet Specifications for all the criteria listed. The Sample Project is an example of what your final report could look like.

Step Three - Choose Your Map Area

Choose any area of the world from https://www.openstreetmap.org , and download a XML OSM dataset. The dataset should be at least 50MB in size (uncompressed). We recommend using one of following methods of downloading a dataset:

- Download a preselected metro area from Map Zen (Note that data obtained from Map Zen is compressed and will usually expand to sizes that meet project requirements.)
- Use the Overpass API to download a custom square area. Explanation of the syntax can found in the wiki . In general you will want to use the following query:`(node(minimum_latitude, minimum_longitude, maximum_latitude, maximum_longitude);<;);out meta;`e.g. `(node(51.249,7.148,51.251,7.152);<;);out meta;` the `meta` option is included so the elements contain timestamp and user information. You can use the Open Street Map Export Tool to find the coordinates of your bounding

box. Note: You will not be able to use the Export Tool to actually download the data, the area required for this project is too large.

Step Four - Process your Dataset

Thoroughly audit and clean your dataset, converting it from XML to JSON format. It is recommended that you start with the Lesson 6 exercises and modify them to suit your chosen data set. As you unravel the data, take note of problems encountered along the way as well as issues with the dataset. You are going to need these when you write your project report. Finally, import the clean JSON file into a MongoDB database and run some queries against it.

Step Five - Document your Work

Create a document (pdf, html) that directly addresses the following sections from the Project Rubric .

- Problems encountered in your map
- Overview of the Data
- Other ideas about the datasets

Try to include snippets of code and problematic tags (see Sample Project) and visualizations in your report if they are applicable.

Evaluation

A Udacity evaluator will review and check your completion of problem set 6 in Data Wrangling with MongoDB, as well as your answers to the questions listed in the rubric.

Be sure to complete all of questions in the problem set and the rubric before submitting your project.

Rubric

Your project will be evaluated by a Udacity reviewer according to this project rubric. Be sure to review it thoroughly before you submit. Your "project meets specifications" if it meets specifications in all the criteria

Submission

Ready to submit your project? Collect the following files:

1. A pdf document containing your answers to the rubric questions. This file should document your data wrangling process.
2. Your Python code for Lesson 6 quizzes, as well as any additional code you used in auditing and cleaning your dataset for the final project.
3. A text file containing a link to the map position you wrangled in your project, a short description of the area and a reason for your choice.
4. An .osm file containing a sample part of the map region you used (around 1 - 10 MB in size). See instructor notes.
5. A text file containing a list of Web sites, books, forums, blog posts, github repositories etc that you referred to or used in this submission (Add N/A if you did not use such resources).

Then go back to the portal, click on the project, and follow the instructions to submit:

- If you want to submit your files through a "Link to Project", upload your project files onto Github and send us the link.
- If you instead want to submit your files through "Upload a Zip", compress your project directory, and submit that zip file.

It can take us up to 2 weeks to grade the project so keep checking back for updates.

If you are having any problems submitting your project or wish to check on the status of your submission, please email us at dataanalyst-project@udacity.com.