Machine Learning Engineering Career Track

Phase 1 Step 3. Data Collection

Summary

Time Estimate: 2 - 4 Hours

To kick-start your capstone project, the first thing you'll need to do is collect your data. In some cases, it can be as simple as downloading a dataset in a zip file or a tarball. In other cases, it can require extracting data using a publicly available API or scraping a website. We urge you to work closely with your mentor to ensure that the data collection process is not too onerous for a capstone project. Also, **if your data collection requires you to write code, it's important that you start early.**

Below, you'll find a few great sources for large datasets that are appropriate for this course (you'll also find these resources listed in the capstone project guidelines.)

- fast.ai research datasets collection
- Google dataset search
- AWS open datasets repository
- Uber Movement
- Yelp dataset

In addition to the resources listed above, you can also explore datasets from Quandl, US Government Open Data, UCI Machine Learning Repository, and Kaggle competitions, or anywhere else you like.

Remember: Your dataset should have at least 15K-20K samples **at a minimum**. We'd like to see you build large-scale applications for this course so we encourage you to work with larger datasets, which means choosing a dataset that's at least 8GB in size or has at least 1 million samples.

Project Submission Steps

Please submit a link to your Github repository, which should contain the following:

1. Code for how you collected the data if applicable
2. The actual dataset: if your dataset is small enough to fit in a CSV, then feel free to include it in the repository. If it's a big dataset or has a lot of binary files (graphics, audio), consider using the Git Large File Storage extension.

This step of your capstone project will be evaluated using this rubric.