Unit 6: Data Collection

Overview and Key Concepts

Lesson 45

Derek Ruths

Overview of unit

Objectives:

- Understand why data collection is a time-consuming part of a project
- Get experience working with API collection
- Become familiar with standard data storage solutions: JSON, csv, and sqlite
- Understand how bias arises and why it’s hard to detect

1. Overview
2. API-based collection
3. JSON
4. Reddit data collection
5. Scraping
6. Sampling
7. Homework 5

Lesson overview

Objectives

- Understand the goals and stages of data collection
- Understand the kinds of data collection
- Understand how time-consuming data collection is

Outline

- What happens during data collection?
- Kinds of data
- The 80/20 rule

The role of data collection

The purpose of the data collection phase is to prepare all raw data for annotation and analysis.

- Remove all imperfections
- Organize in a way that's conducive to analysis

result in data we're confident in

The steps in data collection

1. Collect data

2. Clean data

removing errors/problematic parts
or become aware of them

3. Organize data

↳ nothing or creating a DB Schema or othe

Kinds of data

- Found data – data that was produced for other purposes
  - Social media data
  - Pre-existing surveys
  - Government reports

- Designed data – data created specifically for this project
  - Survey data from surveys run for this project
  - Experimental results

data annotations are designed data

The 80/20 rule

- 80% of time will be spent in data collection
- 20% on everything else

Found data – remove everything that isn't what you need

Designed data – create the data from scratch

both time consuming

Lesson wrap-up

Takeaways

- Data collection is demanding and tricky
- Always allocate enough time for data collection

Up next

- The philosophy of data science coding