Unit 6: Data Collection

JSON

Lesson 47

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. Reddit data collection
4. JSON
5. Scraping
6. Sampling
7. Homework 5

Lesson overview

Objectives

- Be comfortable reading the JSON file format
- Understand the pros and cons for the data format

Outline

- Why JSON?
- The JSON format
- How to read the JSON format
- Why JSON isn’t always the right answer

Why JSON?

CSV/TSV - tabular data (eg clean/simple entities w/ data)

Json - more messy data

The JSON format

numbers

str

bool

lists [, ,]

dict { "k": v, ... }

```
{
  "name": v
  "friends": [ {
        friend
     }
     { friend
     }
  ]
}
```

How to read the JSON format

- import json
- json.load / loads
- json.dump / dumps

load from file

load from python string

eg

```
src = """
{
  ...
}
"""

data = json.loads(src)
```

Issues with JSON

- Bloat

- Hard to detect inconsistency

Lesson wrap-up

Takeaways

- JSON as a transport and storage protocol is great … some of the time

Up next

- Using the Reddit API