Unit 6: Data Collection

API-based Data Collection

Lesson 46

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 how API-based data collection works
- Know some of the key pitfalls with API-based data collection

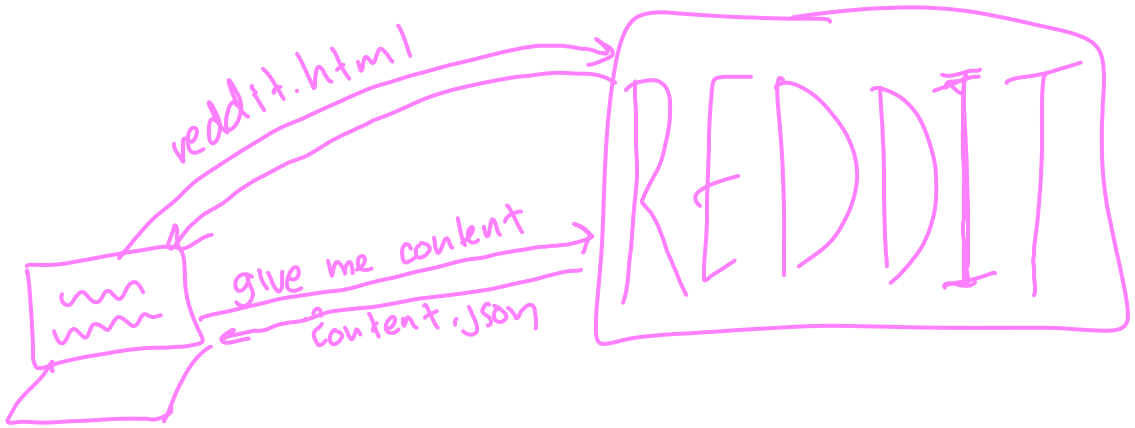
Outline

- What is API data collection?
- Limitations
- Pitfalls

What is a web API?

Application programmer Interface

how a web app gets data from another server



reddit url .json gives you json dump

Web API data collection process

```
import json
data = json.load(open("reddit-page.json", "r"))
print(data["data"]["children"][0]["data"]["title"])

for post in data["data"]["children"]:
    print(post["data"]["title"])
```

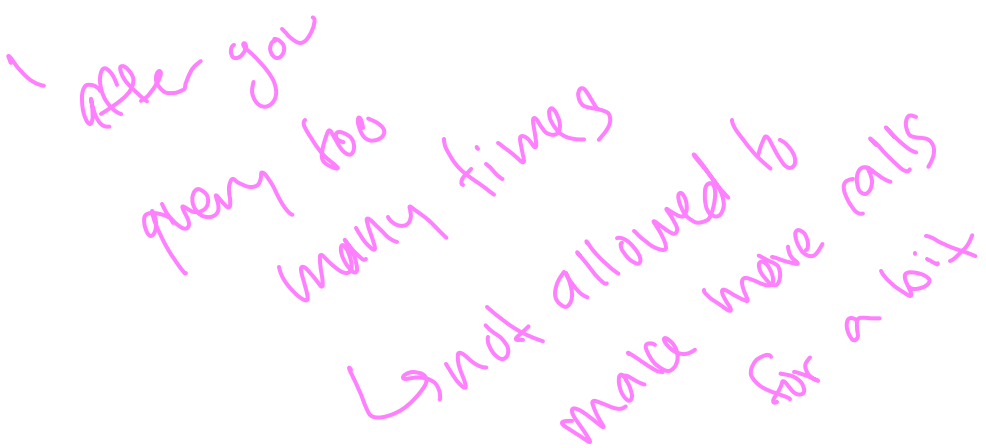
Limitations of APIs

- Do they exist?
- Do they give you what you need?
- Query speed

Challenges & Pitfalls

- Authentication

- Blacklisting



Python library: requests

Lesson wrap-up

Takeaways

- Web APIs are an ideal way to get data… when they exist

Up next

- JSON