Unit 6: Data Collection

Collecting data from Reddit

Lesson 48

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

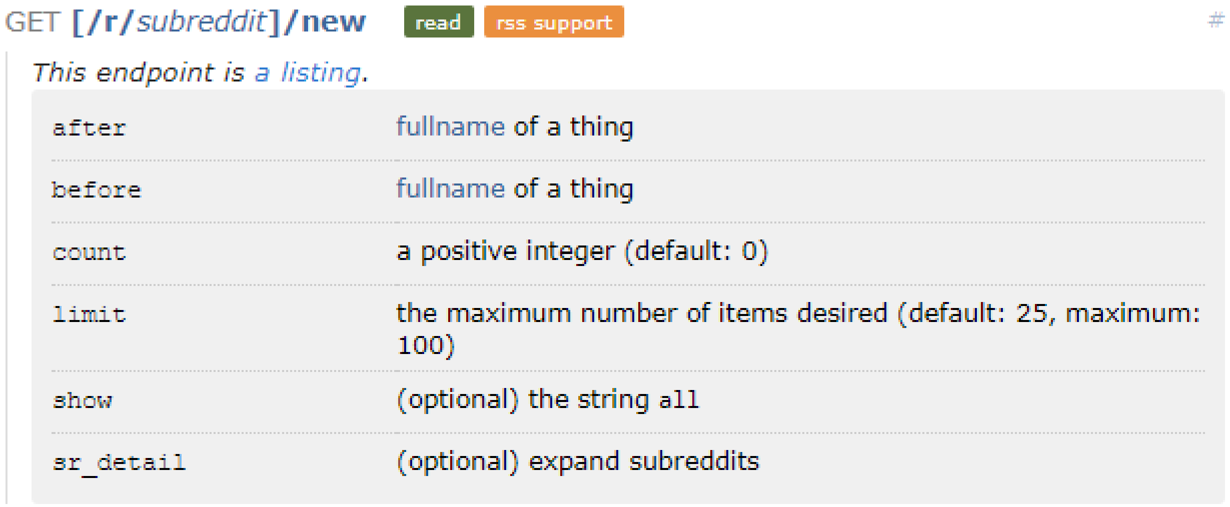
- See how API queries work (in practice)
- Get some experience collecting data from a social media platform

Outline

- “Raw” API queries
- Library-based API queries

How to get data from Reddit

https://www.reddit.com/dev/api/



GET [/r/*subreddit*]/new read rss support #

This endpoint is a listing.

`after`	fullname of a thing
`before`	fullname of a thing
`count`	a positive integer (default: 0)
`limit`	the maximum number of items desired (default: 25, maximum: 100)
`show`	(optional) the string `all`
`sr_detail`	(optional) expand subreddits

Using requests

Lesson wrap-up

Takeaways

- Getting Reddit data is pretty straightforward

Up next

- Scraping