Data Collection for Social Scientists: Leveraging Public APIs with Python

Zach Dickson

MY580

Target Audience: Advanced MSc and PhD students in Social Sciences.

Prerequisites:

- Basic to intermediate proficiency in Python (understanding data types, lists, dictionaries, loops, functions) (see below for resources).
- Familiarity with the Jupyter Notebook or a similar Python IDE (see below for resources).
- A conceptual understanding of social science research methodologies.

Workshop Goal: Equip students with the skills to programmatically access data from public REST APIs, transform this data into analysis-ready datasets, and critically consider dataset design for social science research.

Learning Objectives

By the end of this workshop, students will be able to:

- Understand the fundamentals of APIs and how they can be used in social science research.
- Use Python's `requests` library to interact with REST APIs.
- Fetch, clean, and transform data from the UK House of Commons API.
- Critically evaluate the design of datasets for social science research.
- Apply data cleaning techniques to prepare datasets for analysis.

Workshop Schedule

- Session 1: Foundations & API Interaction (~2 hours)

  - **Module 1: Introduction to APIs** (20 minutes)
    * Welcome and Introduction (5-10 minutes)
    * What are APIs? (5-10 minutes)
    * Why use APIs? (5-10 minutes)

  - **Module 2: Understanding REST APIs & Python's `requests`** (45 mins)
    * URL and Endpoint Basics (10 minutes)
    * HTTP Methods (GET, POST, PUT, DELETE) (10 minutes)
    * Request Parameters and Headers (10 minutes)
    * Response Formats (JSON, XML) (10 minutes)
    * Status Codes and Error Handling (5 minutes)

  - **Module 3: `Requests` Module in Python** (55 minutes)
    * Installing the `requests` library (5 minutes)
    * Making a GET request (10 minutes)
    * Handling JSON responses (10 minutes)
    * Error handling and status codes (10 minutes)
    * Practical exercise: Fetching data from a public API (10 minutes)

Lunch Break (1 hour)

- Session 2: Introduction to the UK House of Commons API (~2 hours)

  - **Module 4: Introduction to the UK House of Commons API** (40 minutes)
    * Overview of the UK House of Commons API (5 minutes)
    * Understanding the API documentation (5 minutes)
    * Authentication
    * Fetching data from the API (10 minutes)
    * *Practical exercise:* Fetching and exploring data from the UK House of Commons API (20 minutes)

  - **Module 5: Data Transformation and Cleaning** (60 minutes)
    * Introduction to data cleaning (5 minutes)
    * Using Python libraries for data manipulation (e.g., pandas) (20 minutes)
    * Structuring data for analysis (10 minutes)
    * Storing and exporting data (10 minutes)
    * *Practical exercise:* Cleaning and transforming data from the UK House of Commons API (15 minutes)

Resources

Installing Python and Jupyter Notebook

To participate in this workshop, you will need to have Python installed on your computer, along with the Jupyter Notebook environment. Python is a versatile programming language that is widely used in data science, and Jupyter Notebooks provide an interactive environment for writing and running Python code.

Python can be installed in various ways, but the most common and user-friendly method is through the Anaconda distribution, which includes Python, Jupyter Notebook, and many useful libraries for data science. If you are comfortable accessing Python through a different IDE or method, feel free to do so.

Video Tutorials for installation:

- Installing Python + Jupyter Notebook on Windows 11
- Installing Python + Jupyter Notebook on MacOS

Intro to Python:

There are many resources available to learn Python. You will not need to be an expert, but you should be comfortable with basic data types (strings, integers, floats), lists, dictionaries, loops, and functions. Here are some resources to get you started. In my experience, the best way to learn is to practice as you go, so I recommend working through some of the exercises in the Python for Everybody course.

There are also many great resources available on YouTube, such as Python Crash Course for Beginners, Python for Data Science, and Learning Python for Data Analysis.