Using LLMs to Code Qualitative Data

Warning: I am new to this!

Andrew Avitabile[1]

[1]University of Virginia

January 30, 2025

Table of Contents

When to Use LLMs?

Researchers have use text-as-data methods for years

Before you set out to use an LLM for coding qualitative data, consider other Natural Language Processing methods:

- Sentiment analyses
- Topic modeling
- Word embeddings

Importantly, consider what information you are giving LLMs (i.e., any personally identifiable information)

- This is new territory for institutional review boards (IRB), so please contact your before starting your project

Download Python and Positron

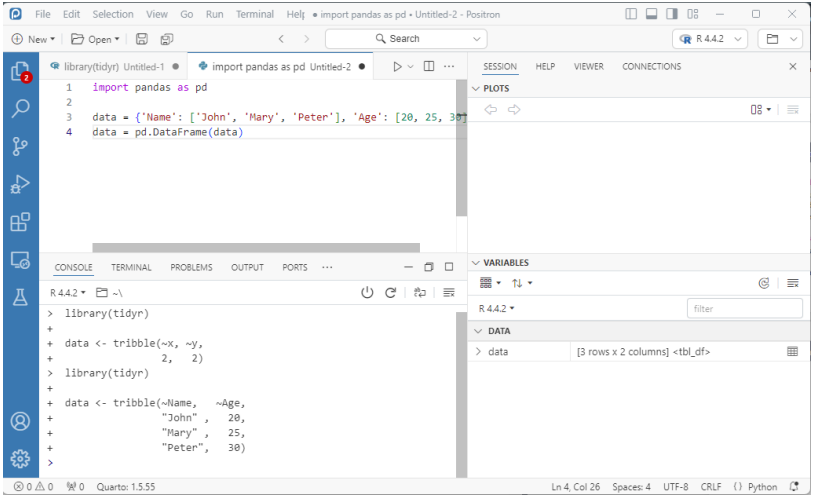
Download Python at this link

- Python is an open source coding language, similar to R. It is the most popular coding language in the world.
- *Disclaimer:* Python's package management system can be really annoying!

Download Positron at this link

- This is the next generation of RStudio. It is called an Integrated Development Environment (IDE).
- Positron is a place to run code in many languages, but specifically R and Python

Quick Walk through of Positron IDE



Quick Intro to APIs

Application Programming Interfaces (API) are rules that allow sets of software to communicate with one another

You can set up an API with OpenAI or Anthropic that allows you to "hit" ChatGPT or Claude with Python code

This allows us to write programs that ask ChatGPT or Claude models to do tasks

Setting up an OpenAI API

Go to this link and set up an account with OpenAI

Note the different models, which have different pricing schemes that frequently change

Get a API key

- This is meant to be kept secret as it allows people to hit an API
- Best practice is to store it in a .env file and not hard coded into your programs
  - This is a program that stores passwords, API keys, etc.
  - Learn more about how to write these files at this link

Tips for Asking LLMs to Code Qualitative Data

Should be an iterative process!

- Imagine you were asking an undergraduate RA to code something from text

Remember to provide context to the model (e.g., "You are a researcher coding…")

You may need to grovel

- Sometimes you need to make it **extremely clear** to get the output you want

Example

I've added a basic example script and dataset on here on GitHub

Let's walk through together!

Feel free to reach out with any questions: yaj3ma@virginia.edu