{reticulate}

PSY-GS 8875 Behavioral Data Science

Overview: Week 1

Readings

- Breiman - 2001 - two cultures (see "articles" on Brightspace or GitHub)

Optional

- none

- Introductions

- Preliminaries

- Generative AI and Retrieval-augmented Generation

- {reticulate}

- Syllabus AI

Introductions

Introductions

1. Preferred name

2. Department

3. What do you want to get out of this course? Any specific methods?

Preliminaries

Preliminaries

- R

- Toolchains?

- GitHub (Desktop)

- Slack?

- NCDS dataset

Core Documents

Syllabus

Schedule

Office hours: alexander.christensen@vanderbilt.edu

Preliminaries

Course Structure

- Technical introduction to method(s)

- Application and activity

- **Activities** will be turned in on Brightspace the following Tuesday by midnight (11:59:59pm)

- **Final project**: application of at least three methods covered in this course to data (preferably your own data!)

Generative AI

- Language generation

- Coding (debugging and generation)

- Image generation

- Video generation

- and more...

- Use it!

- Considered engagement in this course

- Encouraged (and expected) to help you learn and explore materials

- Coding in R will be a significant part of this course

- With generative AI, minimal experience is necessary

- If you've only used the free version of ChatGPT (GPT-3.5), then you might think AI sucks at coding...

- Well, GPT-3.5 does suck at coding

- But newer models (including GPT-4) are much better at coding and are super handy for writing and studying code

- You can "augment" models to make them better too

Retrieval-augmented Generation (RAG)

Retrieval-augmented Generation

Large language models (LLMs) are trained on an enormous amount of text data

- **Pro**: generalizes across many contexts

- **Con**: lacks domain-specific or specialized knowledge

- **Con**: model might "hallucinate" answers because. . .
    - out-of-date information
    - there is no relevant information to retrieve
    - LLMs are sentient and messing with you (not actually. . . yet)

ChatGPT EGA example. . .

ChatGPT EGA example. . .

. . . or use a tool designed for the task: SciSpace

Solution: give the model access to specialized knowledge to let it draw from a more specific context

So... how can we make LLMs better at coding (in R or any other language)?

Solution: give the model access to specialized knowledge to let it draw from a more specific context

So. . . how can we make LLMs better at coding (in R or any other language)?

What about for coding in this course?

{reticulate}

{reticulate}: R package to interface with Python

- seamless integration between R and Python

- do Python without leaving R

- equivalent for Python-to-R: {rpy2}

Syntax

```
import("module")
# import module into environment

module <- import("module")
# assign module to an object (preferred)

module$function
# use function in a module
```

Why bother? Why not learn Python instead?

{reticulate}

Why bother? Why not learn Python instead?

- Most social sciences (and statisticians) use R

- Switching between software is inefficient. . .

- . . . and can lead to reproducibility issues

Syllabus AI

To get some hands-on practice with {reticulate} as well as retrieval-augmented generation, we'll index our syllabus

This Shiny app is going to index this course's syllabus and course schedule as a "quick" look-up resource

Guided Activity. . .

Potential Uses of RAG

- coding help (e.g., R manuals)

- literature search (e.g., relevant references)

- study guides (e.g., develop questions from material)

At Home Activity

- Use one of the following free-to-use models to upload the Syllabus:

    - Perplexity

    - Claude 2

    - Our example's LLAMA-2

- Ask the following questions and submit the responses on Brightspace:

    - Who is the professor for PSY-GS 8875?

    - What is the grading policy?

    - Where can I find data?

Readings for Next Week

Readings

- ESL Chapters: 3.1, 3.2, 4.4, and 4.4.1

- HML Chapters: 4.1-4.5 and 5.1-5.5

Optional

- none