

Intro to Web Scraping and AI-Assisted Programming

Computer Programming for Lawyers, Fall 2024

OCTOBER 15, 2024

Expectations for Next Problem Set

- Starting with this problem set, you are allowed and expected to use generative AI, such as OpenAI's ChatGPT.
- In the header, indicate which AI tool(s) you used.
- **If you use an external chatbot, please also use the "Share chat" button at the top-right to copy and paste the link to your chat transcripts.**
- You may work on this problem set in groups of no more than two. It may be helpful and fun to sit next to a partner as you figure out how to get AI to do what you want it to! If you do this:
 - You must indicate who you worked with in the header.
 - You must each have your own separate interactions with AI.
 - You must submit unique files. That is, no copying and pasting code to each other—only copying and pasting to and from your AI chats.



time_elapsed demo



LLMs are non-deterministic
(random)

What AI is good at

ROUTINE TASKS

Good at creating boilerplate code and achieving simple programming tasks.

PROTOTYPING

AI assists in quickly creating functional code prototypes, reducing development time.

EXPLAINING CODE

AI can explain the function and purpose of code snippets, making it easier to understand complex logic.

STYLE

AI typically produces well-styled, readable code with lots of description.

HANDLING ERRORS

AI can help identify potential bugs, as well as interpret error messages and suggest fixes.

What AI **isn't** as good at

CORRECTNESS

Often produces code that requires human validation for accuracy and security.

MANAGING LARGE TASKS

Large, complex problems need to be broken down into smaller tasks for AI to handle effectively.

BIAS AND FAIRNESS

AI reproduces the biases in its training data, which can impact diversity in programming solutions.

DEEP UNDERSTANDING

AI lacks deep comprehension and might produce inaccurate results without clear context.

AWARENESS

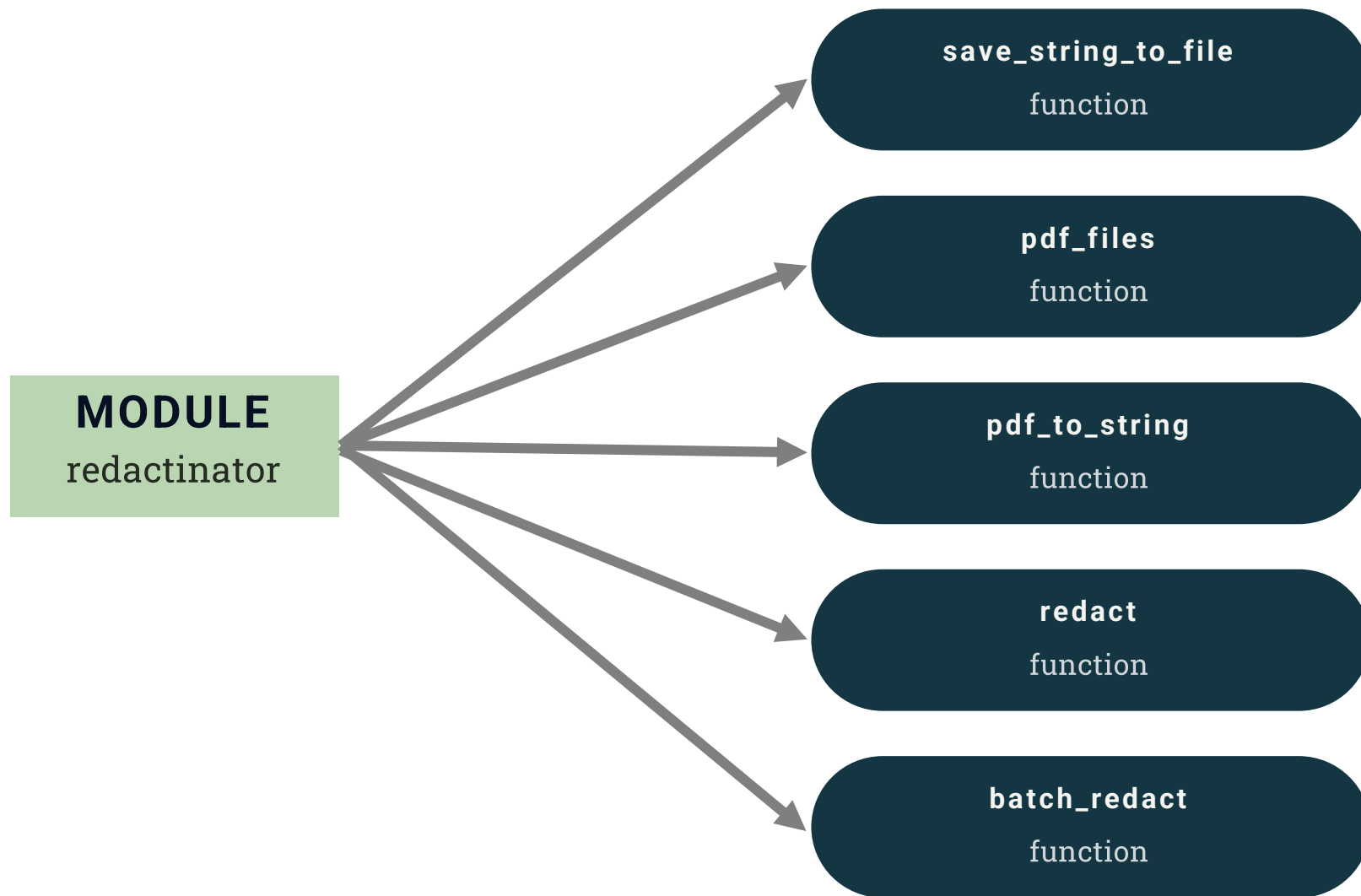
AI doesn't know when it's wrong and often confidently generates incorrect answers.

ATTRIBUTION

AI may give you someone else's code without attributing a source. This can lead to potential ethical or legal issues.



Problem Decomposition





Practice

Write a function signature to randomly assign courtroom roles such as judge, prosecutor, defense attorney, or jury member to a given list of names.

Let Copilot fill in the code.

```
def name_of_function(parameters):  
    """ docstring describing function behavior """
```

Upload the instructions for a past problem set to a LLM, along with your code. Ask it tell you what you did wrong.

You can download the README.md instructions and .ipynb submissions directly from your GitHub repository, or by right clicking on the file name in Codespaces.

Web Scraping

Motivating Web Scraping

- Law firms often need to analyze data "locked" in a website.
- Manual strategies involve copy-and-paste, re-type into Excel, etc. – inefficient for large volumes.
- Web scraping automates downloading of web content.
 - Turns un/semi structured data into structured, usable databases.
 - Reduces repetitive tasks like clicking "Next" or "Save as" multiple times.
- Sample use cases:
 - Download and summarize 150 comments from the FCC's NPRM.
 - Search message boards for potential trade secrets leakers.
 - Review X Corp's filings for acquisition due diligence.
 - Process hundreds of spreadsheets from an FBI FOIA response.



Prerequisite:
Installing External Libraries

Navigating HTML



Inspecting Elements and Understanding HTML Structure

- The webpages you see in your browser are made up of HTML elements.
- To effectively scrape web pages, it's essential to understand how HTML structures content.
- These elements define the structure and content of the page. They are enclosed within tags like `<div>`, `<a>`, `<p>`, `<h1>`, etc.
- **Modern browsers have built-in tools to help inspect the HTML structure.**






Common HTML Tags

- **<div>**: A container for other elements. Often used to group together sections of a webpage.
- **<a>**: Represents a hyperlink, often used to link to other pages or files.
- **<p>**: Defines a paragraph of text.
- **<table>**: Structures data in rows and columns, often used for tabular data.

Common HTML Attributes

HTML tags can have attributes that provide additional information about the element.

- **id**: A unique identifier for an element.
- **class**: Defines a class name that can be used to group elements.
- **href**: The destination URL for a link (<a> tag).
- **src**: The source URL for an image (tag).

 images	2 months ago
 labs	2 days ago
 lecture 	5 days ago
 README.md	2 weeks ago

```
<a title="lecture" class="Link--primary" href="/Computer-Programming-for-Lawyers/Fall-2024/tree/main/lecture">lecture</a>
```