www.imarketing.courses

Doc Prep 101: Reasons Files Don't Work Well (and How to Fix Them)

In this lesson: how to best prepare and organize your documents for chatbot retrieval.

A big mistake people make when uploading docs into AI is assuming the AI can understand them like magic.

On the one hand, they can, but... you can often get substantially better performance out of a doc by pre-processing it.

About Pre-Processing

Pre-processing may involve:

1. Transforming the file type
2. Removing clutter or unnecessary information
3. Adding structure or metadata to improve the AI's understanding

File Types

Large language models like ChatGPT understand most documents based on their text. Indeed, for any inputs other than images, the language model will end up converting it to text to try to understand it.

To get better performance out of your model, the first thing you can do is to convert docs like .DOCX and .PDFs into simpler text formats.

This is often a regular .TXT file.

Some human-readable text files like JSON and CSV are also easily understood by their contained text.

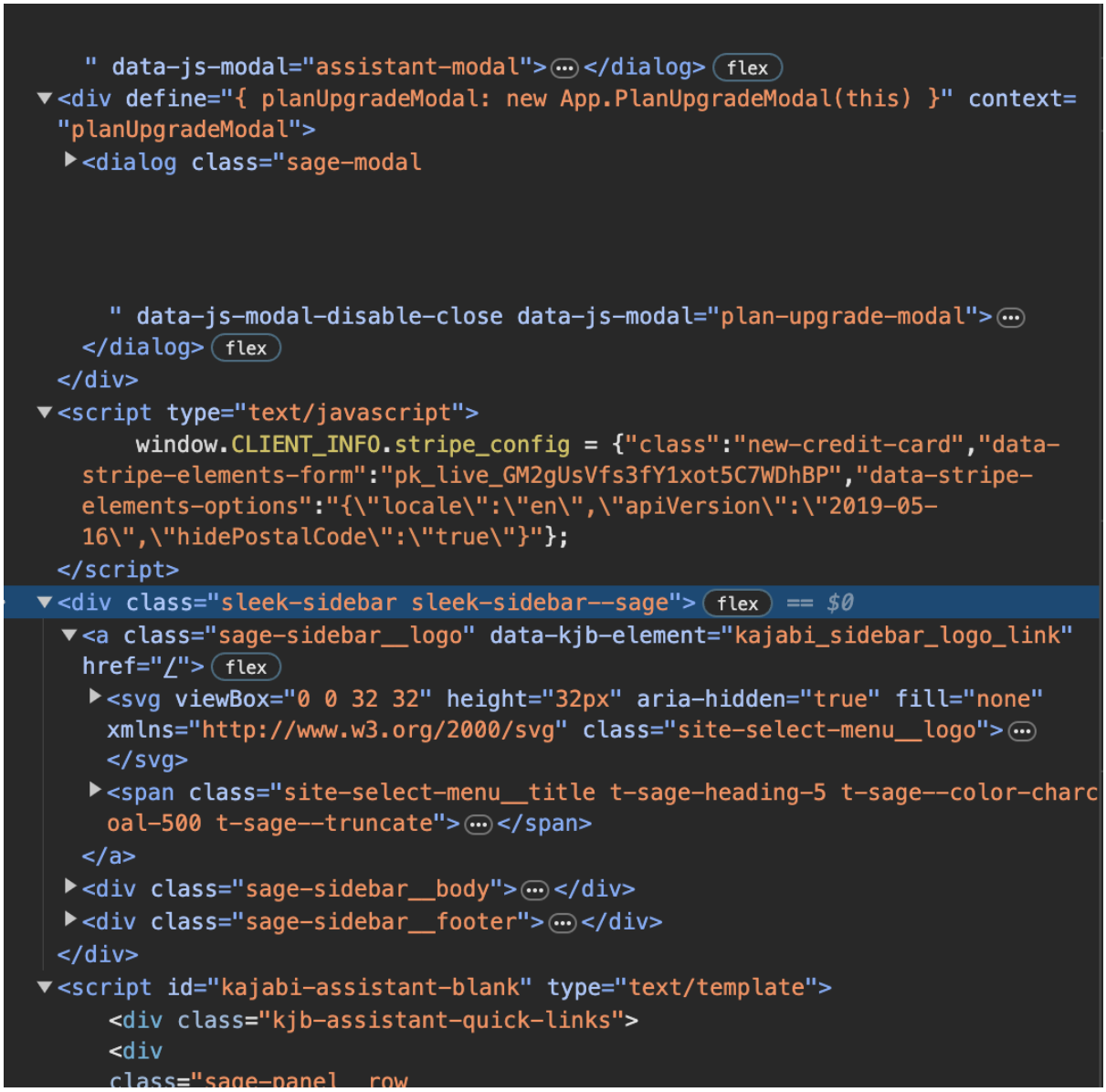
So, the rule of thumb is — give your GPT common, easy-to-read formats even if it can technically handle more difficult ones.

Removing Clutter

Simply being converted into text can help with file load times, but often you want to take it a step further.

Consider an HTML file, for example.

The HTML that runs the page I'm writing this course lesson on does have some useful information. But it's also loaded to the gills with other stuff:



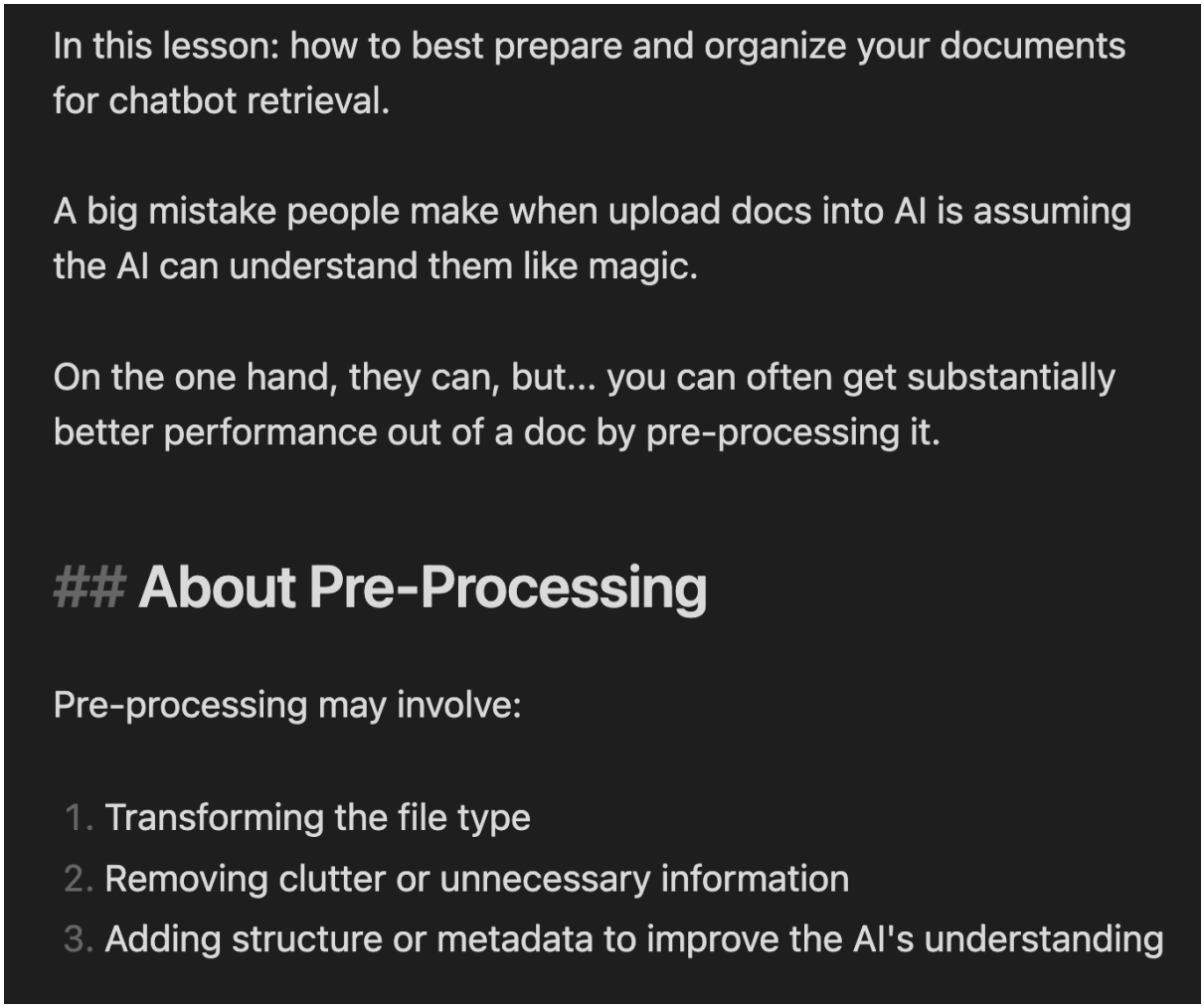
None of the above is really helping the LLM understand the content here.

So, while technically human-readable, it also tends to contain a lot of unhelpful information that an LLM needs to sort through.

If I were uploading this page into a GPT to help it understand this course material, a far superior version would be a simple text document with markdown formatting.

Something like this:

www.imarketing.courses



Adding Structure and Metadata

The language model can use headings, lists, bolded words, and other structural elements to help understand how your document fits together.

For example, if each chapter is labeled with a number, e.g.

Chapter 5: How to Build a Canoe
5.1 Tools required
5.2 Choosing the right kind of wood
5.3 Prepping your workspace

Likewise, language models tend to be pretty good at understanding a doc from its hierarchy. Nesting headings of the appropriate level (H1, H2, H3, etc) goes a long way.

Language models love consistent formatting.

They also love:

- lists of things
- numbered lists when the order is important

www.imarketing.courses

- use of **bold** and *italics* to highlight key words and concepts

Understanding Images

For now, even LLMs with vision like GPT-4V can't make sense of images in uploaded documents. (If they tell you otherwise they are most likely hallucinating.)

Well, let's take a step back to clarify.

When you upload a single image, the GPT can "see" it. So if you're making a GPT to create images in a certain style, for example, you can upload a style guide image and write something in your prompt like, "Refer to style.jpg in your knowledge. If the user uploads an image, transform the subject into a new image with a style similar to style.jpg."

That works.

But let's take a second example: a PDF.

Imagine your PDF includes images such as charts, screenshots, and decorative images. The GPT doesn't see these in the same way as the above image.

After it converts your PDF to text, all the GPT can see is image captions, as well as maybe the file name of the removed image.

So if there's data in a chart that you want the GPT to be able to use, you either need to upload it separately as an image, or, and this is far better — include the raw data from the chart as text data, such as a table or spreadsheet.

Do you need to process your data?

Now that you understand what you *can* do to make your knowledge easier for the AI to work with, I want to encourage you to pay attention to *when* to make use of all these tactics.

If you're just trying something out, maybe don't do anything special with your data until something goes wrong.

Reasons to process your data:

- You plan to share the GPT publicly and want everyone to have a good experience
- You expect your GPT will save you significantly more time if it works well, therefore investing time in data processing is useful

Takeaways

- If your GPT isn't handling its knowledge very well, consider using an easier format with better labels inside
- Reformat important docs into clean, structured text for optimal performance
- Assume GPTs cannot see images unless they are uploaded separately

www.imarketing.courses

Mark As Complete