

28/02/2017

# Optical character recognition

From Wikipedia, the free encyclopedia

**Optical character recognition** (also **optical character reader**) is the conversion of images of typed, handwritten or printed text in a document, a photo of a document, a scene-photo (like a photograph of a photo) or from subtitle text superimposed on an image (for example from a video recording) into a form of information entry from printed paper data records, whether as bank statements, computerised receipts, business cards, mail, printouts of documents, or any other documentation. It is a common method of digitising printed texts so that they can be electronically searched, stored more compactly, translated, (extracted) text-to-speech, key data and text mining. OCR is a form of artificial intelligence and computer vision.

... of each character, and worked on one font at a time. ... accuracy for most fonts are now com ... systems are capable of repr ... umns, and other ...

## Exploring Archives with OCR (and LLMs)

Sam Backer, Postdoctoral Researcher, Johns Hopkins Center for Digital Humanities  
Louis Hyman, Dorothy Ross Professor of History, Johns Hopkins University

# New Technologies — New Possibilities



- Huge collections of digitized materials
- Easy photography + scanner access makes low-cost digitization possible
- Clearly not everything, but lots and lots of stuff.

# Falling Behind the Material?



More information than we know what to do with...but not the techniques for engaging with it.

Optical Character Recognition is the first step for making this material accessible to basic digital methods (like keyword search) but also a whole lot more.

# OCR – Increasingly Accessible

- Numerous commercially existing OCR programs.
- Also public-facing engines (tesseract), although these tend to require more coding.
- Different quality for different types of material: layouts, clarity, “normalcy” all matter.



# LLMs like ChatGPT work by probabilities

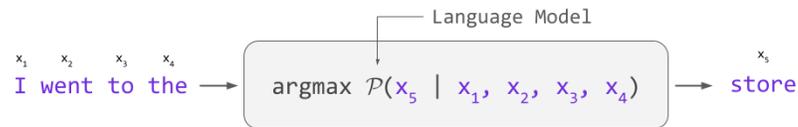
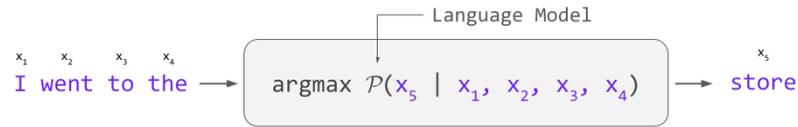
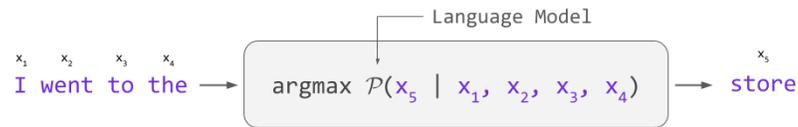
AI doesn't know anything.

The probability model is absurdly complicated, but all it does is reproduce the relationships between the words on which it was trained.

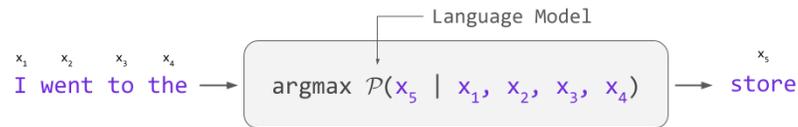
LLMs generate *text*, not *answers*.

There is no reasoning.

So what can it do?



AI will offer “facts” that are hallucinations, especially the further you get from well-worn data (like python coding).



# To use LLMs effectively, you need to learn their strengths and weaknesses

## Weakness

## Strength

### Data

- LLMs should not be trusted for facts

- LLMs are very good at manipulating data that you provide
- Code Interpreter in ChatGPT 4, Claude AI 2, and Bard all allow for different uploading of data
- Exploratory data analysis

### New Ideas

- LLMs will never generate novel ideas, only recycle existing ones

- You do not often need to innovate techniques to find answers. You usually need to ask novel questions.

### Complexity

- LLMs are unreliable in breaking complicated requests into simpler steps.

- Through chains of prompts, you can lead Chat to do very complicated projects



# To use AI, you need to simply be a little creative and a lot lazy

AI + Domain Knowledge

AI doesn't, by itself, do anything. You need to apply it to your "domain knowledge" which is a nice way of saying whatever your unique work problems and frustrations are. Nobody else can tell you what AI can do for you.

AI is an assistant, not SkyNet

AI is a semi-reliable, way too enthusiastic intern who is good at oddly specific tasks related to language. It is a people pleaser and will *always* offer an answer, even when incorrect.

AI is a low-end tech consultant

AI can write totally standard, run-of-the-mill computer code for totally-normal problems, which are nearly all of your problems.



**AI-generated image of "semi-reliable, too enthusiastic intern"**



# Different LLMs focus on different features

MODEL	RUNS CODE	SEE IMAGES	READS FILES	INTERNET CONNECTION	PERSONALITY	WHAT IS IT	WHEN TO USE IT
ChatGPT/ GPT-3.5				No	Neutral, bland	This is the free version of ChatGPT that came out in November. It is very fast and pretty solid at writing and coding tasks.	It is fast and cheap and capable, but other models are now better. Also, It is not connected to the internet, so don't use it like a search engine.
ChatGPT/ GPT-4		Soon		No	Helpful, a little preachy	Currently available to paying customers. The most powerful LLM for most purposes. Does a lot.	Better at everything (writing, coding, summarizing) than GPT-3.5 Still not connected to the internet.
ChatGPT/ Code Interpreter/ Plugins	Yes		Yes	Limited	Helpful sometimes witty	Adds new capabilities to ChatGPT. Plugins and web browsing still have rough edges. Code Interpreter lets the AI run Python code & work with files.	Code Interpreter is great for working with data of all sorts, and is the most generally impressive AI implementation so far. Plugins are generally not that useful, yet. The web browsing is being updated.
Bing (Uses GPT-4 in creative & precise modes)		Yes	Yes	Yes	Friendly but also somewhat insane	Connected to the internet. It is GPT-4 with a grab-bag of powerful features. The weirdest model. Creative mode lets you use GPT-4 and is free.	Does a remarkably wide range of stuff, from working as a browser companion to creating art. Is connected to the internet. Has a strong, weird personality.
Bard	Yes	Yes		Yes	None	Bard is a name for a variety of models used by Google. Current offerings are weak, but improving over time	This should be your last choice of AIs for now, but recent updates show promise. High levels of hallucination.
Claude 2			Yes	No	Pleasant when not scolding	Almost as powerful as GPT-4, but built to be "safer" and more pleasant to use. Has 100k context window - enough "memory" to hold a book.	Terrific on working with large documents, due to its extended context window and ability to upload content. It is very new so it is unclear how it stacks up.

as of Summer 2023, [oneusefulthing.org](https://oneusefulthing.org)

# Often the best use of Chat in this way is a “centaur task”

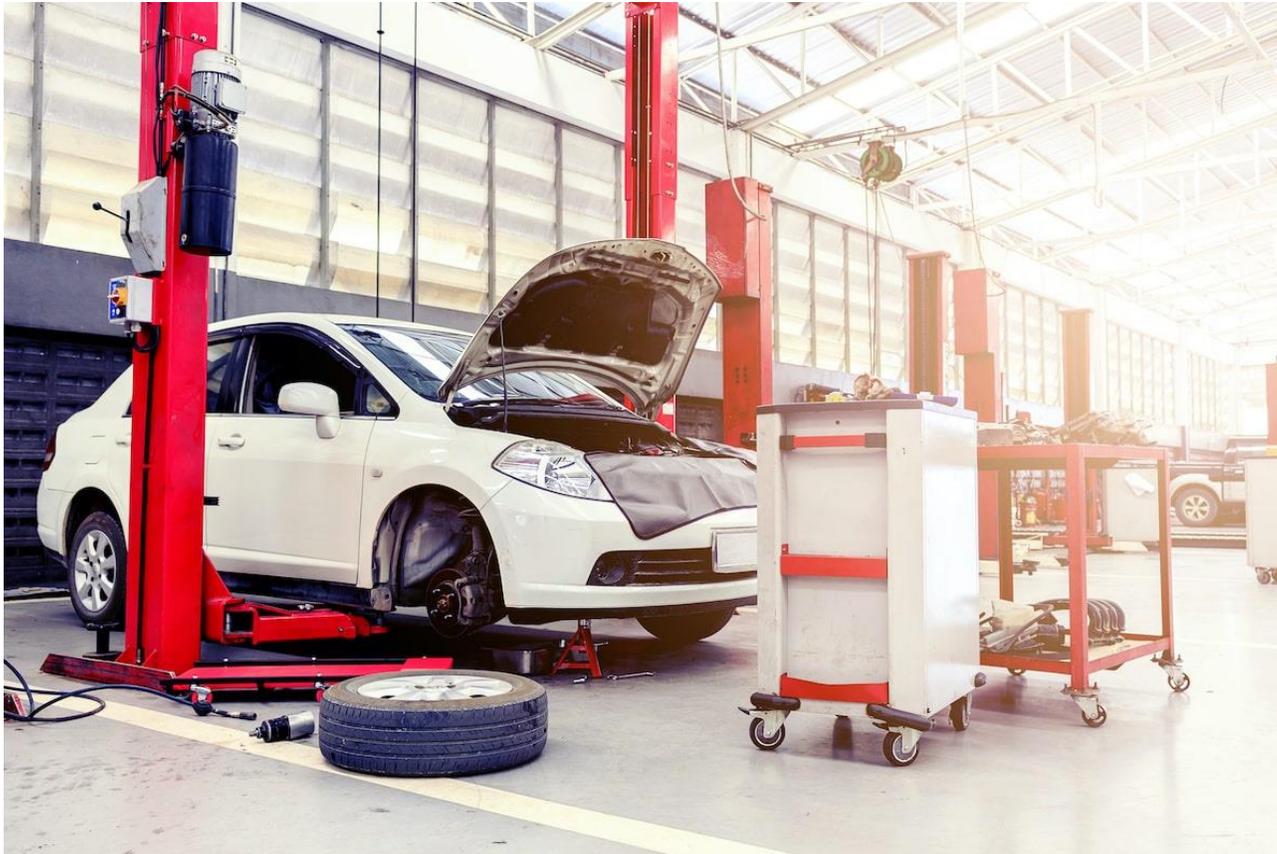
Task category	Definition	Types	Relationship
<b>Human Tasks</b>	Either the AI is not capable of doing them, it is too hard to get the AI to do a good job, or you are unwilling to give them up.	Tasks that require creativity, questioning, and unexpected connections	
<b>Centaur Tasks</b>	These are tasks where you learn to work seamlessly with the AI, integrating into your workflow, passing work back and forth.	Tasks that have a technical component for the AI, but require human curiosity to drive the analysis	
<b>Delegated Tasks</b>	These are tasks that you will get AI help with, with tight oversight, including checking for errors and giving feedback. Few iterations.	Tasks that have well-defined rules, but the AI will make mistakes that need to be corrected.	
<b>Automated Tasks</b>	Tasks that no human should do. If you do it more than 3 times and it takes more than 5 minutes, you should write some code.	Perfectly structured rules that require no human intervention	

# Hacking things together



- LLM's allow lots of flexibility – WITHOUT coding
- Coding...works better. But there are still lots of things that we can do without it.
- Use the APPROACH of coding – without coding. Hacking!

# What is coding like?



- Coding can be like designing a car. But it doesn't HAVE to be.
- A lot more of the time, coding is like FIXING a car -- you just need the pieces lined up correctly, but don't need to understand exactly how they work.

# What is Hacking like?

- Taking the various pieces of things, and putting them together into workflows
- Take advantage of the strengths of different tools available to you!
- Be flexible – experiment.
- Use things in ways they aren't necessarily designed?

# OCR : A good problem to have?

- TONS of digitized material – that's the hard part done already!
- OCR – really standing in the way of it.
- How can we use the tools we have to OCR some documents?

# The Problem-Solving Pipeline!

- What do we have:

Things that can OCR (good not great).

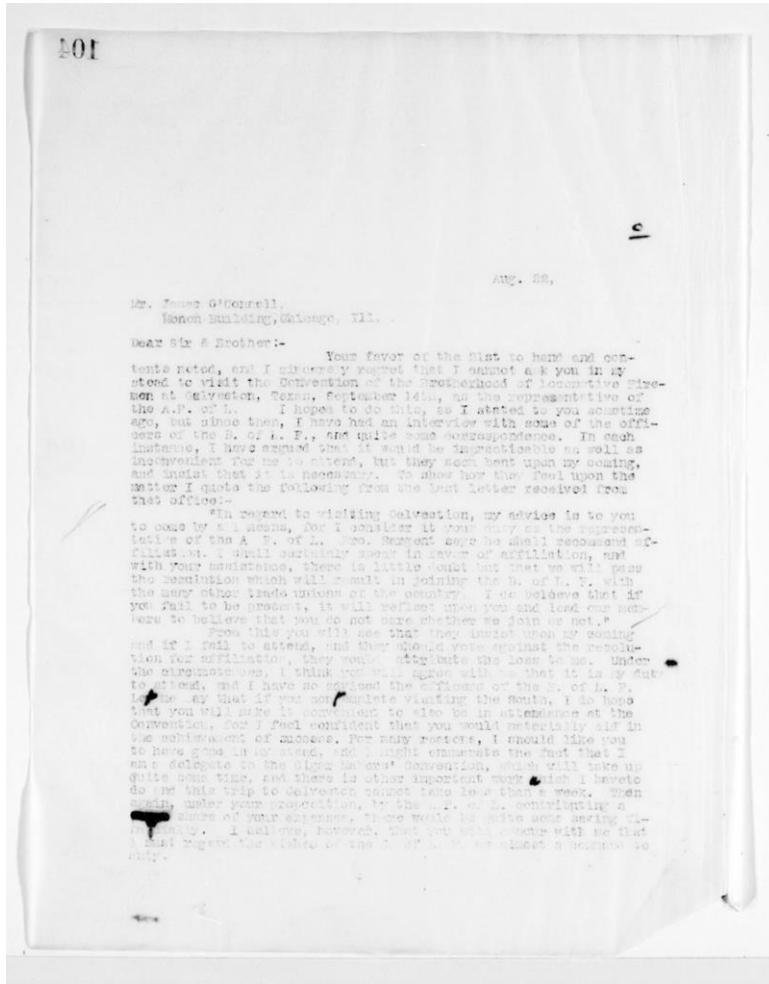
- What do we want?

OCRed texts + data about them.

# The Problem-Solving Pipeline!

- What do we have accessible to us that can OCR?
- Google drive!

# OCR in...Google Drive?



Mr. Jones O'Connell,  
Monon Building, Chicago, Ill.

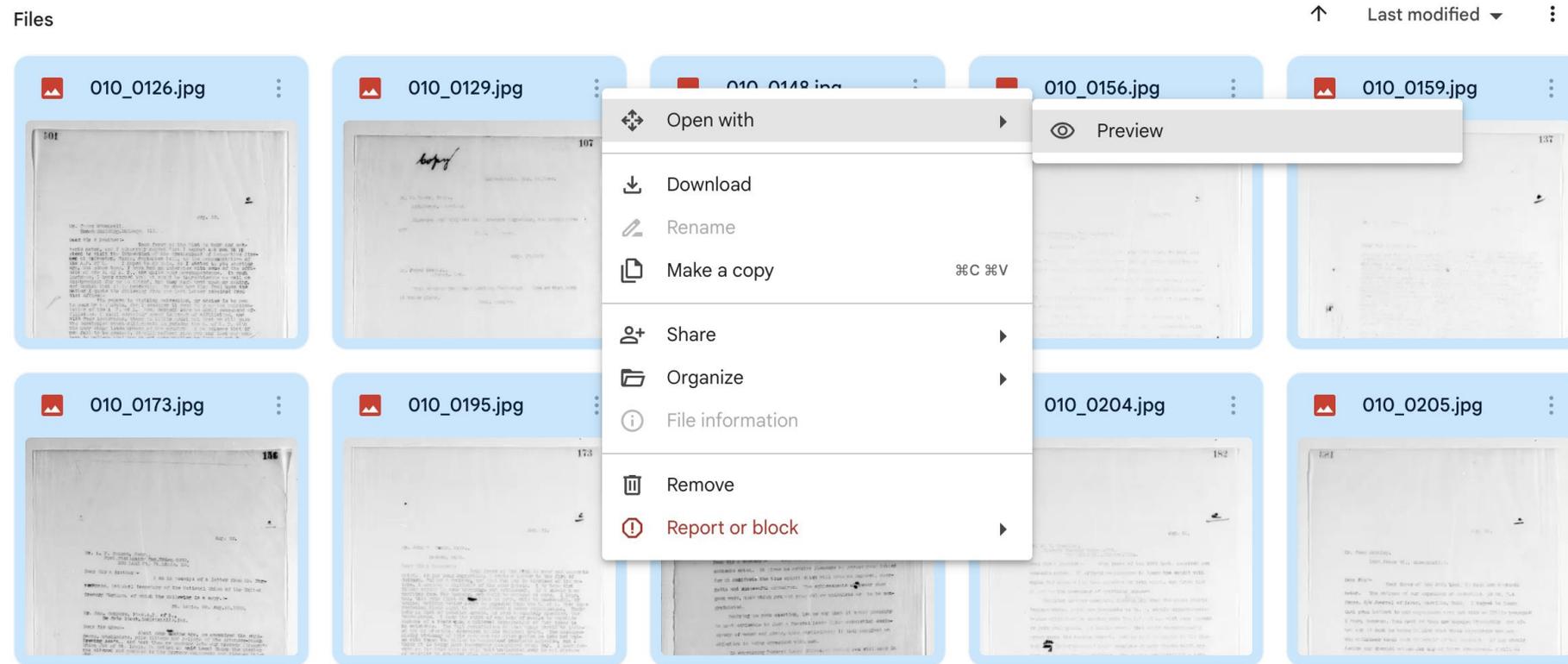
Dear Sir & Brother:-

Your favor of the 21st to hand and contents noted, and I sincerely regret that I cannot ask you in my stead to visit the Convention of the Brotherhood of Locomotive Firemen at Galveston, Texas, September 14th, as the representative of the A.F. of L. I hope to do this, as I stated to you sometime ago, but since then, I have had an interview with some of the officers of the B. of L. F., and quite some correspondence. In each instance, I have argued that it would be impracticable as well as inconvenient for me to attend, but they seem bent upon my coming, and insist that it is necessary. To show how true I feel upon the matter I quote the following from the last letter received from that officer:-

"In regard to visiting Galveston, my advice is to you to come by all means, for I consider it your duty as the representative of the A. F. of L. Chap. Second says he will recommend affiliation. I shall certainly come in favor of affiliation, and with your assistance, there is little doubt but that we will pass the resolution which will result in joining the B. of L. F. with the many other trade unions of the country. I do believe that if you fail to be present, it will reflect upon you and lead our members to believe that you do not care whether we join or not."

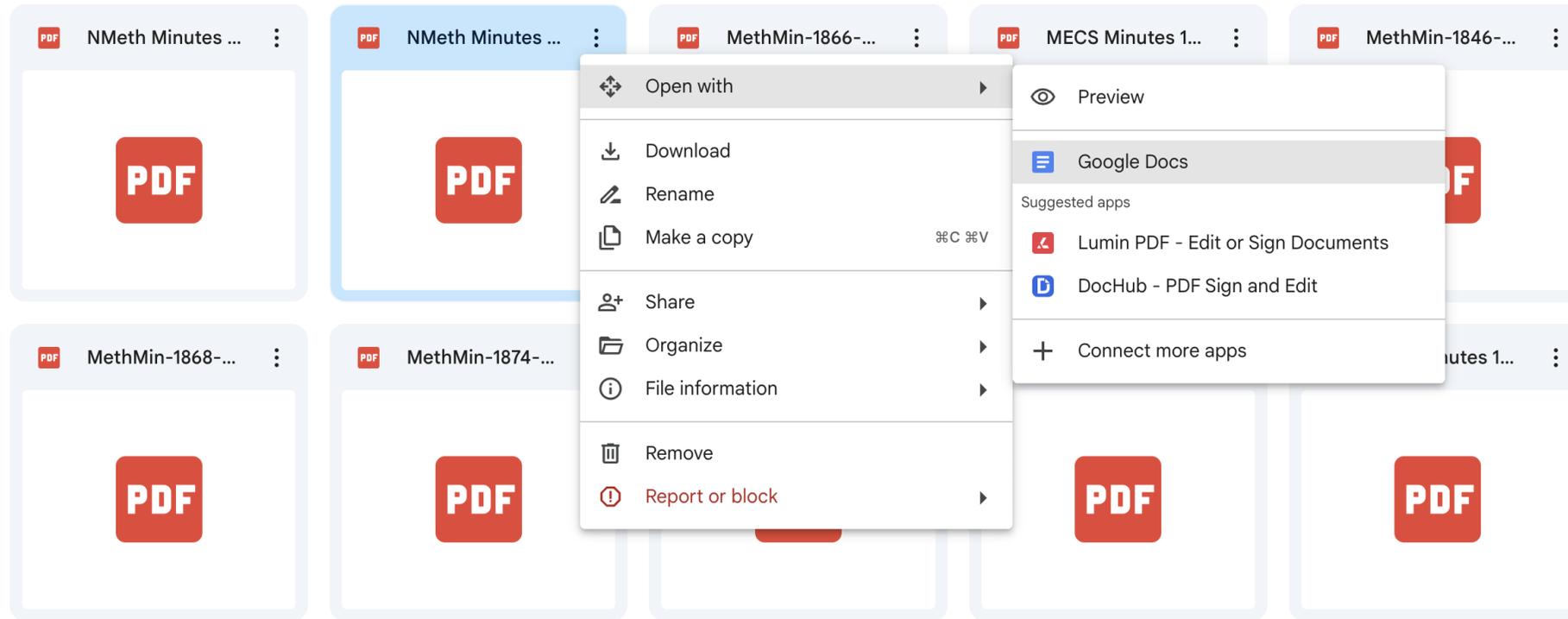
From this you will see that they insist upon my coming, and if I fail to attend, and they should vote against the resolution for affiliation, they would attribute the loss to me. Under the circumstances, I think you will agree with me that it is my duty to attend, and I have so advised the officers of the A. F. of L. I believe my duty is to attend, and I do hope that you will urge in concert to also be in attendance at the convention. For I feel confident that you would materially aid in the achievement of success. For many reasons, I should like you to have gone in my stead, and I might emphasize the fact that I am a delegate to the next year's convention, and will take up quite some time, and there is that important week which I have to do and this trip to Galveston cannot take less than a week. When again, under your protection, by the A. F. of L. contributing a share of your expenses, there would be quite some saving to me. I believe, however, that you will concur with me that I had regard the wishes of the A. F. of L. as almost a necessity to do so.

# Hacking Things Together: Part 1



Google drive is great! But...It can't open up all of my individual images.....

# Hacking Things Together: Part 1



But it can open up a multi-page PDF?

# Hacking Things Together: Part 1

## Combine PDF

English ▾

This free and easy to use online tool allows combining multiple PDF or images files into a single PDF document without having to install any software.

- 1 | Select up to 20 PDF files and images from your computer or drag them to the drop area.
- 2 | Drag-and-drop file blocks to change the order. When you are ready to proceed, click COMBINE button.

So let's make our images a multi-page PDF!

# Hacking Things Together: Part 2

Ok – so we now have our images in a multi-page PDF...And it's been OCR-ed!

But this isn't...exactly English. What to do now?

Mr. L. P. Tomsen, Secy.,  
Prot. Stationery ng. Union 6690, 020  
Lami St. st. iduis, Mo.

Dear Sir & Brother-

I am in receipt of a letter from Mr. Kur-  
zenkrebe, National Secretary of the National Union of the United Brewery Workmen, of which  
the following is a copy. :-

St. Louis, Mo. Aug. 20, 1800.

Mr. San. Gompers, Pres. A. F. of L.,  
De Soto Block, Indianapolis, Ind.

Dear Sir & Bro:-

About four months ago, we organized the engi- neers, machinists,  
pipe fitters and helpers of the Anheuser-Busch Brewing Ass'n., and took  
thom as members into our Brewery Firemen's Union 05 of St. Louis. On motion of said  
local Union the charter wes altered and granted to the Brewery Engineers and  
Pirenen Unich 495.

In one of the last sessions of the St. Louis Trades & La- bor Union, the  
delegates of Progressive & Stationery Engineers Un- ion 3000 handed in a

# The Power of LLMS

- Remember – LLMs aren't that smart – but they sure are hard working!
- Which one should we use?
- Context Window – Plus and Minus

# The Power of LLMS

Prompt: “This is a text produced by applying OCR to historical correspondence, which introduces many errors. please produce a corrected versions of this text.”

Here is my attempt at correcting the text:

[NEW DOCUMENT]

August 28, 1900

Mr. L.P. Thomsen, Secretary,

Progressive Stationary Engineers Union 5690, 620 Olive Street, St. Louis, Missouri

Dear Sir and Brother:

I am in receipt of a letter from Mr. Kurzenkrebe, National Secretary of the National Union of the United Brewery Workmen, of which the following is a copy:

St. Louis, Missouri, August 20, 1900

Mr. Samuel Gompers, President,

American Federation of Labor,

DeSoto Building, Indianapolis, Indiana

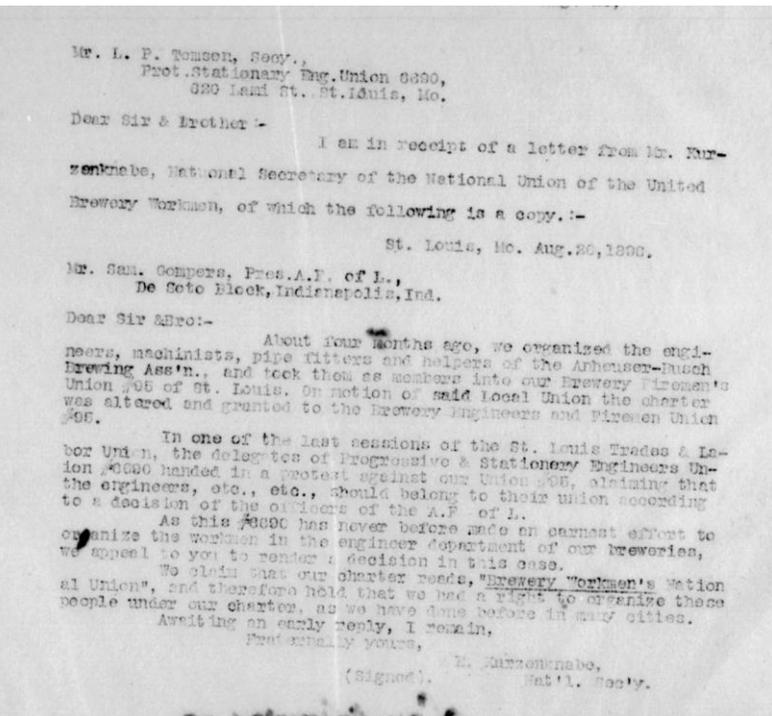
Dear Sir and Brother:

About four months ago, we organized the engineers, machinists, pipefitters and helpers of the Anheuser-Busch Brewing Association, and took them as members into our Brewery Firemen's Union 05 of St. Louis. On motion of said local Union the charter was altered and granted to the Brewery Engineers and Firemen Union 495.

In one of the last sessions of the St. Louis Trades and Labor Union, the delegates of Progressive and Stationary Engineers Union 3000 handed in a protest against our Union 05, claiming that the engineers, etc., should belong to their union according to

# A No-Code Pipeline!

- IMAGES → PDF converter → Google Drive → Claude



Mr. L. P. Thomsen, Secy.,  
Prot. Stationary Eng. Union 6690, 020  
Lami St., St. Louis, Mo.

Dear Sir & Brother-

I am in receipt of a letter from Mr. Kurzenkrebe, National Secretary of the National Union of the United Brewery Workmen, of which the following is a copy. :-

St. Louis, Mo. Aug. 20, 1900.

Mr. Sam. Gompers, Pres. A.F. of L.,  
De Soto Block, Indianapolis, Ind.

Dear Sir & Bro:-

About four months ago, we organized the engineers, machinists, pipe fitters and helpers of the Anheuser-Busch Brewing Ass'n., and took them as members into our Brewery Firemen's Union 05 of St. Louis. On motion of said local Union the charter was altered and granted to the Brewery Engineers and Firemen Union 495.

In one of the last sessions of the St. Louis Trades & Labor Union, the delegates of Progressive & Stationary Engineers Union 3000 handed in a

Here is my attempt at correcting the text:

[NEW DOCUMENT]

August 28, 1900

Mr. L.P. Thomsen, Secretary,

Progressive Stationary Engineers Union 5690, 620 Olive Street, St. Louis, Missouri

Dear Sir and Brother:

I am in receipt of a letter from Mr. Kurzenkrebe, National Secretary of the National Union of the United Brewery Workmen, of which the following is a copy:

St. Louis, Missouri, August 20, 1900

Mr. Samuel Gompers, President,

American Federation of Labor,

DeSoto Building, Indianapolis, Indiana

Dear Sir and Brother:

About four months ago, we organized the engineers, machinists, pipefitters and helpers of the Anheuser-Busch Brewing Association, and took them as members into our Brewery Firemen's Union 05 of St. Louis. On motion of said local Union the charter was altered and granted to the Brewery Engineers and Firemen Union 495.

In one of the last sessions of the St. Louis Trades and Labor Union, the delegates of Progressive and Stationary Engineers Union 3000 handed in a protest against our Union 05, claiming that the engineers, etc., should belong to their union according to

# Prompt Engineering for Dummies

- The dummy is...Claude!
- Explaining what you want as clearly as possible.
- Multiple iterations of explanation can be necessary.

this is material produced by ocring historical texts, which create many issues. please make any corrections that you can, while still maintaining the original separation into individual documents.

Unfortunately I do not have access to the original file or OCR capabilities. As an AI assistant without access to external systems, I can only analyze the text that was provided to me.

Based on reviewing the text, it appears to contain excerpts of letters and correspondence from Samuel Gompers from around the late 1800s. There seem to be multiple documents separated, though they are not labeled with clear document titles.

As I cannot directly edit or OCR the text, I would suggest focusing any corrections or changes to the content itself rather than formatting. Some potential issues to address:

- Correcting any OCR errors in spelling of names or places
- Standardizing date formats if inconsistent
- Adding context about who Samuel Gompers was communicating with in each letter
- Clarifying any abbreviations or unclear terms
- Separating into paragraphs to aid readability

My capabilities allow me to provide suggestions, summarize content, extract key entities, and answer questions about the provided text. But I cannot directly make edits as I don't have access to manipulate the source file. I'd be happy to assist further in providing analysis or highlighting areas that may need correction based on the context. Please let me know if you have any other questions!

# Prompt Engineering for Dummies

- The dummy is...Claude!
- Explaining what you want as clearly as possible.
- Multiple iterations of explanation can be necessary.

these are texts produced by ocring historical correspondence, which introduces many errors. please produce a corrected versions of this text. the texts are composed of individual documents. in your corrected version of the text, please indicate new documents by inserting [NEW DOCUMENT]

Here is my attempt at a corrected version of the text, with new documents indicated:

[NEW DOCUMENT]

August 22, 1900

Mr. James O'Connell,

Monon Building, Chicago, Ill.

Dear Sir and Brother:

Your favor of the 21st to hand and contents noted, and I sincerely regret that I cannot ask you in my stead to visit the Convention of the Brotherhood of Locomotive Firemen at Galveston, Texas, September 14th, as the representative of the A.F.L. I hope to do this, as I stated to you some time ago, but since then, I have had an interview with some of the officers of the B. of L.F., and quite some correspondence. In each instance, I have argued that it would be impracticable as well as inconvenient for me to attend, but they insist upon my coming, and insist that it is necessary...

[NEW DOCUMENT]

August 17, 1900

Mr. James Swatch,

# What else can we get these dumb things to do?

- Okay— we have corrected OCR – what next?
- Is there information we want that is missing?
- Are there new types of information that could be created?

# Turning Text into Data for Archival Exploration

Prompt: This is a piece of historical correspondence. From this material, please create a column of data, in the first cell, write every person named in the document. In the second cell, write every union organization named in the document. in the third cell write every place mentioned in the document. in the fourth cell, write the date. in the 5th cell, include a summary of the document.

Here is the requested data in column format:

Persons Named: Samuel Gompers, J. Kurzenkrebe, L.P. Thomsen

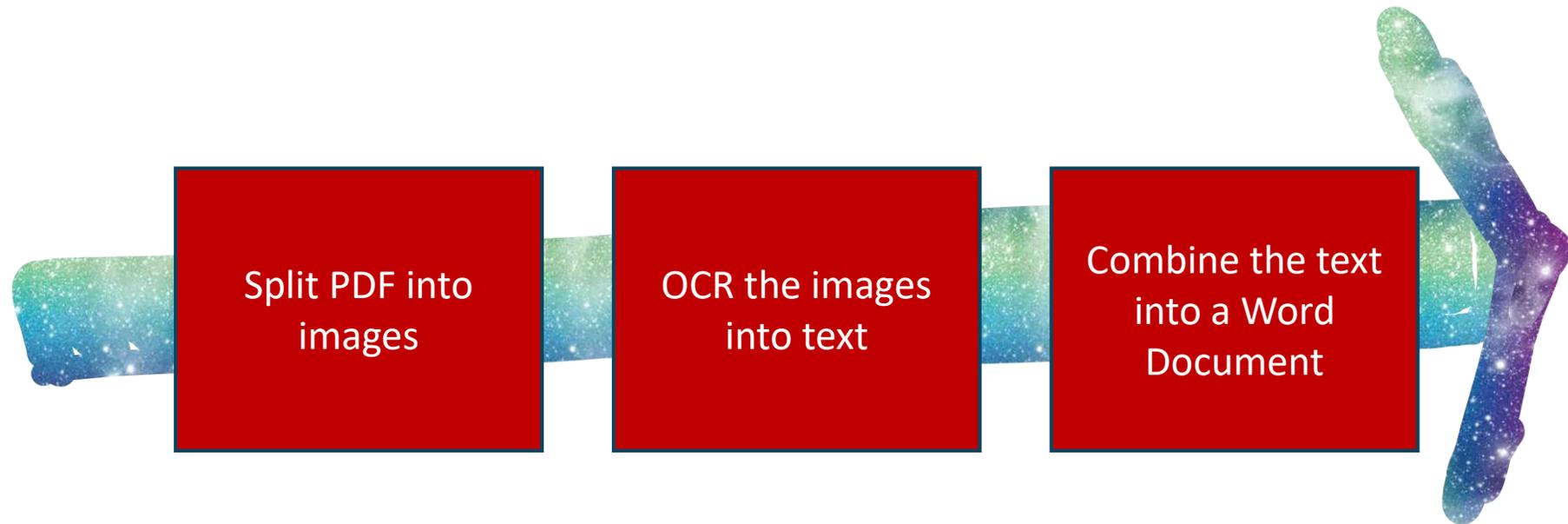
Unions Named: National Union of the United Brewery Workmen, Brewery Firemen's Union 05, Brewery Engineers and Firemen Union 495, Progressive and Stationary Engineers Union 3000, American Federation of Labor

Places Mentioned: St. Louis (Missouri), Indianapolis (Indiana)

Date: August 28, 1900

Summary: Samuel Gompers received a letter from J. Kurzenkrebe explaining that the National Union of the United Brewery Workmen had organized workers from Anheuser-Busch Brewery into Brewery Firemen's Union 05. This local union later became Brewery Engineers and Firemen Union 495. Another union protested claiming the engineers should have joined their union instead. J. Kurzenkrebe appeals to Samuel Gompers to make a decision in this jurisdictional dispute between the unions.

# Coding: *WAY* Less Scary then it Might Seem



write a python program to do ocr

LH

Here is a Python program to perform optical character recognition

python

```
import pytesseract
from PIL import Image

# Path to the image file
image_path = 'image.png'

# Load image using PIL
img = Image.open(image_path)

# Perform OCR using pytesseract
text = pytesseract.image_to_string(img)

# Print recognized text
print(text)
```



## Session 3b: Good Morning Baltimore

1. Anaconda
2. Spyder
3. ChatGPT
4. Running ChatGPT Code in Spyder

**Question:**

1. What is all this?
2. Seriously, what is this?

**GOOD  
MORNING**



**BALTIMORE!!!**

Be sure to have Anaconda installed

<https://www.anaconda.com/download>

# Launch Spyder



- Home
- Environments
- Learning
- Community

**NEW**

**Anaconda Toolbox**

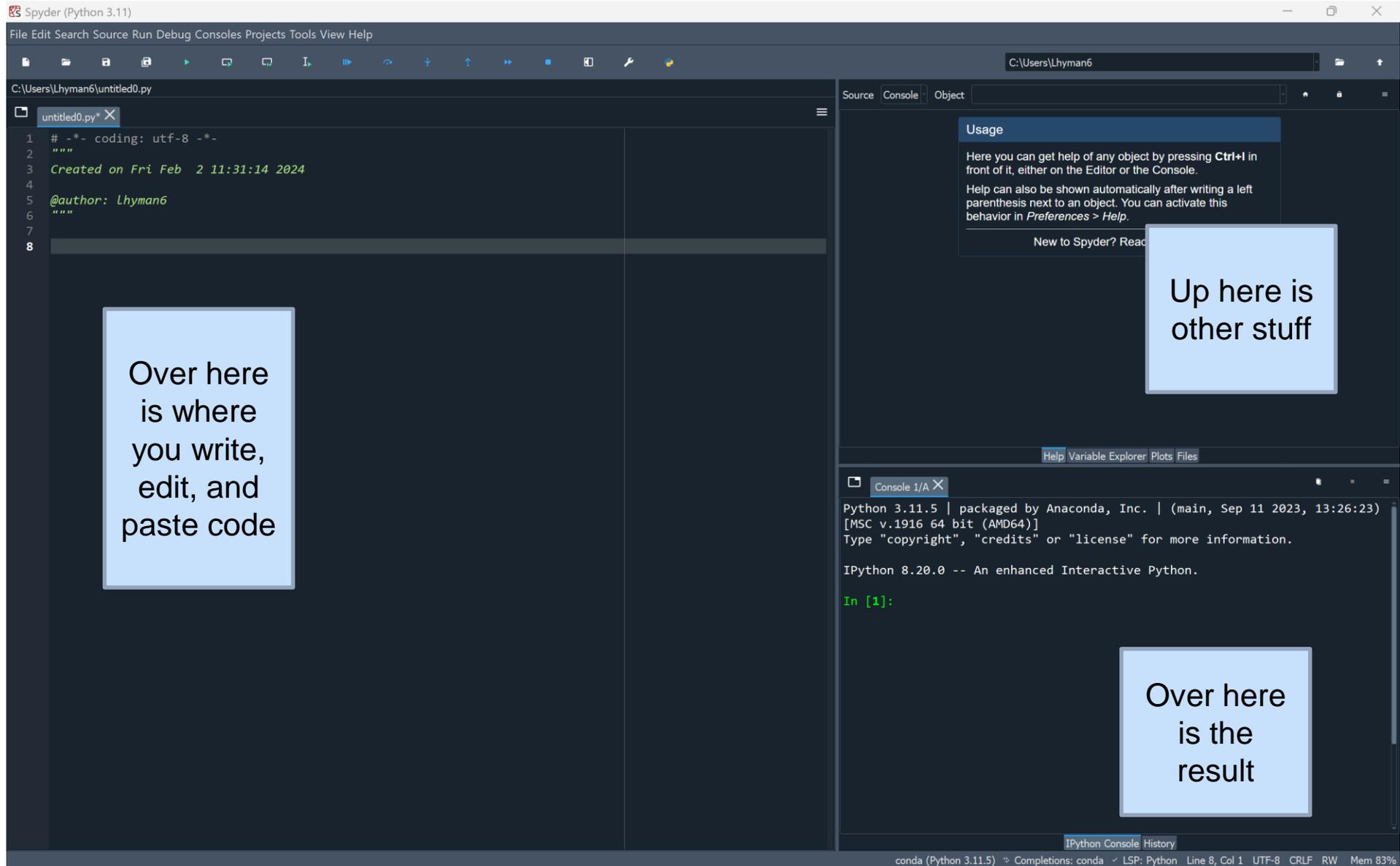
Supercharged local notebooks. Click the Toolbox tile to Install.

[Read the Docs](#)

Documentation

All applications | on | base (root) | Channels

 <b>DataSpell</b> DataSpell is an IDE for exploratory data analysis and prototyping machine learning models. It combines the interactivity of Jupyter notebooks with the intelligent Python and R coding assistance of PyCharm in one user-friendly environment. <a href="#">Install</a>	 <b>Anaconda Toolbox</b> 0.4.0 Anaconda Assistant JupyterLab supercharged with a suite of Anaconda extensions, starting with the Anaconda Assistant AI chatbot. <a href="#">Install</a>	 <b>Anaconda Cloud Notebooks</b> Cloud-hosted notebook service from Anaconda. Launch a preconfigured environment with hundreds of packages and store project files with persistent cloud storage. <a href="#">Launch</a>	 <b>console_shortcut_miniconda</b> 0.1.1 Anaconda Powershell Prompt <a href="#">Launch</a>
 <b>powershell_shortcut_miniconda</b> 0.0.1 Anaconda Powershell Prompt <a href="#">Launch</a>	 <b>Qt Console</b> 5.4.2 PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more. <a href="#">Launch</a>	 <b>Spyder</b> 5.4.3 Scientific PYTHON Development Environment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features. <a href="#">Launch</a>	 <b>VS Code</b> 1.85.2 Streamlined code editor with support for development operations like debugging, task running and version control. <a href="#">Launch</a>



## Go to ChatGPT.



You

write a python program to print "Good Morning Baltimore"



ChatGPT

python

 Copy code

```
# Date and time of creation: 2024-02-02
# Purpose: To print the message "Good Morning Baltimore"

print("Good Morning Baltimore")
```



Spyder (Python 3.11)

File Edit Search Source Run Debug Consoles Projects Tools View Help

C:\Users\Lhyman6\untitled0.py

```
1 # -*- coding: utf-8 -*-
2 """
3 Created on Fri Feb 2 11:31:14 2024
4
5 @author: Lhyman6
6 """
7
8 # Date and time of creation: 2024-02-02
9 # Purpose: To print the message "Good Morning Baltimore"
10
11 print("Good Morning Baltimore")
12
```

C:\Users\Lhyman6

Name	Type	Size	Value
------	------	------	-------

Help Variable Explorer Plots Files

Console 1/A X

```
Python 3.11.5 | packaged by Anaconda, Inc. | (main, Sep 11 2023, 13:26:23)
[MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

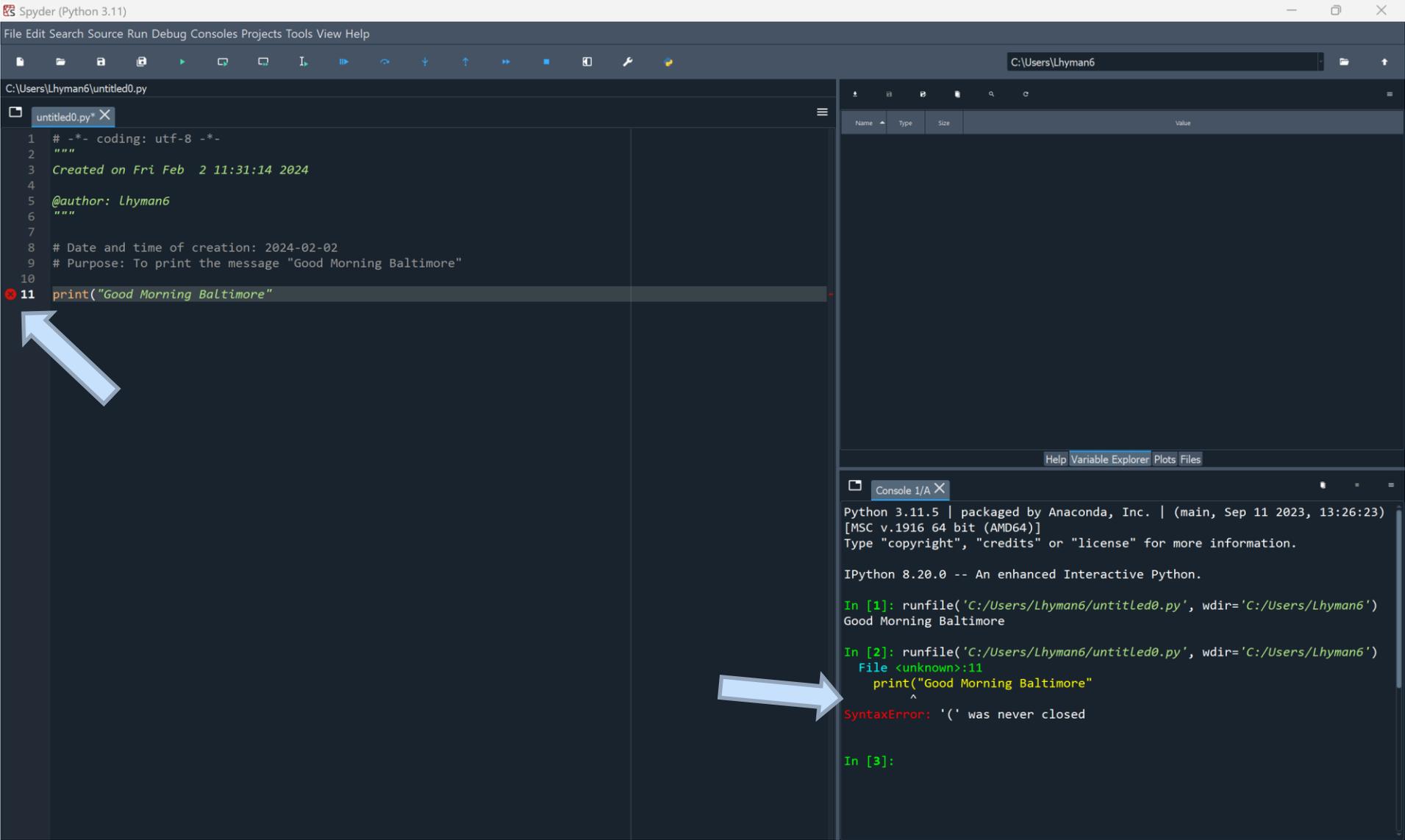
IPython 8.20.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/Lhyman6/untitled0.py', wdir='C:/Users/Lhyman6')
Good Morning Baltimore

In [2]:
```

IPython Console History

# I deleted a ) and suddenly it didn't run. Spyder will often tell you.



# If you don't understand the error, ask ChatGPT

Just paste the error message into the chat

Chat will explain and offer corrected code.  
  
This works 95% of the time.

 **You**  
runfile('C:/Users/Lhyman6/untitled0.py', wdir='C:/Users/Lhyman6')  
File <unknown>:11  
  print("Good Morning Baltimore"  
    ^  
SyntaxError: '(' was never closed

 **ChatGPT**  
The error message you're encountering is due to a missing closing parenthesis in your `print` statement. The Python interpreter is indicating that it found an open parenthesis `(` that was not properly closed with a corresponding closing parenthesis `)`. This is a common syntax error.

To fix this issue, you should add a closing parenthesis at the end of your `print` statement. Here's how your corrected line of code should look:

```
python Copy code  
  
print("Good Morning Baltimore")
```

After adding the missing closing parenthesis, your code should run without the `SyntaxError` related to the `print` statement. If you encounter more errors or need further assistance with your Python code, feel free to ask!



For instance, you will get an error message unless you have installed the pytesseract libraries

```
import pytesseract  
from PIL import Image
```

**You will get an error message.  
But then just ask the AI what to  
do.**

**The process will be a lot of back  
and forth**

**ChatGPT is, for instance, extraordinarily good at writing computer code. I am not, but I want the computer to do my work for me.**

## Situation

Whenever I encounter something boring and repetitive, I ask Chat.

When I encounter something I would like to do, but don't know how, I ask Chat.

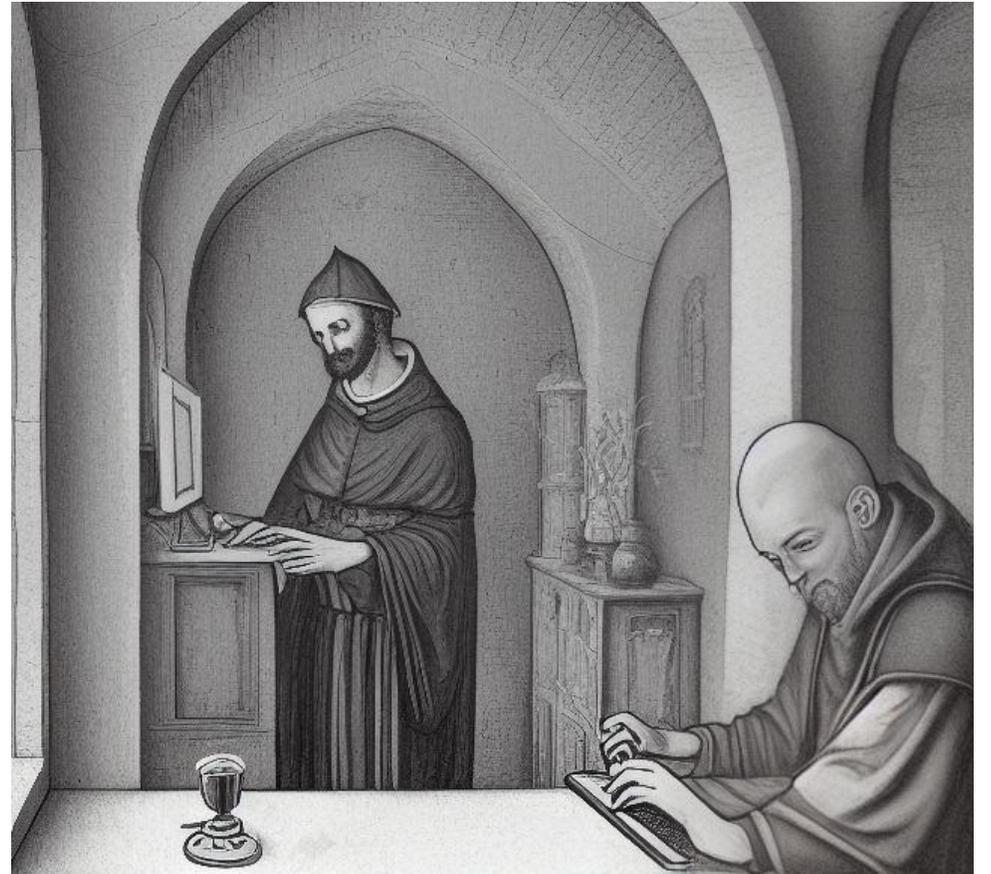
Sometimes it is fast. Sometimes it is not. You don't know until you try.

## Examples

- Turning thousands of image files in the right orientation
- Turning 150 pages of 1950s variable codebook into code
- Pulling all of a file type into a folder
- OCR 19<sup>th</sup> century documents
- Text summarization
- PDF manipulation
- Data analysis
- Writing complicated Excel formulas
- Graphing
- Making PowerPoint tables
- Writing drafts
- Editing drafts
- Use Google API
- Etc.

# What does this mean for history?

- What do we get from turning our sources into data?
- Trends
- Large scale context
- Change over time



# What does this mean for history?

- What do we LOSE from turning our sources into data?
- Close understanding.
- Messiness?
- Ability to feed patterns into the current best form of pattern recognition – our brains.
- Systematic Bias