PDF Analysis + Text Mining

Matilda Gaddi

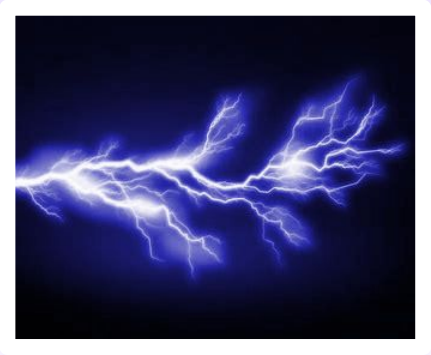
Your experience

Programming?

Python?

Text mining?

The Power of Programming



(In this context)

- ★ Ctrl F on steroids
- ★ Automated dataset/spreadsheet building
- ★ Standardized output
- ★ No need to manually go back through every PDF to extract a new variable
- ★ Adjusting and re-doing work is orders of magnitude less time consuming (hours vs minutes)
- ★ Don’t fear making slight improvements and corrections
- ★ Allows for iteration
- ★ Replicability

Demo: OCDO AIODAWG RFI Analysis

https://github.com/matildagaddi/RFI-analysis/blob/main/RFI-analysis.ipynb

Regex Applications

Exact matches

- Precise words
- Labeled items

Patterns

- Phone numbers
- Addresses
- Email addresses
- Dates
- Web server logs

What else can you think of?

Resources

Youtube

- Tons of specific tutorials

Similar tutorial in R: https://www.charlesbordet.com/en/extract-pdf/#use-pdftoolspdf_text

Reading PDFs

- “pdftotext” Python library: https://pypi.org/project/pdftotext/
- “PyMuPDF” (import fitz) Python library: https://pymupdf.readthedocs.io/en/latest/the-basics.html

ChatGPT

- Interactive instruction and programming

Regex

- “re” Python library: https://docs.python.org/3/howto/regex.html
- Syntax cheatsheet: https://www.rexegg.com/regex-quickstart.php
- Test your regular expression syntax: https://regex101.com/
- UCSD lecture code-along: https://dsc-courses.github.io/dsc80-2024-wi/resources/lectures/lec11/lec11-filled.html

Matilda, I have no Python experience, and so many questions

ChatGPT prompts to help you

- Write a python script to extract addresses from several locally saved pdfs
- How do I run python?
- How do I get jupyter notebook?
- How can I change the code to exclude addresses with apartments?
- What does this error mean and how do I fix it? [include entire error message]
- How do I know what the “path” to my pdf folder is?

“Python” can be replaced with “R”

Just scratching the surface

Extra fun(ctional) facts:

- You can upload datasets to Chat GPT to do initial analysis for you.

Automation is applicable to many repetitive mundane tasks

- I developed a similar project last year at BEA to scan hundreds of tables for missing and illogical values in SQL

What tools/solutions could help make your job easier?

If you could have anything researched/developed, what would it be?

Thanks!
Questions?