How to Use Large Language Models for the Analysis of PDFs

Introductions

This brief guide explains how to analyze text in a batch of PDFs using Claude or other large language models.

Methods

I explain a process that includes extracting text from a directory full of PDFs on your system and uploading it to your Google Sheets. We extract PDFs 50,000 characters at a time because this is a limitation of Google Sheets. We will put the PDF file name in the first column and the extracted test in the second column, and then it can be used for Claude.

Requirements: Google account (Google Sheets and Google Cloud); Python; an LLM model account that functions on Google Sheets (for example Anthropic's Claude)

1. Set up a Google Cloud API and share the Sheets with the Cloud project

- Go to the Google Developers Console.
- Create a new project.
- Enable the Google Sheets AND Google Drive API for your project.
- Create credentials (Service Account) and download the JSON file containing your keys.
- Share your target Google Sheet with the email address of your service account.

2. Run the Python code

- Put the path of the JSON file in the Python code (you can get the path by dragging and dropping the file in the terminal on Mac).
- Put the PDFs in a single folder and put the path of the folder in the Python code.
- Put the name of the shared Google Sheet in the Python code.
- Run the Python code.

3. Use the cloud API to analyze the data.

You can create an account and use any large language model to analyze the sheet. Here Anthropic's Claude 3 is explained:

First, you need to install "Claude for Sheets™" on your Google Sheets.

For now, we have the first column containing PDFs file names, and the second column containing texts from the PDFs. You can run the following formula in the third column:

```
=claude("code this text"&B2, $E$2, "system",$D$2)
```

The first section is the prompt.

The second section is the Claude model, for example, E2 can contain: *claude-3-opus-20240229*

The third section is the main section of the prompt telling the LLM what to do. For example, screening (explaining the eligibility criteria), or data extraction, or analysis.