How to Extract Words from PDFs with Python

As I mentioned in my previous article: [How to Connect to Google Sheets with Python](), I've been working with a client to help them parse through hundreds of PDF files to extract keywords in order to make them searchable.

Part of solving the problem was figuring out how to extract textual data from all these PDF files. You might be surprised to learn that it's not that simple. You see, PDFs are a proprietary format by Adobe that come with their own little quirks when it comes to automating the process of extracting information from each file.

Adobe：創意、行銷和文件管理解決方案

https://www.adobe.com/hk_zh/

Adobe 正在透過數位體驗改變世界。我們協助客戶建立、提供及最佳化內容和應用程式。

[PDF 閱讀器]() · [產品]() · [下載和安裝說明]() · [下載]()

Luckily, we have the right language for the job: [Python](#).

Now, I've made my love for Python clear: It's easily human-readable, it has a ton of awesome libraries that allow you to do basically anything. It's the perfect tool in your utility belt. As I've mentioned before, it makes you batman.