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.

Luckily, we have the right language for the job: Python. Now, I’ve made my love for Python clear. It’s easily readable and 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.