



Unleashing the power of Python in eDiscovery

The challenge - One of PwC's eDiscovery clients approached us with a problem. They had the documents which they wanted to load into the eDiscovery platform so their teams could review it.

However, they weren't able to provide each document separately - instead, all documents were provided as one file. And on top of that - They had only a scanned version of these documents.

Dealing with scanned documents is a common task here at PwC eDiscovery team. We use state of the art OCR technologies to transform the documents into searchable documents.

But how would the client be able to properly review thousands of documents which are all bundled together in one document? Not only that the file's size was hundreds of MB's, how could someone navigate through thousands of pages and review each document separately?

Our solution - PwC implemented an algorithm which was based on Python programming language. The solution identified where each document begins and ends. By using that logic, we were able to transform one gigantic document into thousands of separate documents, thus allowing the legal team to review each document separately.

PwC's Forensic Technology Solutions work closely with PwC's Data Analytics team. This combination of forces turns our powerful eDiscovery platforms to even more powerful, by harnessing the might of the scripting languages such as Python.

