Srinandhan Ramakrishnan

12/12/2018

**GENE 312 - Project Abstract**

**Automating the process of extraction and printing of keywords using Python**

This project was started with a goal to extract information from KGI’s Team Master’s Project (TMP) contracts. These contracts are secure PDF documents accessible by Corporate Partnerships, an outward looking group responsible for industry relations at the school. A need was felt to extract important information from these TMP contracts (like payments and purchase dates). This information would ultimately help the school keep track of present orders (life science companies sponsor projects with KGI’s MBS students) within their Salesforce database while also enabling quick lookup of historical data. A significant milestone in this regard would be populating information in spreadsheets with “keywords” from these TMP contracts. Being PDF (Portable document Format) documents, copy-paste was the only method existing to extract information. This was a time-consuming task and in case of scanned documents, even copy-paste was not possible. Hence, a python program to extract information from these PDF documents that would then be input into databases like Salesforce, or local spreadsheets was scripted. Three modules were majorly made use of in this endeavour: PyPDF2 (the first-line pdf extraction tool used in python), textract (specializes in OCR and can parse scanned images as well) and NLTK (natural Language Toolkit used in statistical natural language processing). Installation of these modules also required intensive research into OS-specific system behaviours (Windows/Apple), Python-specific package managers like Chocolatey, as well as PDF file encoding and debugging. Using a combination of the modules mentioned earlier, keywords were successfully extracted from parsed PDF documents and printed to a separate spreadsheet for later use.