Reference: automatetheboringstuff.com/chapter13/

From command shell:

python -m pip install --upgrade pip

python -m pip install python-docx

python -m pip install matplotlib

python -m pip install PyPDF2

In Python:

import docx -- for word docs

import PyPDF -- for pdfs

pdfFileObj = open('pdfFileName', 'rb') - opens file ''pdfFileName' in read binary mode

pdfReader = PyPDF2.PdfFileReader(pdfFileObj)

pdfReader.numPages -- return the number of pages in 'pdfFilename'

pageObj = pdfReader.gePage(n) -- creates an object representing the page 'n', zero based index (page 1 is 0 etc)

pageObj.extractText() - returns the text, if possible, in the page object

pdfReader.isEncryped -- checks if file is encrypted

pdfReader.derypt('password')