DOC READING ALGORITHM

Purpose

This program could read several format of files (CV) stored inside the resume folder. If you run the algorithm you can easily capture information like name, email id, address, educational qualification, and experience in seconds from a large number of documents.

Techniques used

This program could read several format of files (CV) stored inside the resume folder.

# Procedure

1. In order for Python to extract information from them, our first step would be to convert them to .txt format. We will use pdfminer (from pdf to text) or Rtf15Reader (from rtf to text) and docx (from docx to text)
2. Breaking entire document into sentences, lines, tokens.
3. Using Regular Expressions in order to capture email, phone, experience and names in the CV.