Project 2: Data extraction using Regex

M. Hassan Shaikh• 10.04.2024

Overview

- Problem statement
- Assumptions
- Solution
- Results

PROBLEM STATEMENT

Part I

- extract candidate names, phone numbers, and email addresses from resume documents.

Part II

- Designing a CSV file format with columns for 'Candidate Name', 'Phone Number', and 'Email ' to store the extracted contact information.

ASSUMPTION

I

- File must be in docx and pdf format. Else an exception mechanism is implemented.
- Contain no images else give error.

II

- Name format:

  David James Austin
- Number format:

  +91 98700 95725
- Email id format:

  ABCD123@XYZ.com

Solution Algorithm

Inputting the resume and installing required libraries

Based on format call function to extract information using regex searching

Check its format whether it is in pdf or doc, raise error if a unsupported format is used

Creating a csv format file and uploading details here

Search algorithm in Regex form

Name

r"([A-Z][a-z]+[\s-]+[A-Z][a-z.]+[\s-]+[A-Z][a-z.])+

Contact

r'''(\d{3}[-\.\s]??\d{3}[-\.\s]??\d{4}|\(\d{3}\)[-\.\s]*\d{3}[-\.\s]??\d{4}|\d{5}[-\.\s]??\d{4})'''

Email

r"[a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+"

Result

Given resume photo

RESUME

Personal Information:

Name: David Jon Clauss
Mobile Number: +91 12345 67890
Email ID: Davidjon2@abmaill.com
Date of Birth: April 10, 2024

Educational Qualifications:

• Graduated with a **Bachelor in Commerce(B. Com)** from Ladhidevi Saraf College of Commerce, *(2020-23)*

• Passed **Higher Secondary Board (HSC)** with 71% at Saraf College of Science and Commerce. *(2018-20)*

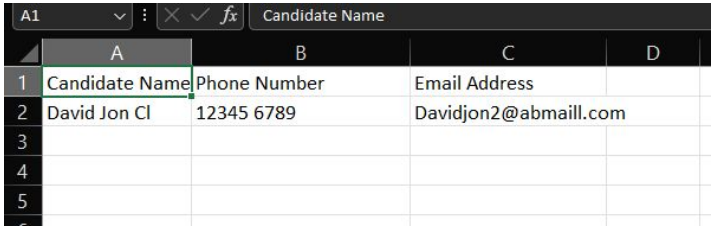
• Passed **Secondary school board(SSC)** with 64% at St. Francis High School. *(2017-18)*

Experience:

Tele Access: *(June 2022-October 2023)*

Worked as **SPOC** at Tele Access Company Private Limited.

Csv pic



DEMO

Thank you

Github link:
https://github.com/hassan-byt0/Mentorness