High-Level-Design (HLD)
&
Low-Level-Design (LLD)

Resume OCR

V1.0
Creator: Sarita Ray

Contents-

1. Introduction:

1.1. Why this HLD & LLD-

The purpose of these HLD & LLD documents is to add the necessary details to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding and can be used as a reference manual for how the modules interact at a high-level as well as low level.

1.2. Scope-

It refers to the component-level design process. A detailed description of each and every module means it includes actual logic for every system component and it goes deep into each module's specification.

2. General Description:

2.1. Product Perspective-

This project is meant to demonstrate machine learning algorithms and techniques to implement an OCR with high accuracy by making use of learning techniques and feature reduction algorithms to make it more efficient.

2.2.Scenario-

Recruiters receive hundreds of CVs in various formats daily, but most of them are not always readable or searchable. A significant chunk for recruiters is the scanned resumes
which are a challenge for recruiters. Building a Resume OCR, a ton of person-hours is saved for the recruiter to cater to potential candidates better. A company can track the quality of applicants over time. Meaningful analytics on candidates can be generated. The main objective here is -

1. Extract relevant data from the resume.
2. After OCR keep it in an excel format for further use.
3. Maintain a database to store each and every data.

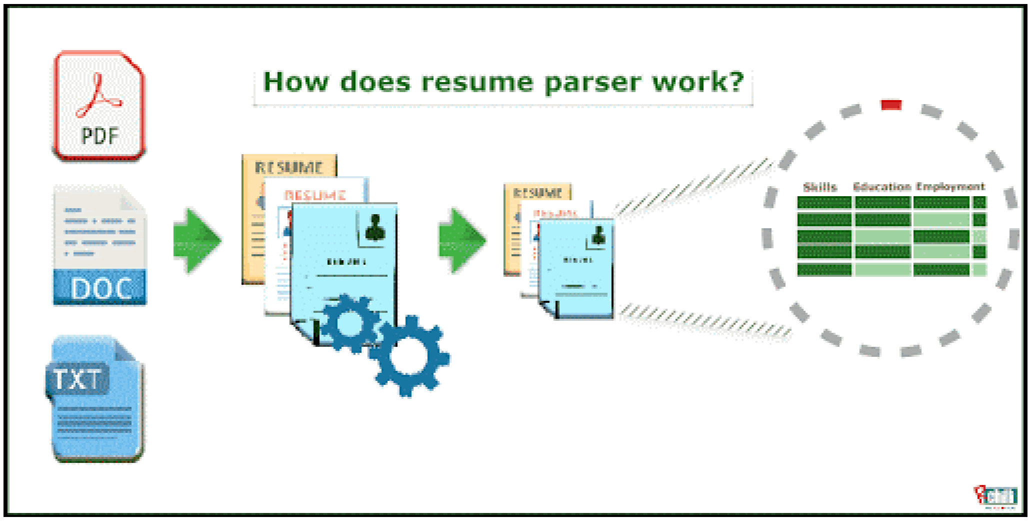
2.3. Resume OCR:

A Resume OCR used for extracting inFormation from Resumes.

2.4. Features-

- Extract name
- Extract email
- Extract mobile numbers
- Extract skills
- Extract total experience
- Extract college name
- Extract degree
- Extract designation
- Extract company names

3. Architecture-



4. Installation-

- You can install this package using

```
pip install pyresparser
```

- For NLP operations we use spacy and nltk. Install them using below commands:

```
# spaCy
python -m spacy download en_core_web_sm


# nltk
python -m nltk.downloader words
```

5. Supported File Formats-

- PDF and DOCx files are supported on all Operating Systems
- If you want to extract DOC files you can install textract for your OS (Linux, MacOS)
- Note: You just have to install textract (and nothing else) and doc files will get parsed easily

6. Usage-

- Import it in your Python project

```
from pyresparser import ResumeParser
data = ResumeParser('/path/to/resume/file').get_extracted_data()
```

7. CLI-

For running the resume extractor you can also use the `cli` provided

```
usage: pyresparser [-h] [-f FILE] [-d DIRECTORY] [-r REMOTEFILE]
                   [-re CUSTOM_REGEX] [-sf SKILLSFILE] [-e EXPORT_FORMAT]


optional arguments:
  -h, --help            show this help message and exit
  -f FILE, --file FILE  resume file to be extracted
  -d DIRECTORY, --directory DIRECTORY
                        directory containing all the resumes to be extracted
  -r REMOTEFILE, --remotefile REMOTEFILE
                        remote path for resume file to be extracted
  -re CUSTOM_REGEX, --custom-regex CUSTOM_REGEX
                        custom regex for parsing mobile numbers
  -sf SKILLSFILE, --skillsfile SKILLSFILE
                        custom skills CSV file against which skills are
                        searched for
  -e EXPORT_FORMAT, --export-format EXPORT_FORMAT
                        the information export format (json)
```

8. Notes-

- If you are running the app on windows, then you can only extract .docs and .pdf files.

9. Result-

The module would return a list of dictionary objects with result as follows:

```
[
  {
    'college_name': ['Marathwada Mitra Mandal’s College of Engineering'],
    'company_names': None,
    'degree': ['B.E. IN COMPUTER ENGINEERING'],
    'designation': ['Manager',
                    'TECHNICAL CONTENT WRITER',
                    'DATA ENGINEER'],
    'email': 'omkarpathak27@gmail.com',
    'mobile_number': '8087996634',
    'name': 'Omkar Pathak',
    'no_of_pages': 3,
    'skills': ['Operating systems',
              'Linux',
              'Github',
              'Testing',
              'Content',
              'Automation',
              'Python',
              'Css',
              'Website',
              'Django',
              'Opencv',
              'Programming',
              'C',
              ...],
    'total_experience': 1.83
  }
]
```

10. References that helped me get here-

- https://www.kaggle.com/nirant/hitchhiker-s-guide-to-nlp-in-spacy
- https://www.analyticsvidhya.com/blog/2017/04/natural-language-processing-made-easy-using-spacy-%E2%80%8Bin-python/
- https://medium.com/@divalicious.priya/information-extraction-from-cv-acec216c3f48
- Special thanks to dataturks for their annotated dataset