Folio3 Software

AUTO-FILLER

FORM AUTO-COMPLETION USING OCR

GROUP MEMBERS:

S M BILAL ARSHAD
MUHAMMAD MUSTAFA BAWANY
MUHAMMAD YASEEN AMIR
IBAD SALEEM

Introduction:

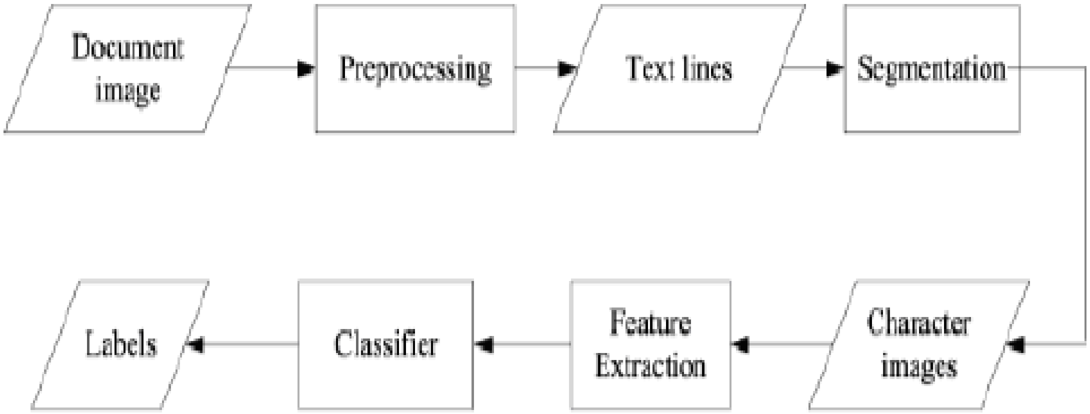
This project is dedicated to creating an OCR software that recognizes text and patterns on resumes and CVs to auto-complete application forms. It is found that many job applicants bounce off job portals with manual forms, so to address this problem we will parse out their resumes.

Problem Statement:

Many job portals still don't have any AI auto-form completion; even if they do, they fail on resumes made with templates. This causes job applicants to fill out and correct the forms, creating a time-consuming activity.

Proposed Solution:

Create an OCR-based AI model, trained on multiple templates from the internet which can parse resumes and can fill out the information like Email, Phone Number, Github, LinkedIn, Education, and Experience.



1. Data gathering and collection
2. Data cleansing and extraction
3. Data manipulation
4. OCR engine training to segment text-lines
5. Feature extraction and classification of text-data
6. Validate via tests

Future Roadmap:

- Create a chrome extension or a web-based application
- Add certifications, honors, and awards sections
- Making code more scalable
- Adding an additional feature of evaluating resumes based on grammar and design, and providing recommendations to improve.

Type of Dataset:

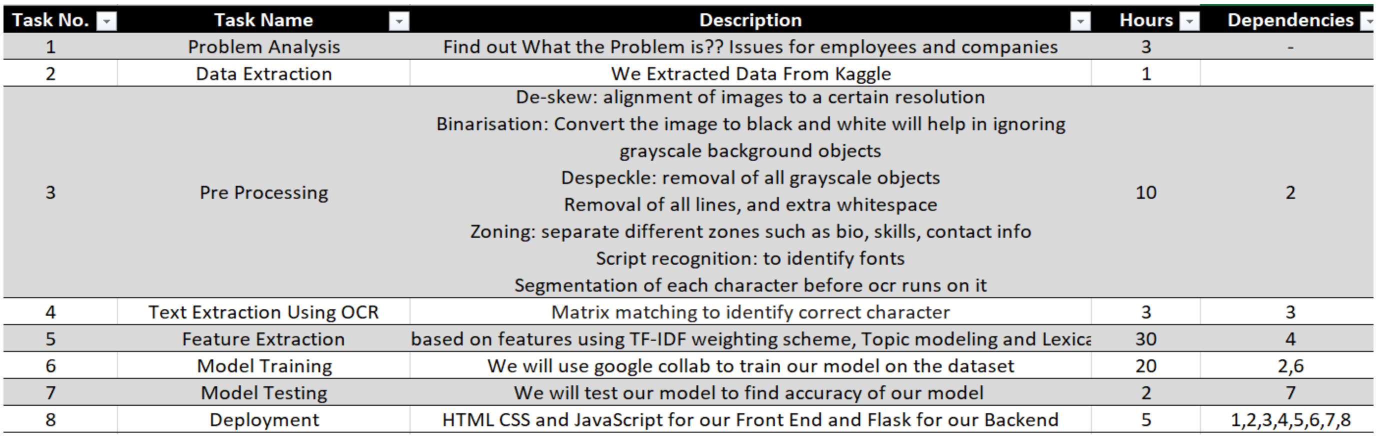
A Filebase Dataset, containing resumes in the form of pdf or word documents. Converting them into images and parsing text from those images through OCR.

Source of Dataset:

The dataset has been extracted from Kaggle. Below is the link to the dataset.

https://www.kaggle.com/datasets/aishikai/resume-dataset

Work Breakdown Structure:



Task No.	Task Name	Description	Hours	Dependencies
1	Problem Analysis	Find out What the Problem is?? Issues for employees and companies	3	-
2	Data Extraction	We Extracted Data From Kaggle	1	
3	Pre Processing	De-skew: alignment of images to a certain resolution Binarisation: Convert the image to black and white will help in ignoring grayscale background objects Despeckle: removal of all grayscale objects Removal of all lines, and extra whitespace Zoning: separate different zones such as bio, skills, contact info Script recognition: to identify fonts Segmentation of each character before ocr runs on it	10	2
4	Text Extraction Using OCR	Matrix matching to identify correct character	3	3
5	Feature Extraction	based on features using TF-IDF weighting scheme, Topic modeling and Lexica	30	4
6	Model Training	We will use google collab to train our model on the dataset	20	2,6
7	Model Testing	We will test our model to find accuracy of our model	2	7
8	Deployment	HTML CSS and JavaScript for our Front End and Flask for our Backend	5	1,2,3,4,5,6,7,8