Data Collection and Preprocessing Phase

Date	15 Nov 2024
Team ID	739938
Project Title	AI Enabled Candidate Resume Screening
Maximum Marks	6 Marks

Preprocessing Template

The images will be preprocessed by resizing, normalizing, augmenting, denoising, adjusting contrast, detecting edges, converting color space, cropping, batch normalizing, and whitening data. These steps will enhance data quality, promote model generalization, and improve convergence during neural network training, ensuring robust and efficient performance across various computer vision tasks.

Section	**Description**
Data Overview	Here we used the resume which consists of information about candidate's profile,Skills,educational qualifications
Resizing	Resizing typically refers to adjusting and standardizing the format of the resumes (often in PDF, Word, or image formats) to ensure consistency and compatibility with the AI system.
Normalization	Standardizing Text formats: Resumes often come in different fonts, sizes, and styles. Normalization involves converting all text into a consistent format, such as lowercase text and uniform fonts, to ensure that the AI system treats all content equally, regardless of formatting. Removing Noise: This includes stripping out irrelevant symbols, extra spaces, and unnecessary characters (e.g., header/footer details or images) to focus on the core content.
Data Augmentation	Data augmentation in AI-enabled candidate resume screening enhances the training dataset by creating synthetic variations of real resumes. Techniques like text paraphrasing, noise injection, and synthetic data generation help the AI model handle diverse resume formats, improve robustness, and

	generalize better to unseen candidate profiles.
Denoising	Denoising in AI-enabled resume screening is crucial for improving the accuracy and effectiveness of the screening process. It involves removing irrelevant, noisy, or low-quality data—such as unnecessary formatting, spelling errors, and non-essential details
Edge Detection	Section Detection: Resumes are typically divided into sections such as Work Experience, Education, Skills, and Contact Information. Edge detection involves detecting the boundaries between these sections, which could be marked by headings like "Experience" or "Education," and helps the AI model structure the resume into meaningful segments for further analysis. Title and Content Seperation: The AI identifies where a section title (e.g., “Skills”) ends and the content (list of skills) begins, ensuring accurate extraction of relevant data.
Color Space Conversion	Simplifying the image: Most resumes submitted in image formats (e.g., PNG, JPEG) may contain color elements such as logos, headers, and various design features. In many cases, color information isn’t necessary for text extraction and may even complicate OCR accuracy. Therefore, converting the image to grayscale(black and white) simplifies the document by reducing unnecessary color details, making it easier for OCR systems to focus on textual content Highlighting the text: Grayscale images help accentuate dark text on light backgrounds, which improves the accuracy of text detection in scanned resumes.
Image Cropping	Cropping Non-Text Areas: Resumes often have large margins or background decorations that do not contain useful text. AI models can automatically identify and crop these areas to reduce the amount of non-essential data. This ensures that the OCR system is not distracted by unnecessary background details. Margin Detection: Some advanced AI models can also detect when a resume has excessive whitespace or irrelevant images and can crop out these areas for improved focus on the resume's textual content.

Batch Normalization	Batch normalization in AI-enabled candidate resume screening is used to standardize the input data, stabilize the training process, and improve model performance. By normalizing the data fed into the neural network, it helps the AI system learn more efficiently, reduces the impact of noisy or inconsistent data (such as varied resume formats), and accelerates the convergence of the model. This leads to a more robust and accurate resume screening process, particularly in systems that handle diverse and complex resume data.
Data Preprocessing Code Screenshots	

Loading Data	`data = ResumeParser('PradeepthiResume.pdf').get_extracted_data()`
	`print(data)`
	`# grab the name`
	`name = data['name']`
	`# grab the Email`
	`email = data['email']`
	`# grab the Skills`
	`skills = data["skills"]`

Resizing	-
Normalization	`# lowercase the skills` `actual_skills = [i.lower() for i in skills ]`
Data Augmentation	-
Denoising	-
Edge Detection	-
Color Space Conversion	-
Image Cropping	-
Batch Normalization	-