Data Collection and Preprocessing Phase

Date	25 Jan 2025
Team ID	740678
Project Title	Amazon Kindle Store Reviews Analysis
Maximum Marks	6 Marks

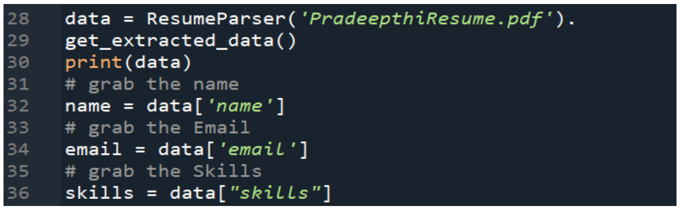
Preprocessing Template

The user reviews will be preprocessed by cleaning, normalizing, tokenizing, filtering noise, lemmatizing, detecting sentiment cues, and converting data formats. These steps enhance text quality, improve model learning, and ensure efficient, robust natural language understanding across diverse review inputs.

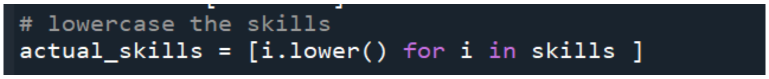
Section	**Description**
Data Overview	Collected review data from the Amazon Kindle Store. Reviews include free-form text content, star ratings, reviewer metadata, and timestamps. The dataset is largely unstructured and varies widely in length and quality.
Resizing	Removal of HTML tags, emojis, special characters, and redundant whitespace.Converting all text to lowercase for uniformity.
Normalization	Standardizing review formats by removing non-informative content (e.g., promotional content or automated disclaimers). Unifying punctuation and token structure for model compatibility.
Data Augmentation	Generating synthetic variations of reviews using paraphrasing or back-translation for model robustness.

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	-