**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 9 JULY 2024 |
| Team ID | INDIVIDUAL |
| Project Title | CareerCraft: ATS-Optimized Resume Analyzer using Gemini Model |
| 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 | The project uses resumes and job descriptions in text format to identify and extract key features like skills, experience, and keywords that are critical for matching resumes to job descriptions. |
| Resizing | Not applicable, as the data consists of text documents rather than images. |
| Normalization | Text data will be normalized by converting it to lowercase, removing special characters, and standardizing whitespace. This ensures uniformity and consistency across the dataset, which is essential for accurate analysis. |
| Data Augmentation | To diversify the text data and improve model generalization, techniques such as synonym replacement, random insertion, and random deletion may be applied. This helps create variations in the text, enhancing the model's ability to handle different inputs. |
| Denoising | Denoising involves cleaning the text data by removing irrelevant information, including extra spaces, HTML tags, special symbols, and other non-text elements that may interfere with processing. |
| Edge Detection | Not applicable for this project, as it deals with text data rather than images. |
| Color Space Conversion | Not applicable, as this project is focused on text data. |
| Image Cropping | Not applicable for text data.. |
| Batch Normalization | Not applicable for this text-based project. |
| **Data Preprocessing Code Screenshots** | |
| Loading Data | The code to load the resume and job description datasets will involve reading PDF and text files into the Python environment using libraries such as PyPDF2 or Textract. |
| Resizing | Not applicable. |
| Normalization | Show a code snippet that converts text to lowercase and removes special characters. |
| Data Augmentation | Include a code snippet demonstrating synonym replacement or other text augmentation techniques. |
| Denoising | Provide a code snippet for cleaning the text data, removing irrelevant characters, and standardizing whitespace. |
| Edge Detection | Not applicable. |
| Color Space Conversion | Not applicable. |
| Image Cropping | Not applicable. |
| Batch Normalization | Not applicable. |