Chandubhai S. Patel Institute of Technology

Computer Science & Engineering Department

Weekly Internship Report

Student ID: 20CS080 Student Name: Jekeel Mayur Shah .
Sem: 7 Div.: CSE - 2 Type of Internship: Project Internship .

Week No.: 2 from: 29th May to 4th June Working Hours / week: 30 (Virtual) .

Company Name: Jio Platforms Limited (JPL) Technology worked on: Python (Machine Learning).

Describe your principal assignments and responsibilities for this period.

During this period, my principal assignment was Data preparation. Our next step was text extraction from invoices in our dataset. I explored different tools such as OCR (Optical Character Recognition), PDF parsers like Tesseract and PyMuPDF to extract text from the invoices. After extraction, I carried out text preprocessing tasks, including removing special characters, escape sequences, tokenization, removing stopwords, and lemmatization. Additionally, I obtained an official email ID from the ril.com domain.

What experiences were particularly rewarding during this report period?

During this report period, I found the experience of exploring various tools for text extraction, such as OCR and PDF parsers, to be particularly rewarding. It allowed me to learn about different techniques and their application in extracting text from invoices efficiently. Additionally, performing text preprocessing tasks to clean and prepare the extracted text for further analysis was satisfying, as it enhanced the quality of the data for subsequent machine learning tasks.

What experiences were particularly difficult during this report period?

One of the challenges I faced was working on virtual desktops connected to the intranet of the company. To gain access to different applications and tools required for the project, a series of procedures had to be completed. This process presented some difficulties initially, as it involved navigating through the company's network and ensuring proper access rights. However, with the guidance of the IT department, I was able to overcome these challenges and successfully set up the necessary environment for my work.

Describe principal tasks and duties to be performed and accomplishments during the upcoming week.

- Vendor Identification based on Logo and template matching of different vendors using deep learning or image processing techniques.
- Documenting the progress made, challenges faced, and any notable accomplishments during the week

Learning Outcomes: (in brief)

- Gaining practical experience in using OCR and PDF parsing techniques for text extraction from invoices.
- Developing proficiency in text preprocessing tasks, including cleaning, tokenization, removing stopwords, and lemmatization.
- Understanding the importance of data quality and preprocessing for machine learning tasks.
- Enhancing skills in data exploration, experimentation, and documentation.

Signature of Internal Guide
Prof. Akshita Kadam

Signature of External Guide
Mr. Milind Mulye