INTERNSHIP REPORT

TITLE: Alstom- Data Preparation

Report by:
Rekadi Varshini
Corp ID – vr80820
Duration: 09-05-24 to 09-07-24

Internship Team Details:

Reporting Manager – Rekadi Varshini

Supervisor – Madhavi Arelli

Technical Lead – Kavya Brungi

Software Engineer – Naveen rao Kasavaram

Software Engineer – Sandeep Sakalabhaktula

Table of Contents:

Company Overview:

CYIENT is a global engineering and technology solutions company headquartered in Hyderabad, India, with a significant presence across North

America, Europe, Asia Pacific, and the Middle East. Founded in 1991 as Infotech Enterprises Limited, the company rebranded as CYIENT in 2014 to reflect its evolution into a comprehensive provider of engineering, manufacturing, geospatial, networks, and operations management solutions. They leverage the power of digital technology and advanced analytics capabilities, along with domain knowledge and technical expertise, to solve complex business problems.

As a Design, Build, and Maintain partner, they take solution ownership across the value chain to help their clients focus on their core, innovate, and stay ahead of the curve. With more than 15,000 employees in 22 countries, they partner with

clients to operate as part of their extended team, in ways that best suit their organization's culture and requirements. Their industry focus spans aerospace

and defence, medical, telecommunications, rail transportation, semiconductor, utilities, industrial, energy, and natural resources.

Acknowledgement:

I am Rekadi Varshini, and I would like to extend my heartfelt gratitude to everyone who has played a role in the successful completion of my internship and the preparation of this report.

First and foremost, I am deeply thankful to Madhavi Arelli, Senior Technical Manager - Automation, for their continuous support, guidance, and valuable feedback throughout my internship period at Cyient Ltd. Their expertise and mentorship have been instrumental in shaping my professional growth and understanding of Data Science.

I would also like to express my sincere appreciation to Kavya Brungi, Data Scientist, for their exceptional leadership, technical expertise, and unwavering support throughout our projects. Their guidance was pivotal in navigating challenges and achieving project goals, and their mentorship has significantly

contributed to my development in Data Science and Natural Language Processing.

I am also grateful to the entire team at Cyient Ltd for their warm welcome, cooperation, and willingness to share their knowledge and experiences. Their support has provided me with invaluable insights into the practical aspects of Data Science.

I am deeply grateful to Gaurav Rohilla, my Reporting Manager at Cyient, whose guidance and support have been invaluable throughout my internship. Gaurav's strategic direction and encouragement enabled me to tackle challenges effectively and grow professionally in the field of Data Science. I appreciate Samit's dedication to fostering a supportive environment where I could learn and contribute meaningfully to the team's objectives.

Thank you all once again for your contributions and support, which have been integral to the successful completion of this internship report.

Sincerely,

Rekadi Varshini

Team automation (AI & ML)

Cyient Ltd Madhapur

Introduction:

Machine learning is a branch of AI focused on creating algorithms that enable computers to learn from and make predictions based on data. It involves techniques like supervised, unsupervised, and reinforcement learning. Applications include predictive analytics, natural language processing, and computer vision. Advances in computational power and data availability drive its continuous innovation.

Data science is an interdisciplinary field that uses statistical, computational, and domain-specific techniques to extract insights from data. It involves data collection, cleaning, analysis, and visualization to support decision-making. Key applications include predictive analytics, machine learning, and data mining. Advances in technology and data availability continue to enhance its impact.

Natural Language Processing (NLP) is a field of AI that enables computers to understand and generate human language. It combines linguistics, computer science, and machine learning to process text and speech data. Applications include chatbots, sentiment analysis, and language translation. Advances in computational power and large datasets drive innovations in NLP.

Internship Responsibilities:

During my tenure as an intern, I undertook the following responsibilities:

1. I was tasked to engage in comprehensive study of NLP materials to understand practical applications in projects, covering methods, formulas, and detailed introductions.

2. Applied NLP and ML techniques to analyze and categorize a dataset of amazon product reviews.

3. Attained proficiency in Python programming to actively contribute to developing automation test scripts.

4. Utilized Visual Studio Code (VS Code) as the primary IDE for authoring, debugging, and executing automation test scripts.

5. Leveraged GitHub Co-pilot, an AI-driven code completion tool, to streamline code generation and automate repetitive coding tasks.

6. I was part of the Design Automation Inputs Siemens.

7. I was a part of the Alstom- Data Preparation.

Learning and Development:

1. **Mastery in Python Programming:** During my internship, I achieved mastery in Python programming by delving into advanced concepts and techniques. This included efficient code development, data structure handling, algorithm implementation, and utilizing Python libraries for data manipulation and analysis.

2. **NLP Foundation:** I established a strong foundation in Natural Language Processing (NLP), focusing on techniques like tokenization, text preprocessing, and sentiment analysis. These skills enabled effective handling of textual data and application of NLP models for tasks such as language modeling and extracting insights from large datasets.

3. **Working on Real-Time Projects:** Working on real-time projects was a highlight, providing hands-on experience in applying Python, Data Science and NLP skills to solve practical challenges. From data extraction to visualization, I contributed to the projects, enhancing technical proficiency and teamwork in delivering impactful solutions.

4. Integration of Code: I prioritized mastering code integration in collaborative environments, using tools like Git and GitHub for version control. This ensured code quality, change tracking, and seamless collaboration, supporting efficient project workflows and team success.

5. Real time use of different packages in python: During the internship, I leveraged a variety of Python packages to tackle diverse challenges effectively. These included OpenCV (cv2) for computer vision tasks like image processing and object recognition, numpy for efficient numerical computations and data manipulation, and pytesseract for optical character recognition, enabling text extraction from images and documents. Additionally, pandas facilitated structured data handling and analysis, while openpyxl streamlined interactions with Excel files for seamless data management. The os module supported platform-independent operations such as file handling. These packages played pivotal roles in real-time projects, enhancing workflow efficiency and analytical capabilities across different domains, highlighting their essential contribution to modern software development practices.

6. Design Automation Inputs Siemens: I was given a compressed file that contains files that are generated by the computer in the pdf format. My task was to extract data from the pdf using the required data extraction techniques for the required folder called testlog register and the goal of the project was to automate the process.
This project after sometime was put on hold.

7. Alstom- Data Preparation: I was given a compressed file that contains 6 excel files with large data and was tasked to extract a particular type of data from the excel files.

After extracting the data using unique keywords specific

to the particular data, I was tasked with the extraction of the remaining data from that particular data after which I was assigned with its preparation after which I was tasked with extracting data by comparing the 2 excel files and extracting the compared data into another excel file. I was also tasked to learn and help whenever together with the AL & ML team to build an image extraction and conversion part of the project.

Challenges and Solutions:

I have encountered several challenges throughout my internship journey and successfully navigated through them. Here are some of these challenges, and the solutions to resolve them.

1. Developing Regex Patterns for PDFs in Different Formats:

Challenge: Defining Regex Pattern for PDF Extraction.

Solution:

To address PDF variability at Siemens, we analysed layout patterns, identified key data points, and crafted flexible regex patterns. These were rigorously tested and integrated into Python-based automation tools for streamlined extraction.

2. Complex Data Extraction from Excel Files:

Challenge: Extracting intricate data from Excel files

Solution:

The challenge of extracting intricate data from Excel files was tackled by leveraging Python's pandas library. This powerful tool allowed for targeted data extraction based on specific criteria, such as filtering rows or columns based on conditions, aggregating data, and handling complex data structures efficiently. Libraries in pandas like openpyxl, numpy and cv2 were used ensuring accurate extraction of detailed information from Alstom's extensive Excel datasets.

3. Verification of Expected vs. Actual Output:

Challenge: Verification of Expected vs. Actual Output

Solution:

Manual validation processes were implemented. This included meticulous manual comparison of data elements, considering case sensitivity and formatting variations. By conducting thorough manual checks and comparisons, discrepancies between expected and actual outputs were identified and resolved to ensure accuracy and reliability in project outcomes.

4. Image Extraction and OCR from Multiple Image Files:

Challenge: Image Extraction and OCR from Multiple Image Files.

Solution:

Image Preprocessing: Each image may vary in quality, requiring preprocessing techniques such as thresholding and contour detection to isolate regions of interest (textual content).

Text Extraction: Utilizing OCR (Optical Character Recognition) with Tesseract to extract text from the detected regions in the images. Handling variations in text size, orientation, and clarity adds to the challenge.

Data Structuring: Once text is extracted, organizing it into a structured format suitable for Excel involves aligning text from various regions (contours) into rows and columns.

Conclusion:

My internship journey at Cyient Ltd. has been a transformative experience, providing me with invaluable opportunities to delve into the realms of Data Science, Machine Learning, and Natural Language Processing (NLP). Throughout this period, I have not only enhanced my technical skills but also gained practical insights into real-world applications across diverse projects.

At Cyient, I had the privilege of working alongside talented professionals who provided continuous guidance and support. Special thanks to Madhavi Arelli and Kavya Brungi for their mentorship in Automation and Data Science, respectively. Their expertise and encouragement were instrumental in my professional growth and understanding of complex technologies.

The challenges I encountered, from developing regex patterns for PDF extraction to tackling complex data structures in Excel files, further strengthened my problem-solving abilities. Each challenge was met with a systematic approach and innovative solutions, showcasing my adaptability and commitment to delivering high-quality results.

Moreover, my involvement in projects like Design Automation Inputs for Siemens and Data Preparation at Alstom allowed me to apply theoretical knowledge to

practical scenarios. I honed my skills in Python programming, data manipulation with libraries like pandas and numpy, and image processing using OpenCV and Tesseract OCR.

As I reflect on this journey, I am grateful for the collaborative environment at Cyient, where teamwork and knowledge-sharing were integral. The experience has not only equipped me with technical proficiency but also instilled in me a deeper appreciation for continuous learning and innovation in the field of technology.

In conclusion, my internship at Cyient has been a significant stepping stone in my career, preparing me to tackle future challenges with confidence and enthusiasm. I look forward to applying the skills and insights gained here to contribute meaningfully to future endeavors in Data Science and AI.