Project Weekly Progress Report Agile – Scrum

|  |  |
| --- | --- |
| **Semester** | **W2023, SEM-2** |
| **Course Code** | **AML-2404** |
| **Section** | **Section 2** |
| **Group Name** | **D** |
| **Student names/Student IDs** | **Jash Vaghasiya - C0884733**  **Nivedini Kathagonda - C0872720**  **Keval Parmar - C0882386**  **Monil Rupawala - C0882370**  **Sai Divya Madhuri Guntupalli - C0882360** |
| **Reporting Week** | **6** |
| **Team Lead for the reporting week** | **Keval Parmar** |

**1. Progress Made in Reporting** **Week:**

In our pursuit to extract data from the PDF document obtained from the European nations' website, we have made notable progress despite the challenges posed by the missing side borders in the tables. We have implemented a range of strategies and techniques to tackle this issue and improve the extraction process.

One of the most significant advances we have made is the effective application of optical character recognition (OCR) technology. We were able to retrieve the material more effectively by converting the scanned PDF into editable text using OCR. This phase was critical in converting unstructured data into a more organised format, establishing the framework for subsequent research.

**Flowchart of Extracting data from PDF:**

A picture containing text, screenshot, font, design

Description automatically generated

Additionally, our investigation into image processing methods has produced encouraging outcomes. We have been able to infer and recreate the missing side borders to a considerable extent using techniques like edge detection and contour analysis. Although the success rate varied depending on how complicated the table layouts were, this method has given us useful insights and helped us get closer to accurately extracting the data.

Our efforts to implement machine learning algorithms have also been promising. We have been able to forecast the missing side borders by using data that is currently accessible and training models using existing table structures. This method has gradually increased the precision of our extraction findings, hence boosting the correctness of the extracted ingredient data.

Overall, our progress in the data extraction process has been significant. We have successfully employed OCR, image processing, and machine learning techniques to overcome the challenges presented by the missing side borders. These advancements have brought us closer to achieving our goal of extracting the ingredient data from the PDF document obtained from the European nations' website.

**2. Difficulties Encountered in Reporting Week:**

Even while we have made significant progress in extracting data from the PDF file, we have run into a few problems, mostly because the tables' missing side borders. The accurate extraction and organisation of the ingredient information have been significantly hampered by these problems.

The tables' lack of full side borders has thrown off the data's alignment and organisation. It has been difficult to precisely separate the table columns due to the lack of distinct borders. As a result, it has been challenging for our extraction methods to distinguish between neighbouring columns and precisely classify the retrieved data.

In conclusion, the process of extracting data has been significantly hampered by the absence of entire side borders in the tables of the PDF document. The difficulties in effectively classifying and organising the data, especially in intricate table layouts, have been a recurring difficulty. Even though OCR, image processing, and machine learning approaches have helped, these issues still affect the dependability and accuracy of the information that is recovered. As we work to improve the extraction procedure and raise the calibre of the extracted ingredient data, addressing these difficulties is still a top focus.

**Picture of PDF:**

A picture containing text, number, receipt, parallel

Description automatically generated