Project Weekly Progress Report Agile – Scrum

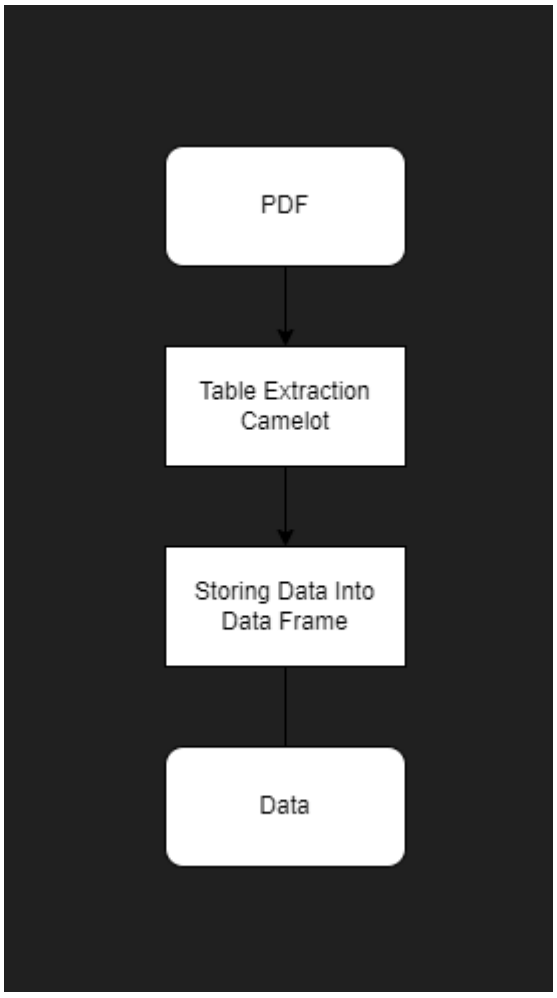
Semester	**W2023, SEM-2**
Course Code	**AML-2404**
Section	**Section 2**
Group Name	**D**
Student names/Student IDs	**Jash Vaghasiya - C0884733** **Nivedini Kathagonda - C0872720** **Keval Parmar - C0882386** **Monil Rupawala - C0882370** **Sai Divya Madhuri Guntupalli - C0882360**
Reporting Week	**6**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



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:

INCI name	INN name	Ph. Eur. Name	CAS No	EINECS/ELINCS No	Chem/IUPAC Name	Restriction	function
ABIES BALSAMEA EXTRACT			85085-34-3	285-364-0	Abies Balsamea Extract is an extract of the sprouts of Abies balsamea, Pinaceae		Film forming/hair conditioning
ABIES PECTINATA EXTRACT			92128-34-2	295-728-0	Abies Pectinata Extract is an extract of the bark and needles of the silver fir, Abies pectinata, Pinaceae		Tonic/deodorant
ABIES PECTINATA OIL			92128-34-2	295-728-0	Abies Pectinata Oil is the volatile oil obtained from the needles of the silver fir, Abies pectinata, Pinaceae		Tonic/masking
ABIES SIBIRICA OIL			91697-89-1	294-351-9	Abies Sibirica Oil is the volatile oil distilled from the needles and branches of Abies sibirica, Pinaceae		Tonic/masking
ABIETIC ACID			514-10-3	208-178-3	Abietic acid		Emulsion stabilizing
ABIETYL ALCOHOL			666-84-2	211-544-4	[1R-(1.alpha.,4a.beta.,4b.alpha.,10a.alpha.)]-1,2,3,4,4a,4b,5,6,10,10a-decahydro-7-isopropyl-1,4a-dimethylphenanthren-1-methanol		Viscosity controlling
ACACIA CATECHU			8001-76-1	232-291-7	Acacia Catechu is the dried, crushed core of Acacia catechu, Leguminosae		Hair dyeing/astringent
ACACIA CONCINNA EXTRACT			202148-87-6		Acacia Concinna Extract is an extract of the fruit of Acacia concinna, Leguminosae		Skin conditioning
ACACIA DEALBATA EXTRACT			165800-52-2		Acacia Dealbata Extract is an extract of the leaves of the wattle, Acacia dealbata, Leguminosae		Skin conditioning
ACACIA DECURRENS EXTRACT			98903-76-5	308-877-4	Acacia Decurrens Extract is an extract of the sprouts of the acacia, Acacia decurrens, Leguminosae		Tonic
ACACIA FARNESIANA EXTRACT			89958-31-6	289-655-3	Acacia Farnesiana Extract is an extract of the flowers and stems of the acacia, Acacia farnesiana, Leguminosae		Viscosity controlling/astringent
ACACIA FARNESIANA GUM			2593593	232-519-5	Acacia Farnesiana Gum is a plant material derived from the dried, gummy exudate of acacia, Acacia farnesiana, Leguminosae		Viscosity controlling/astringent

5.4.2006 EN Official Journal of the European Union