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	**8**
Team Lead for the reporting week	**Jash Vaghasiya**

1. Progress Made in Reporting Week:

The preprocessing stage of the data received from PDF documents of European skincare items saw substantial advancements during week 8 of the project. Due to the mismatch between our recommendation system and intended recommendations, we chose to use this resource. We used the Camelot library to extract text from the PDFs as part of the data preprocessing. Row by row, the extraction procedure produced separate rows for constituents with descriptions spanning more than two lines. This development has created the framework for more data analysis and improving our recommendation system.

	0	1	2	3	4	5	6	7	8	9	10
0	ABIES BALSAMEA			85085-34-3	285-364-0	Abies Balsamea Extract'nis an extract ofinthe		Film forming/hair conditioning	NaN	NaN	NaN
1	EXTRACT					balsamea, Pinaceae			NaN	NaN	NaN
2	ABIES PECTINATA			92128-34-2	295-728-0	Abies Pectinata Extract'nis an extract ofinthe		Tonic/deodorant	NaN	NaN	NaN
3	EXTRACT					ofinthe silver'infir, Abies pectinata, Pinaceae			NaN	NaN	NaN
4	ABIES PECTINATA OIL			92128-34-2	295-728-0	Abies Pectinata Oil'nis the volatile oil obtai		Tonic/masking	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...
14534	EXTRACT					Zizyphus joazeiro, Rhamnaceae		NaN	NaN	NaN	NaN
14535	ZIZYPHUS JUJUBA				90045-99-1\n289-966-4	Zizyphus Jujuba Extract'nis an extract ofinthe	Skin conditioning	NaN	NaN	NaN	NaN
14536	EXTRACT					jujube, Zizyphus jujuba, Rhamnaceae		NaN	NaN	NaN	NaN
14537	ZIZYPHUS SPINA CHRISTI					Zizyphus Spina Christi'nis the dried crushed l	Skin conditioning'	NaN	NaN	NaN	NaN
14538						Zizyphus spina christi, Rhamnaceae		NaN	NaN	NaN	NaN

2. Difficulties Encountered in Reporting Week:

However, dealing with the problem of separating ingredient information into different rows is now proving difficult. Camelot's text extraction process resulted in other rows for ingredients on numerous lines, which caused the data to be represented in pieces. This makes it challenging to match the correct elements with the relevant goods. We investigate various options, including text processing methods and NLP tools, to detect and reconstruct the fragmented ingredient information reliably. It is imperative to address this issue to guarantee the integrity and dependability of the data used for the recommendation system.

Prepared by William Pourmajidi

```
array([nan, 'Skin conditioning', '', 'Tonic/masking',
       'Emollient/tonic/masking', 'Stabilising', 'Emollient', 'Solvent',
       'Emollient/skin conditioning', 'Uv absorber/uv filter',
       'Propellant', 'Solvent/surfactant/viscosity con-', 'trolling',
       'Preservative/solvent', 'Antistatic/emollient/skin condi-',
       'tioning/hair conditioning', 'Emulsifying/emollient/solvent',
       'Emulsifying', 'Emollient/emulsion stabilising',
       'Viscosity controlling/emulsifying', 'Hair dyeing',
       'Emollient/skin conditioning/sooth-', 'ing'
       'Emollient/moisturising', 'Film forming/skin conditioning/',
       'hair conditioning', 'Antistatic/skin conditioning/hair',
       'conditioning'], dtype=object)
```

The ingredient information must also be included in the recommendation system's later phases and the data preparation step. It affects the feature extraction process since splitting the ingredients across numerous rows makes extracting significant characteristics difficult. It also makes it harder to provide recommendations that accurately match user input with the right products. To manage the fragmented ingredient data effectively and guarantee the development of reliable suggestions for consumers, the solution to this challenge necessitates thorough analysis and experimentation with various methodologies and algorithms.