Stibo Recommendation Framework

Eram Khan, Ashley Stroup, Kumar Subramanyan

MGT 6748 – Group 4

ABSTRACT

Stibo's goal of automating the process of selecting certifications based on product data is a challenging problem. Each product can be unique in underlying characteristics, so it is difficult to map all attributes of a product directly with a certificate. However, leveraging term frequency – inverse document frequency, which is an information retrieval/text mining method, allows for a faster approach. The key idea in the approach is to extract relevant/unique features from structured product data and unstructured text data on ESG (Environmental, Social, and Governance) certificates. Certificates are then shortlisted for each product based on similarities in terms of category/sub-category, product type, usage, and country of origin.

This approach also prompts the user to see which initial certifications should be investigated further. User feedback will enhance the process by incorporating the feedback into the recommendation system.

1. PROJECT OVERVIEW

1.1 Background

Stibo Systems "unifies and governs data with master data management" (Stibo A/S). This organization works with manufacturers, distributors, retailers, and service providers to manage data transparency. With the master data, it allows companies to have an accurate view of their data to make better business decisions and help with digital transformation.

The goal for this project is to empower Subject Matter Experts (SMEs) by building a user-friendly sustainability certificate recommendation system for various products. This will reduce the amount of time it takes to produce the correct certifications while reducing the risk of making an incorrect assessment. The recommendation engine needs to be an automated process that does not rely heavily on SMEs; however, SME feedback will be incorporated to evolve the recommendation engine over time. The first phase of this project had a focus on electronic products only.

1.2 Literature Review

Other companies like KPMG are also looking into certificate suitability. They recommended a stage-based algorithm to determine which certifications are relevant out of the more than 300 certifications worldwide (Bornhauser). They recommend segmenting the certifications into tiers based on markets, brands, customers, governance, people, etc. While Stibo is helping companies determine which certifications apply to what products, companies have an increased focus on electronics and related environmental outputs like carbon footprint and maintaining CO_2 neutral operations (Helmold).

As well as understanding company best practices and latest innovations, software methodologies were researched. Natural language processing (NLP) is a great tool to help scan through certification data. Nadeau & Sekine provided their input on Named Entity Recognition (NER) techniques and off the shelf capabilities. A supplementary technique investigated was a string match. This allows for the ability to join strings based on a fuzzy match to identify ways to match product related features with certificate related features (Wang, Li, & Feng). This approach will be used and discussed in further detail.

2. METHODOLOGY

2.1 Exploratory Data Analysis

2.1.1 Exploratory Data Analysis – Product Data

Stibo provided over 120,000 example product data to build the recommendation engine in the domestic appliance JSON files. All product data that was empty was removed, and the product data was combined into one readable data frame. This allows entities to be extracted faster that fit into a particular category. The top three categories from the JSON product data were washing machines, dishwashers, and fridge-freezers. A clustering approach was considered for certifications that are more data driven like Energy Star. For example, Energy Star provides a list of refrigerators and whether they are Energy Star certified. As shown in Figure 1, a clustering approach does not work well with this data set and goal.

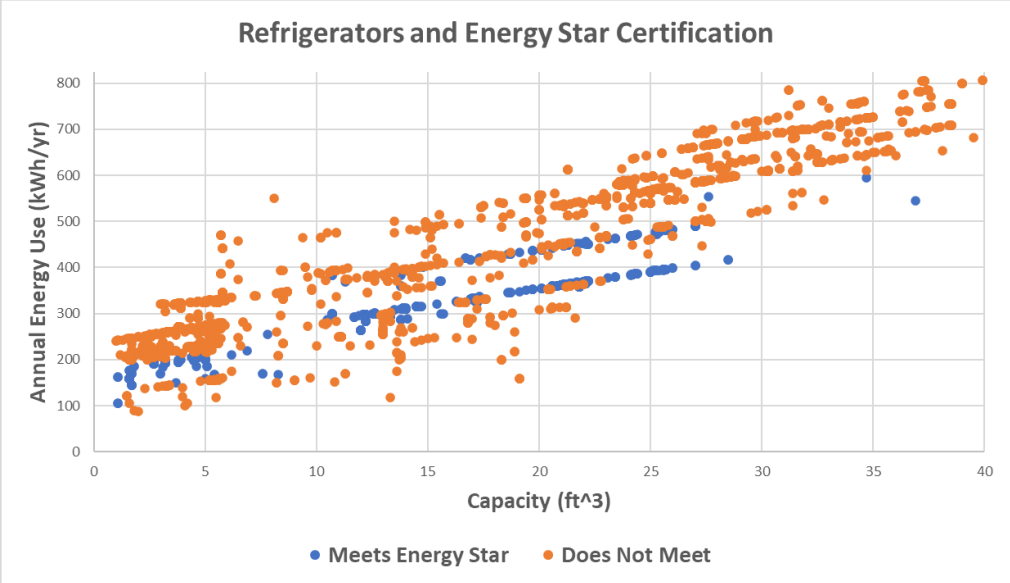

Figure 1: Clustering Example

2.1.2 Exploratory Data Analysis – Certification Data

Certification data and product relevance were explored from Ecolabel as well as individual certificate documents. Certifications related to electronics were pulled out as certifications to focus on. The certifications were then classified in specializations and applicable industries, products, or processes. Unique keywords were determined for each certification based on term frequency – inverse document frequency. The category and applicable country were also extracted from Ecolabel. During this analysis, it was identified that many certificates, particularly linked to manufacturing has strong demarcation based on Geography, for example, Energy Label is focused primarily on China and Taiwan. On parsing though different certifications related to Electronics it was identified that most were applicable in certain countries – United States of America (9), Canada (5), Germany (4)

2.2 Model Survey

Multiple techniques were used to identify relevant text from both structured and unstructured data sources.

- **TF-IDF Scores**: Term Frequency – Inverse Document Frequency is a statistic used for information extraction from text. Typically, it is used to identify the importance of a word within a document relative to other documents in the analysis.
    - TF (Term Frequency = $\frac{Word\ Freq.\ in\ Doc.}{Total\ \#\ of\ Words}$

- IDF (Inverse Doc. Freq.) = $\log\left(\frac{\#Docs.\ under\ Consideration}{\#Docs.\ Containing\ the\ Word}\right)$
- TF-IDF = TF * IDF

- **Named Entity Recognition**: It is an NLP technique used to identify distinct types of entities within a text and classify them into defined categories. Post tokenization, Bidirectional Encoder Representations from Transformers or BERT is leveraged to identify complex patterns within text.
- **Fuzzy String Match**: Multiple algorithms were explored to determine match confidence between keywords related to product and certifications. TFIDF based match was then selected based on initial performance.
  - **Sequence Matcher:** It is based on sequence-based match between two strings in terms of characters or symbols.
  - **TFIDF:** TFIDF vector is generated for the sequences to be compared. Match is then determined using cosine similarity.
  - **Fuzzy Matcher:** Levenshtein Distance is used to determine similarity between strings. It refers to the number of single character changes needed to convert one string to another.

3. IMPLEMENTATION AND RESULTS

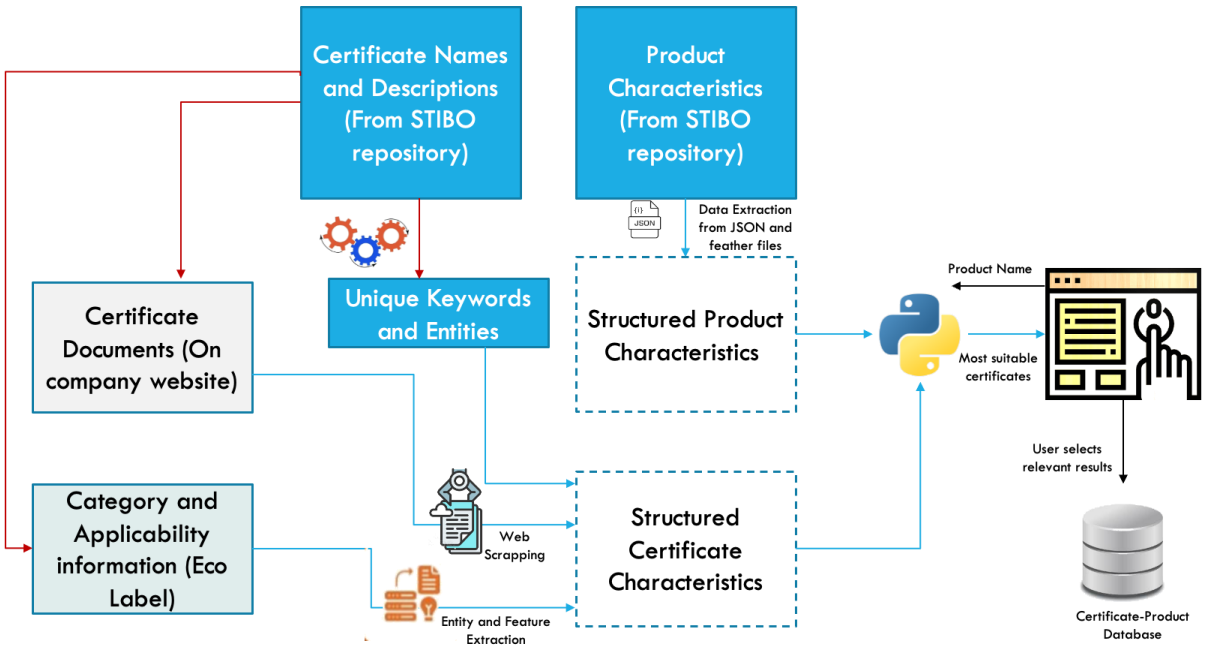

Figure 2: Solution Overview

3.1 Data Input

Data is read in from two different csv files containing product data and certification information, respectively. For the product data, each feather or JSON file was iteratively read with the information extracted into one data frame. Figure 3 shows how the data is accumulated in the data frame. Each product is one row with the name, title, and category being consistent across all entities. The remaining columns are for unique arguments per product; however, if products belong to the same category, they tend to have the same arguments listed. The argument gives what attribute the value belongs to, and the value provides the details. For example, the argument could be product size, and the value would be 60 cm.

Name	Title	Category	arg1	val1	arg2	val2	arg3	val3
WEG 675 WPS	Miele WEG 675 WPS washing machine Front-load 9...	Washing Machines	Appliance placement	Freestanding	Loading type	Front-load	Product colour	White
DFO 3T133 A F X	Indesit DFO 3T133 A F X Freestanding 14 place ...	Dishwashers	Appliance placement	Freestanding	Product size	Full size (60 cm)	Custom panel-ready	N
WGG24400FR	Bosch Serie 6 WGG24400FR washing machine Front...	Washing Machines	Appliance placement	Freestanding	Loading type	Front-load	Product colour	White
RT35A5930S8	Samsung RT35A5930S8 fridge-freezer Freestandin...	Fridge-Freezers	Appliance placement	Freestanding	Product colour	Stainless steel	Built-in display	N

Figure 3. Product Data Extraction

For the certification data, unique keywords were extracted based on the certificate descriptions and related documents using their respective TFIDF score. Entities were extracted using Named Entity Recognition. Also, product and category relevance from Ecolabel was acquired using web scrapping. All data points were then combined. The final data used for the certification data frame can be seen in Figure 4.

	Unnamed: 0	certification	text	star_char	end_char	label_
0	0	R2ÊCertificationÊ	REUSE & RECYCLING	30	47	ORG
1	1	R2ÊCertificationÊ	R2	50	52	PRODUCT
2	2	R2ÊCertificationÊ	STANDARD	54	62	ORG
3	3	R2ÊCertificationÊ	Version	67	74	ORG
4	4	R2ÊCertificationÊ	July 1,	96	103	DATE
...	...	...	...	...	...	...
12923	19	B Corporation	Ecolabel	4845	4853	PERSON
12924	20	B Corporation	Standard	5396	5404	PERSON
12925	21	B Corporation	Organizational	6571	6585	ORG
12926	22	B Corporation	year	6604	6608	DATE
12927	23	B Corporation	Advertisement Subscribe	6783	6806	PERSON

TF-IDF Score

Certification	Keyword	TFIDF Score
R2 Certification	electronics	0.34363801
R2 Certification	practices	0.29398769
R2 Certification	repair	0.29398769
R2 Certification	used	0.25602615
R2 Certification	equipment	0.22909201
R2 Certification	recycling	0.19113047
R2 Certification	common	0.14699385
R2 Certification	disassembling	0.14699385
R2 Certification	documentation	0.14699385
R2 Certification	exported	0.14699385
R2 Certification	general	0.14699385

On comparing descriptions of all certifications, keywords like "electronics", "recycling", "disassembling" etc have got the highest tfidf score for R2 Certification

Entities Extracted

Certification	Text
R2 Certification	REUSE & RECYCLING
R2 Certification	STANDARD
R2 Certification	Version
R2 Certification	SERI
R2 Certification	SERI Version
R2 Certification	U.S. Copyright Office
R2 Certification	the R2v3 Standard
R2 Certification	EULA
R2 Certification	SERI
R2 Certification	DATA SECURITY
R2 Certification	The Sustainable Electronics Reuse & Recycling
R2 Certification	IT Asset Disposition
R2 Certification	the R2 Technical Advisory Committee

Based on Documentation of different certifications, all entities mentioned were extracted using Named Entity Recognition

Ecolabel Tags

Certification	Ecolabel Tags
Carbon Reduction Label	Appliances
Carbon Reduction Label	Building products
Carbon Reduction Label	Carbon
Carbon Reduction Label	Cleaning products
Carbon Reduction Label	Cosmetics / Personal care
Carbon Reduction Label	Electronics
Carbon Reduction Label	Energy
Carbon Reduction Label	Financial services
Carbon Reduction Label	Food
Carbon Reduction Label	Forest products / Paper
Carbon Reduction Label	Health care services & equipment
Carbon Reduction Label	Machinery & Equipment
Carbon Reduction Label	Textiles

Data was scrapped from Ecolabel website to identify tags associated with different certifications.

Figure 4. Certificate Extraction

Once all the data is obtained, the user will select the relevant product attributes to start the matching process. The recommendation will be for name, title, and category. The user is then asked which product they would like to certify. It could be a product name or title. With all relevant information, a new data frame is created by collating at the product and certificate level.

3.2 Certification Identification

Once all relevant data and certificates are collected into a dataframe, the certification process for a product will start. The user will enter the product for which the user is looking for a certificate, with an example being WEG 675 WPS. Under normal circumstances, the person performing this certification will have a challenging time, as there is nothing clear from the name of the product as to what this product is and what kind of certification can be relevant for this product. The same algorithm based on inverse document frequency is used to find the product

entered by the user in the database created in Section 3.1. The process will analyze the dataset and will look up the names of the certificates. It will give a confidence score for each of the matches.

There is a threshold of zero that was set for the confidence level. As the algorithm and usage becomes more mature, the confidence level can be increased naturally over time or can be increased rapidly by user input. In this example, the algorithm returns five possible certifications for this product with a confidence level higher than zero as shown in Figure 5.

```
                                      certification  \
846                                Carbon Reduction Label
4174                             e-Stewards Certification
6308   Carbon Neutral Product and Carbon Neutral Service
11775                                     Energy Star
12651                                         80 PLUS

                                    concatenated_text  conf_tfid_score
846    3.1 3.1 Textile Exchange 2                  ...         0.000816
4174   Ethical Responsible Reuse, Recycling Dispositi...         0.000320
6308   CarbonNeutral 1 July 2023 Community Reforestat...         0.000396
11775  September 2021 U.S. The Licensed Professional'...         0.000867
12651  80 PROGRAM OVERVIEW  Doug McIlvoy June 2013 2 ...         0.006885
```

Figure 5. Confidence Scores

Product	Certificate Names	Confidence Metric
Hisense RB390N4AC21 fridge-freezer Freestanding 300 L E Grey	B Corporation	0.03111
	Energy Star	0.00515
	Cradle to Cradle Certified(CM) Products Program	0.00498
	AENOR Medio Ambiente	0.00382
	Carbon Neutral Product and Carbon Neutral Service	0.00080
Beko SLMP07W1 tumble dryer Freestanding Front-load 7 kg A+ White	80 PLUS	0.02130
	e-Stewards Certification	0.00869
	Energy Labels	0.00319
	Carbon Neutral Product and Carbon Neutral Service	0.00130
Whirlpool AKM 291/IX Stainless steel Built-in 116 cm Gas 4 zone(s)	Cradle to Cradle Certified(CM) Products Program	0.00856
	UL GREENGUARD Certification	0.00468
	Carbon Neutral Product and Carbon Neutral Service	0.00424
	Energy Star	0.00127
	e-Stewards Certification	0.00067
Siemens SX56P596EU dishwasher Semi built-in 13 place settings	UL GREENGUARD Certification	0.01311
	Cradle to Cradle Certified(CM) Products Program	0.00367
	Energy Star	0.00169
	Carbon Reduction Label	0.00124
BKM-220D - Decoder	e-Stewards Certification	0.00869
	Energy Labels	0.00319
	Carbon Neutral Product and Carbon Neutral Service	0.00130

Figure 6. Sample Results

4. DISCUSSION

4.1 User Input

This application requires user input and feedback to work appropriately. The user input is the product being certified and the relevant product attributes to consider. However, the user is also being asked to provide feedback on which certification is the best fit for the product listed. This can evolve into choosing the appropriate certifications from the list provided or even including a text box on why certain certifications are not appropriate for the product input. This feedback can be incorporated into the product dataframe to be used for future certifications involving specific brands or categories.

4.2 Feasibility of Current Approach

The application explored the beginning stages of a recommendation engine to ease the burden on Subject Matter Experts. This looked at finding relevant certifications based on term frequency – inverse document frequency. This is an important first step; however, many certifications can only be applied if certain product attributes are applicable. Using Energy Star and refrigerators as an example, the current criteria level is 10% less measured energy use than the federal efficiency standards. The efficiency of the refrigerator would need to be readily on hand in the product dataframe or would need to be calculated. If certifications are more text based, this recommendation engine would be a great tool.

4.3 Recommendations for Future Application Iterations

The recommendation for future iterations would be to include more certified products. If the engine can also match the user input product with a certified product that has similar attributes, the output would be the certifications of the certified product. Also, the product dataframe would need additional product attribute data for certain certifications. The example provided in Section 4.2 shows how effective the engine could become if more product data is available. Once sufficient data is collected, a classification model can be trained based on user feedback and incorporated in the current framework. If there is sufficient user feedback available for a certain product, certification can be predicted using the text-based classification model.

5. CONCLUSION

Correctly certifying products is important not only for the companies but also for the consumers of these products. By creating a more automated process, Subject Matter Experts could focus their attention on ensuring that the engine was performing as expected. The current recommendation engine uses a text mining approach via term frequency – inverse document frequency to match products with certifications. The user will then pick the certification that most closely matches their expectations for the product.

6. REFERENCES

[1] A/S, S. S. (n.d.). Cloud-Native Master Data Management Solutions. www.stibosystems.com Retrieved October 22, 2023, from https://www.stibosystems.com/solution

[2] Bornhauser, F. (n.d.). *Sustainability standards and labels*. Retrieved October 22, 2023, from https://assets.kpmg.com/content/dam/kpmg/ch/pdf/kpmg-ch-eco-labels-sustainability-standards-labels.pdf

[3] Helmold, M. (2023). Quality Management (QM). *Management for Professionals*, 1–13. https://doi.org/10.1007/978-3-031-30089-9_1

[4] Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticæ Investigationes. International Journal of Linguistics and Language Resources*, *30*(1), 3–26. https://doi.org/10.1075/li.30.1.03nad

[5] Wang, J., Li, G., & Feng, J. (2014). Extending string similarity join to tolerant fuzzy token matching. *ACM Transactions on Database Systems*, *39*(1), 1–45. https://doi.org/10.1145/2535628