**LCA Data Augmentation**

**Executive summary**

The LCA data augmentation tool aims to the query the existing internal database using AI/ML based model and retrieve the data with the most similar component description matches and provide the similarity score. With current intelligent solution manual effort of LCA engineer can be reduced from 2hrs to 5mins per LCA.

**State of the art/ shortcomings**

Life Cycle Assessment (LCA) is a systematic method used to evaluate the environmental impacts of a product, service, or process throughout its entire life cycle. The goal is to assess all relevant stages, from raw material extraction to production, use, and disposal, commonly referred to as the "cradle-to-grave" approach. The increase in Environmental Product Declaration (EPD) requests from distributors and customers reflects a growing demand for transparency in the environmental impacts of products. To meet this demand, companies need to adapt their processes to efficiently generate and publish EPDs without compromising on accuracy or compliance. one of the compliance requirements is that the error in actual weight of the assembly and the sum of the weight of parts in the LCA inventory must be within a margin of ±5%. This requirement ensures that the weight calculated in the LCA inventory closely matches the actual, physical weight of the final product. In Life Cycle Assessment (LCA), both the weight and the material of a part have a significant impact on the overall environmental footprint.

When an engineer receives a new LCA (Life Cycle Assessment) request, they must manually search the existing internal database to find exact or most similar parts in order to extract data such as weight, material, vendor etc. Users must manually search through large datasets and interpret the results themselves using a lookup table. If they need related terms or synonyms, they must manually adjust their queries and repeat the search. Excel does not learn from previous searches or user preferences, treating each search as independent without improving results over time. This lack of flexibility makes searching for relevant matches in large datasets especially time-consuming, particularly when the dataset is complex. While manual Excel searches are a simple and cost-effective solution for small, structured datasets where users know exactly what they're looking for, they become inefficient, slow, and error-prone in larger, more complex datasets like those used in LCA. Human error, such as overlooking key data or misinterpreting results, is also more likely in these scenarios.

**Invention and how it solves problems**

NLP-based semantic search offers a significantly more powerful, efficient, and accurate method for retrieving data, especially within large and complex datasets like those found in LCA databases. This approach is particularly useful for addressing challenges related to out-of-vocabulary (OOV) words, synonyms, and context-sensitive information.

Traditional keyword searches often struggle with variations in terminology and context. NLP-based semantic search, however, leverages word embeddings to understand the nuanced relationships between words. For example, the term "Capacitor" can be linked to related units like "uF" and "pF", as well as associated materials like "Cer". word Embeddings are vector representations of words that capture their meanings and relationships. In the current application, these embeddings are generated using the FastText model. FastText, developed by Facebook AI Research (FAIR), is an open-source, free library renowned for its efficiency in training word vector models. It can process up to 1 billion words in under 10 minutes, making it exceptionally fast. FastText employs unsupervised learning to create vector representations for words. This means it can learn from large amounts of text data without requiring labeled examples, making it versatile and adaptable to different contexts and datasets.

FastText improves on traditional word embedding techniques by incorporating sub-word information. This allows it to capture the morphological relationships between words and their components (sub-words), providing a more detailed and accurate representation of word meanings. one of FastText’s significant advantages is its ability to handle misspellings and variations in morphologies of a word. By using sub-word based embeddings and an n-gram technique, FastText can recognize and accurately process words even when they are misspelled or presented in different forms. This robustness makes it particularly valuable in datasets where typographical errors or alternate spellings are common.

FastText can generate embeddings for words that are rare or not present in the training data (out-of-vocabulary words). This capability is crucial for applications where new or infrequent terms need to be understood and incorporated into the semantic search. even if extensive pre-processing is not performed, FastText maintains high accuracy in generating word embeddings. This flexibility is beneficial in practical applications where pre-processing may be limited or infeasible.

Once the model has been trained, it is integrated with a user-friendly Graphical User Interface (GUI) to ensure ease of use and accessibility for users. This integration allows users to interact with the model efficiently without requiring specialized technical knowledge. when a new LCA (Life Cycle Assessment) request is submitted, the model performs an automated search through the existing database. It systematically examines each component description from the database in a loop to identify the most similar matches to the components listed in the new request. For each component in the new request, the model provides not only the most similar match from the database but also a similarity score indicating how closely the match aligns with the requested component.

In this specific case, the model was trained using the UPS database, and the Centaur BOM (Bill of Materials) was used as the new BOM to be evaluated. The process involves the model searching each entry in the Centaur BOM against the UPS database. It then auto-fills the results with the most similar component descriptions and their respective similarity scores, streamlining the process of identifying corresponding components. The results show that approximately 65% of the Centaur BOM entries have a similarity score of 70% or higher when matched with the UPS database. This indicates that the model effectively identifies relevant matches for a majority of the BOM components.

For electronics components specifically, the model achieved an accuracy of 81%. However, since electronics components contribute only a small portion to the overall weight of the BOM, their impact on the cumulative weight is minimal. As a result, the accuracy for the total cumulative weight of the BOM is exceptionally high, at 99%. This level of accuracy ensures compliance with the requirement that the error in weight must remain within a ±5% margin.

Moreover, the new approach has significantly reduced manual efforts involved in the process. Previously, manually searching through the data would take approximately 2 hours. With the implementation of this model, the time required has been reduced to just 5 minutes. With an annual projection of 200 LCAs, annual effort 400Hrs can be reduced to approx. 17Hrs. This drastic reduction in time demonstrates the efficiency and effectiveness of the automated system in handling LCA requests.