Accelerate Product Similarity Optimization

Version 1.0.0

September 2023

Accelerate Product Similarity

The Clustering Accelerator provides an easy way to create clusters in order to enrich data by creating data-driven labels on any dimensions of the transactions.

Overview (Product Similarity)

Purpose

It is usually difficult to handle a large number of products, yet at the same time defining a pricing strategy for the right subset of products is key, as it is not realistic to have a strategy for each and every products.

In this scenario Product Similarity Accelerator is a science brick that provides a **similarity score** between products and possibly **regroups** them into **similar groups**. These groups can then be leveraged to apply a pricing strategy.

In addition, Similarity groups help enrich data for further processes, such as *Clustering* or *Negotiation Guidance* and similar products can be offered as an alternative product in a quote.

Other typical use cases:

- For a Pricing Manager to understand relationships between products and to group them appropriately in order to steer pricing strategy at this level.
- For a Data Manager to match competitive products with their own portfolio.
- For a Spare Part Manager to match newly created parts within a meaningful product category.

Pricefx Solution

Product Similarity Accelerator walks you through the steps to easily compute a similarity score and regroup products by similarity based on product specifications. To do that, the model relies on 3 types of information used to define the products:

- **Text Attributes** - Any textual data, such as product name or product description, that describe the product. Several fields can be used (and will be combined). The resulting text will be encoded into a set of numbers by a "Transformer" which is the **T** of the famous ChatGP**T**. This step encapsulates the meaning and then compares texts, including synonyms. What is provided:
  - English text transformer
  - Multilingual text transformer, including Arabic, Chinese, Dutch, English, French, German, Italian, Korean, Polish, Portuguese, Russian, Spanish, Turkish.

- **Categorical Attributes** – Can be any kind of specification that defines the product, such as product category, brand, type, color...
- **Numerical Attributes** – Can be any kind of numerical specification, such as size, power... or even price (you can define a price threshold to avoid comparison of products that have prices too far apart).

From those attributes, a **similarity score** is computed, with the possibility to give a specific weight to each attribute type. Then **similarity groups** are created based on the relationships and similarity among all products.

Outputs

The outputs of the model are:

- List of products and similar product, including the similarity score between them.
- List of products with their similarity group.

A set of dashboards is also provided in order to review and assess the outputs.

Limitations

- **Number of products** – The current implementation can use up to 150 000 products at once.
- **15 languages supported** – Arabic, Chinese, Dutch, English, French, German, Italian, Korean, Polish, Portuguese, Russian, Spanish, Turkish.
  There is a possibility to use only English which increases the accuracy slightly. Several languages can be mixed and still be compared, possibly with a decrease of accuracy.
- **No out-of-the-box way to manually adjust outputs**, such as reassigning products to another product group. If needed, this should be configured separately e.g. in Custom Forms.
- **No predefined extension point** – There is no out-of-the-box extension point defined for now. If you intend to use your own metric, custom code should be written. (But then the accelerator becomes specific so it cannot be updated without extra effort to port those modifications.)
- **Data requirements** – See Data Requirements (Product Similarity).

Business User Reference (Product Similarity)

- Usage (Product Similarity)

Usage (Product Similarity)

Take the following steps to configure a model:

- Introduction
- 1. Create a Model Based on ProductSimilarity Model Class
- 2. Set the Scope of Products (Definition Step)
- 3. Set the Scope of Transactions (Definition Step)
- 4. Configure the Similarity Model (Definition Step)
- 5. Adjust Weights (Similarity Weighting Step)
- 6. Get Overview on Similarities (Product Similarity Step)
- 7. Explore Similarity per Product (Product Similarity Step)
- 8. Configure Similarity Grouping (Product Similarity Step)
- 9. Explore Similarity at Group level (Product Grouping Step)
- 10. Explore Products in Their Groups (Product Grouping Step)

- Best Practices

Introduction

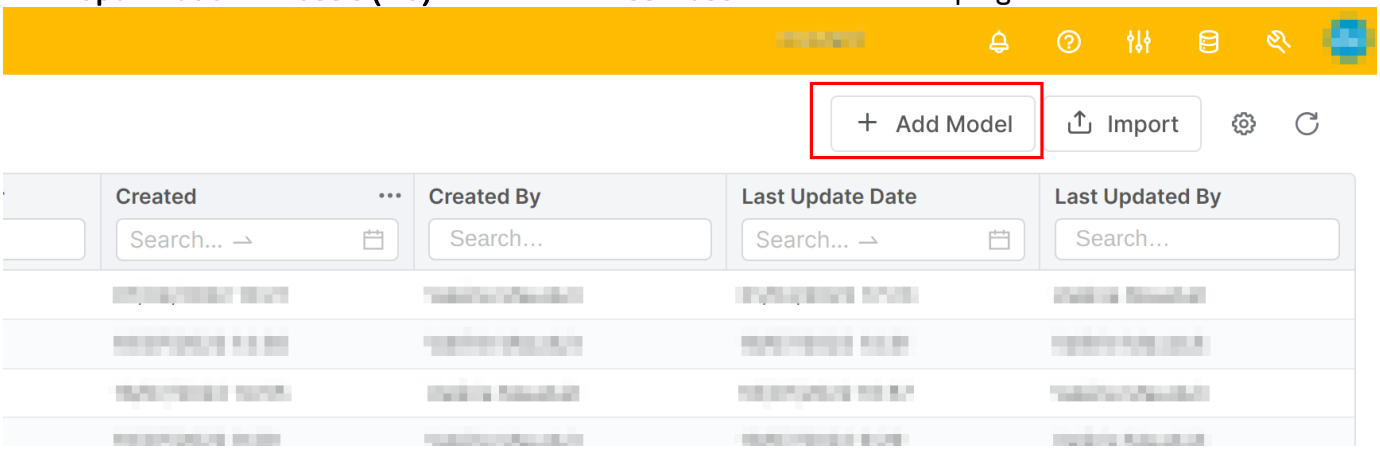
Product Similarity accelerator provides data on similarity between products (through a similarity score) and then groups products based on that score. The solution is powered by machine learning models. This documentation will guide you through setting up, configuring, and using this optimization model.

Main Features

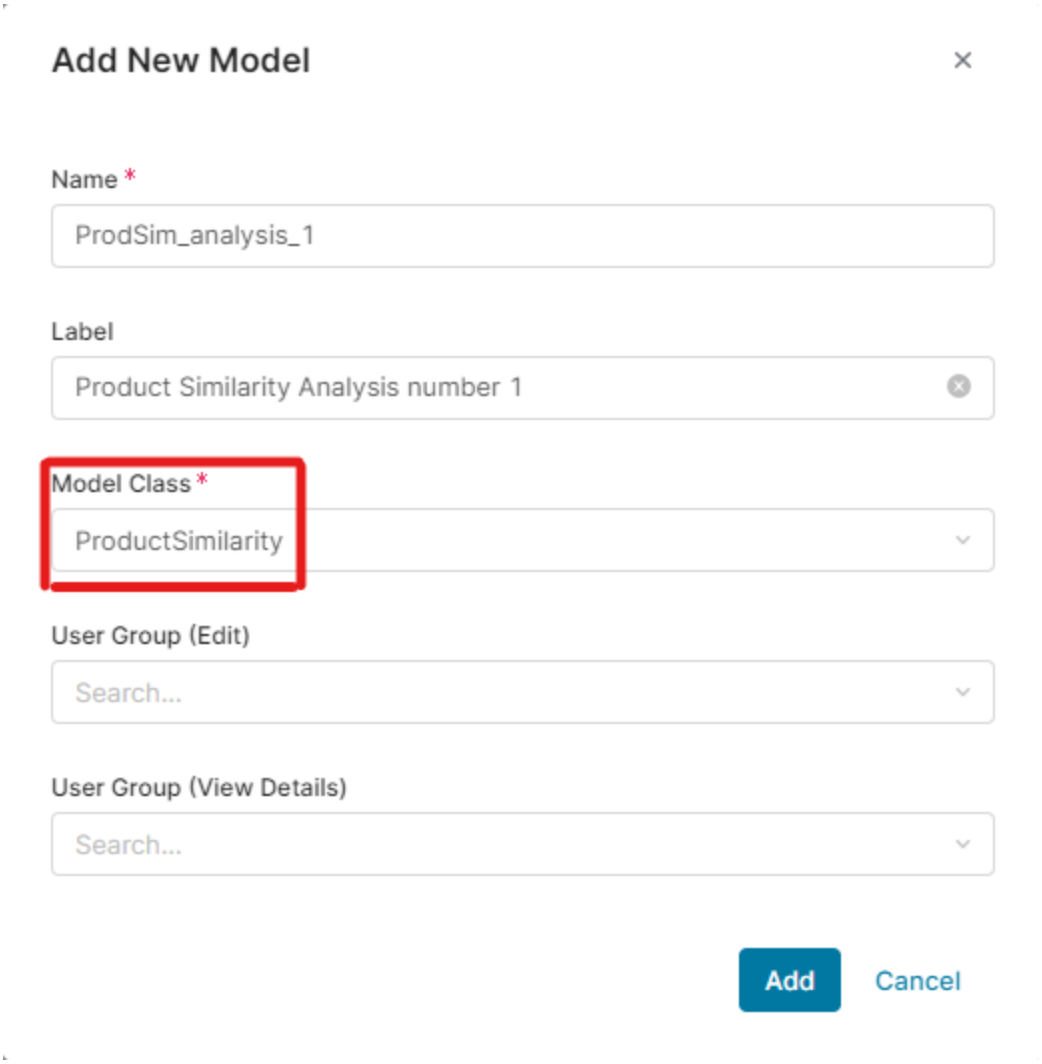
1. **Product Similarity Step** - Uses a combination of product attributes in order to compute the similarity between products. Each type of attribute (textual, categorical, numerical) provides a score that can be weighted in order to give more importance to some attributes.
2. **Product Grouping Step** - Builds a connected graph based on computed similarities and detects product groups.

1. Create a Model Based on ProductSimilarity Model Class

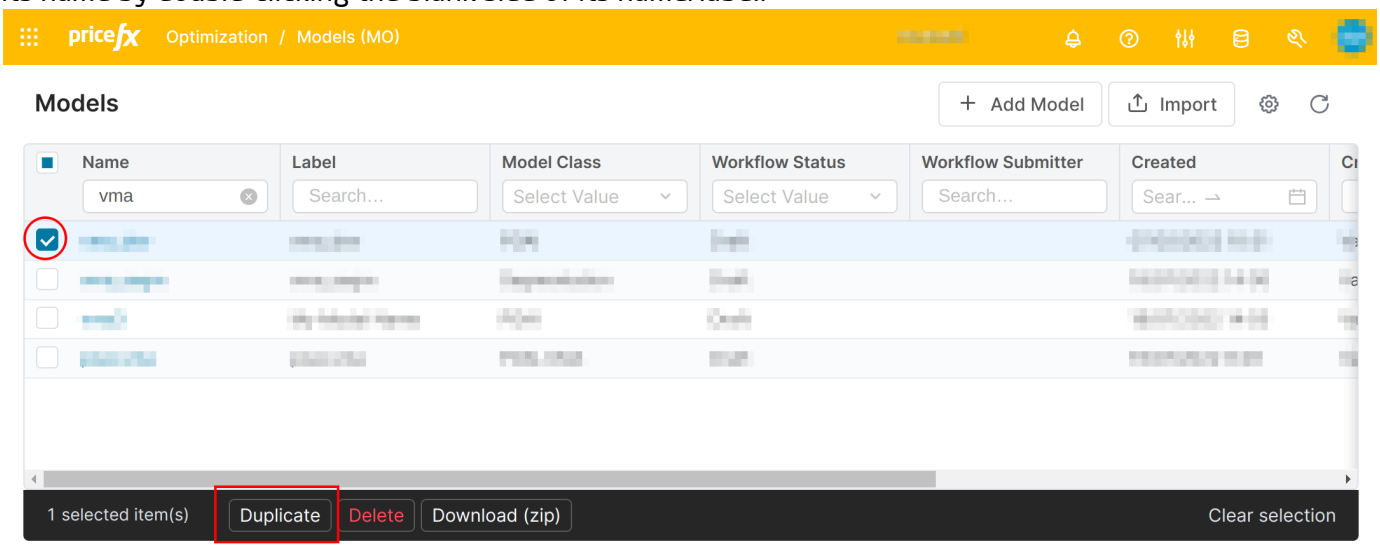
Go to **Optimization** > **Models (MO)** and click the **Add Model** button at the top right.



A pop-up is shown where you provide a name for your model, and the Model Class, which is *ProductSimilarity*, where you also have the possibility to configure which user groups can edit *ProductSimilarity* models and which user groups can view the details of computed *ProductSimilarity* models. Another Model Class would belong to another kind of optimization model.



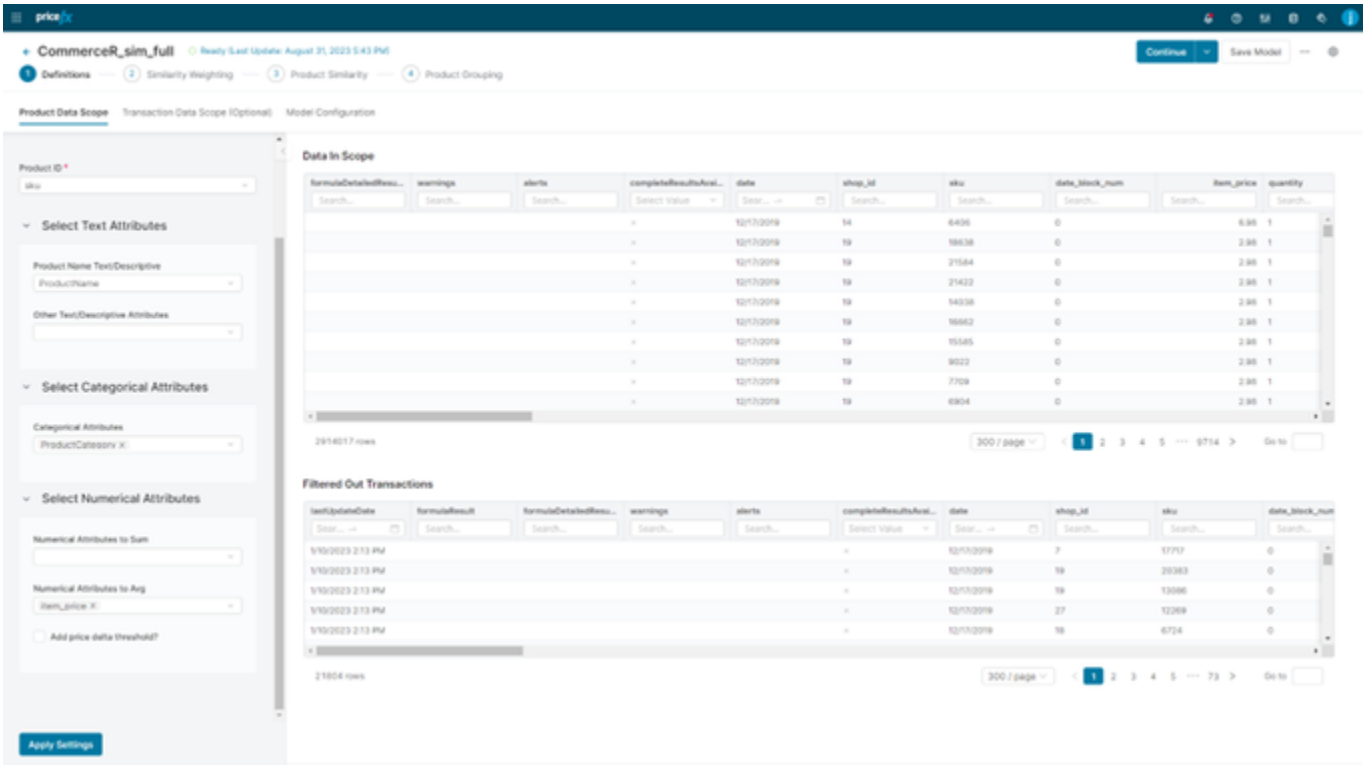
You can also duplicate an existing model. In this case, you will keep all the inputs of the previous model and you will have to rerun all the steps to get the outputs. Once you have copied a model, you can change its name by double-clicking the blank side of its name/label.



The same model class, *ProductSimilarity*, can be used by many models. Use informative names for your models, providing information on your dataset, and your calculation case.

The model you are configuring aims at identifying **Products** that appear similar in your business and regrouping them under data-driven labels, i.e. **similarity groups**. Similarity is estimated through product specific data, descriptions, hierarchical levels etc. and eventually transactional data. To process information related to similarity, two machine learning models will be implemented: the first one to manage textual descriptions, and the second one to regroup products in meaningful groups. You will have to set some parameters to make these algorithms works nicely on your data.

2. Set the Scope of Products (Definition Step)



This step is pivotal as it lays the foundation for your analysis. To ensure the accuracy and relevance of the similarity assessments, you need to define a clear scope for your products based on the following attributes.

Define the range of products you want to analyze:

1. Go to Product Data Scope tab under the Definition step.
2. Select your **Product table** with the correct name under **Data Source Input** (numerical, textual, or categorical).
3. Using filters to narrow down the scope of this analysis, you can easily exclude products based on every descriptor or combination of product's descriptors you want.
   - **Product ID** – This field is mandatory. Every product should have a unique identifier to differentiate it from others. This identifier can be a Product ID, SKU, or any other unique code.
   - **Text Attributes** – This is any textual information of the product providing context or description. Common attributes include product names, descriptions, or other labels.

     Even if a product name is not mandatory, it is highly recommended for the rest of this analysis.

     These columns will be processed using sentence transformers to compute textual similarity. Texts over 255 characters will not be considered.
   - **Categorical Attributes** – Categorical attributes are essential for classifying products into different categories or hierarchies. This data helps narrow down product pairs for similarity computations.

To set it up, go to the Categorical Attributes section and choose columns that represent categorical data, such as product category, brand, colors, types, etc.

- **Numerical Attributes** – These attributes represent a variety of metrics, such as sales volume, weight, or dimensions. Depending on their nature, they can be summed up or averaged during the analysis.
  To set it up, go to the Numerical Attributes subsection and for each numerical column, choose whether you want it to be summed up or averaged during analysis.
- **Price Delta Threshold** – Products that have significantly different prices would probably make no sense to be consider similar even if the product names are similar (like a spare part mentioning the original product name), so you can select only products within a defined price range.
  - In the Price Delta Threshold section, enter a value. Products with price differences exceeding this value will be excluded from similarity computations. Periodically review the set price delta threshold for its relevance.
  - Click **Apply Settings**.
  - Review and confirm the uploaded data, check the selected columns to ensure they match your intent.
  - Click **Continue** to proceed to the next step or go to the next tab to go further in the configuration.

Remember, clarity and correctness of this scope directly influence the subsequent steps and the accuracy of your results. Therefore, ensure you thoroughly understand your data and their attributes.

3. Set the Scope of Transactions (Definition Step)

Product Data Scope | **Transaction Data Scope (Optional)** | Model Configuration

Use Transaction Source Data?

Data Source Input *

Data [DS]

Data Source Filter

+ Set Filter

Product ID *

Product sku

Select Text Attributes

Product Name Text/Descriptive

Other Text/Descriptive Attributes

Select Categorical Attributes

Select Numerical Attributes

Incorporating transactional data can add depth and richness to the analysis, allowing the system to gauge product similarity not only on inherent product attributes. That is probably where you can get an average selling price to define a range of acceptable price. Here is a detailed guide on how to integrate and structure this data:

1. Go to Transactions Scope in the Definition tab.
2. Select the transactional Data Source [DS] with relevant data (e.g., quantity, revenue, margin).
3. Using filters to narrow down the range of transactions you want to consider for your analysis can improve the quality of results and speed up the processing. Applying filters (for example, the last two years) can help in focusing on recent trends and making the analysis more relevant to current market conditions.
4. Select the mandatory Product ID (unique identifier) among the columns from your transactional data that correlates with the product's unique identifier. It is crucial for consistency that this matches the unique identifier chosen in Step 2.
5. Select Textual, Categorical and Numerical attributes the same way it has been described above.
6. Click **Apply Settings**.
7. Review and confirm the transactional data.

Recommendations

- **Data Consistency** – Ensure that the unique product identifiers in the transactional data match exactly with those in the product data from Step 2. Any discrepancy can lead to missing or inaccurate insights.
- **Regularly Update Transactional Data** – The more recent your data, the more relevant and actionable the insights. Schedule periodic updates to keep the system's analysis current.
- **Consider Seasonal Variations** – If you are dealing with products that have seasonal variations in sales (e.g., winter clothes or summer accessories), consider this when choosing your date range (e.g. using last 12 months) and interpreting results.

4. Configure the Similarity Model (Definition Step)

Finetune the similarity model for best results:

1. Select the Model Configuration tab.
2. Select the right **Text Transformer** – English or Multilingual – for the content of the text columns selected in the Product Data Scope and Transaction Data Scope tabs. The text transformer is a crucial component when dealing with textual attributes. It is responsible for converting raw text

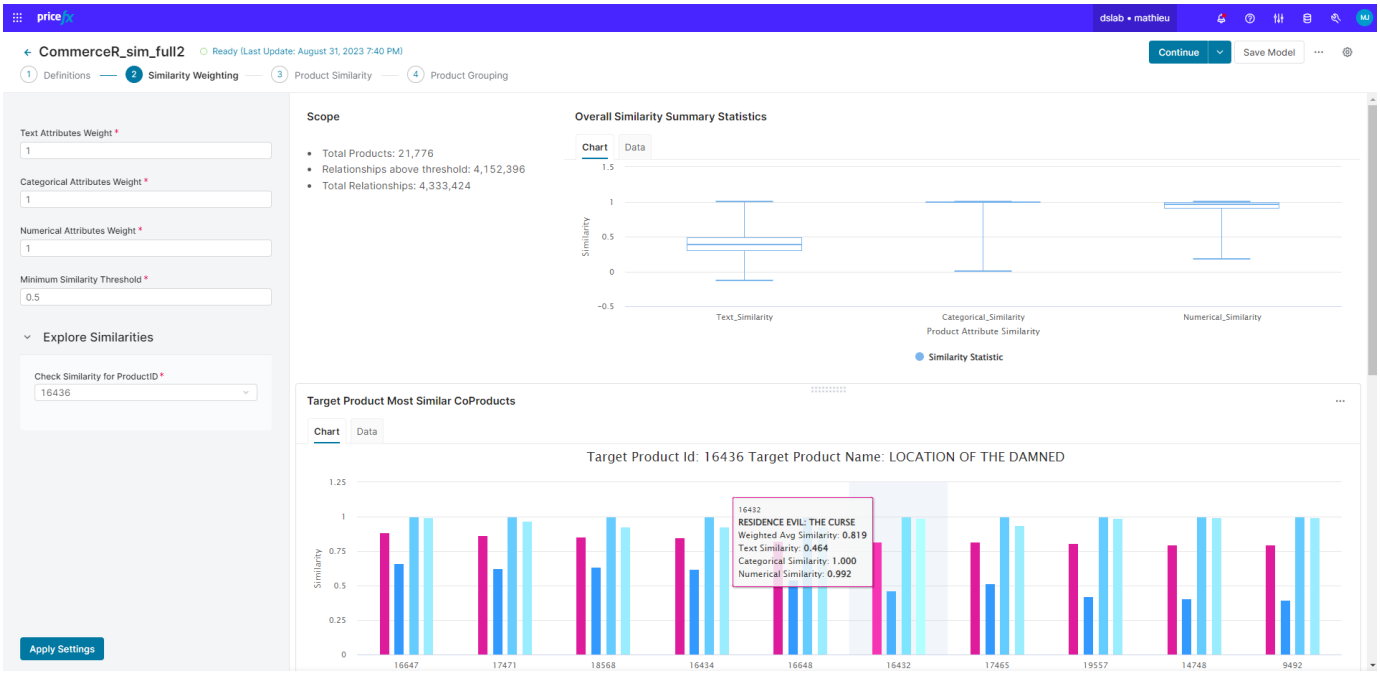
into numerical vectors that can be used for similarity computations. Depending on the language and scope of your product descriptions, you have two choices.
Multilingual supports: Arabic, Chinese, Dutch, English, French, German, Italian, Korean, Polish, Portuguese, Russian, Spanish, Turkish.

3. The **Refresh Text Transformer** option allows you to skip some computations which can be long, if encoding was already done sooner and with the same scope. If some parameters (like other categorical or numerical attributes) are modified, you want to re-execute the model. Nevertheless, be careful that the products are the same, if in doubt, it is recommended to re-execute the text transformers by selecting this check box.
4. Set **Maximum Number of Similar Products** that are kept in next steps of the analysis. This defines the maximum number of similar products that will be retained for each product in the dataset. For instance, if you enter '10', for every product, the system will keep data on its top 10 most similar counterparts based on the computed scores.
   Post transformation, the system will compute similarity scores between most of all possible pairs of products. However, for efficiency and clarity, you might want to limit the number of similar products that are kept for further, deeper analysis: a large number (more than 20) will make the computation longer and complexify the further grouping analysis; a too small number (below 5) will create a risk of producing less meaningful results.
5. Click **Continue**.

Recommendations

- **Know Your Text Data** – Before selecting a text transformer, ensure you are familiar with the languages present in your product descriptions. While the multilingual transformer is powerful, using it unnecessarily can be computationally intensive.
- **Start with a Moderate Number** – For Maximum Number of Similar Products, initially opt for a moderate number, 10 is a good start. Too few can miss out on relevant similarities, while too many can overwhelm the analysis. Once you are familiar with the system's outputs, you can adjust this number for subsequent runs.

5. Adjust Weights (Similarity Weighting Step)



In this step, the user has the opportunity to adjust the influence of each attribute type on the overall similarity score. The previous computation provides a similarity score up to 1 (perfectly matching) for each type of attribute:

- **Text Attributes** – Text Similarity based on cosine distance on the transformer encoded vector.
- **Categorical Attributes** – Categorical Similarity based on Hamming distance of the categorical values.
- **Numerical Attributes** – Numerical Similarity based on Mahalanobis distance of the numerical values.

A weighted approach allows for flexibility in how product similarities are determined based on the importance and relevance of each attribute type. Every product's attribute does not hold equal importance. Depending on the business context, some attributes may play a more significant role in determining product similarity than others.

Setting Attribute Weights

Default weights are displayed in the left menu for each type of attributes previously selected. Those weights can be adjusted and when you click **Apply Settings**, the dashboard on the right will reflect those parameters.

For instance, in most cases, text attributes provide more information and differentiate the products, so a higher weight is probably a good option.

For selected pairs of products, the system will calculate a Weighted Average Similarity (WAS) using the following formula:

$$\text{Weighted Average Similarity} = \frac{\text{wt} \times \text{St} + \text{wt} \times \text{Sc} + \text{wn} \times \text{Sn}}{\text{wt} + \text{wc} + \text{wn}}$$

Where:

- *wt* is the weight for textual attributes.
- *St* is the similarity score for textual attributes.
- *wc* is the weight for categorical attributes.
- *Sc* is the similarity score for categorical attributes.
- *wn* is the weight for numerical attributes.
- *Sn* is the similarity score for numerical attributes.

Setting the Similarity Threshold

To further refine the results,

1. Use the **Maximum Similarity Threshold** input field to set a value between 0 and 1.
2. Any product pairs with a weighted average similarity below this threshold will be ignored in subsequent analyses. This ensures you are focusing only on the most relevant and significant similarities.

Setting a too low threshold might end up with products that are considered similar but are not similar enough from a business point of view. Default value is 0.6 and we do not advice to have a threshold below 0.4.

Scope

Gives you insights on:

- Total number of products

- Total number of relationships
- Number of relationships above Maximum Similarity Threshold

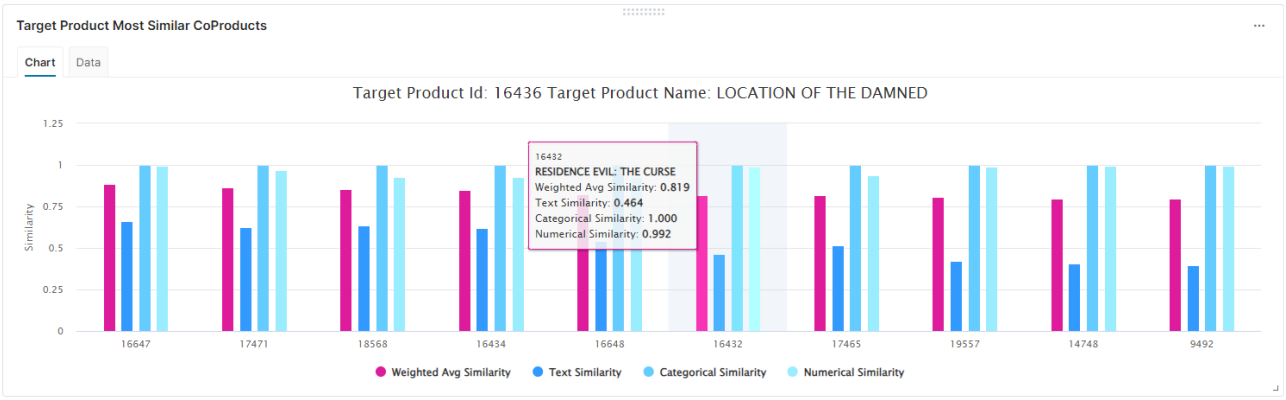
Visualize Similarity Distributions

Overall Similarity Summary Statistics can provide insights into how the three kinds of computed similarities are distributed across the dataset. These box plot charts show the distribution of values for textual, categorical, and numerical similarities. This visualization helps you understand the spread and central tendency of similarity scores, which can impact weight adjustments.

Explore Similar Products

Get a hands-on feel of the similarity results:

1. Navigate to the Explore Similarities section.
2. Select a product from the dropdown list. The system will display similar products based on the computed weighted average similarities.

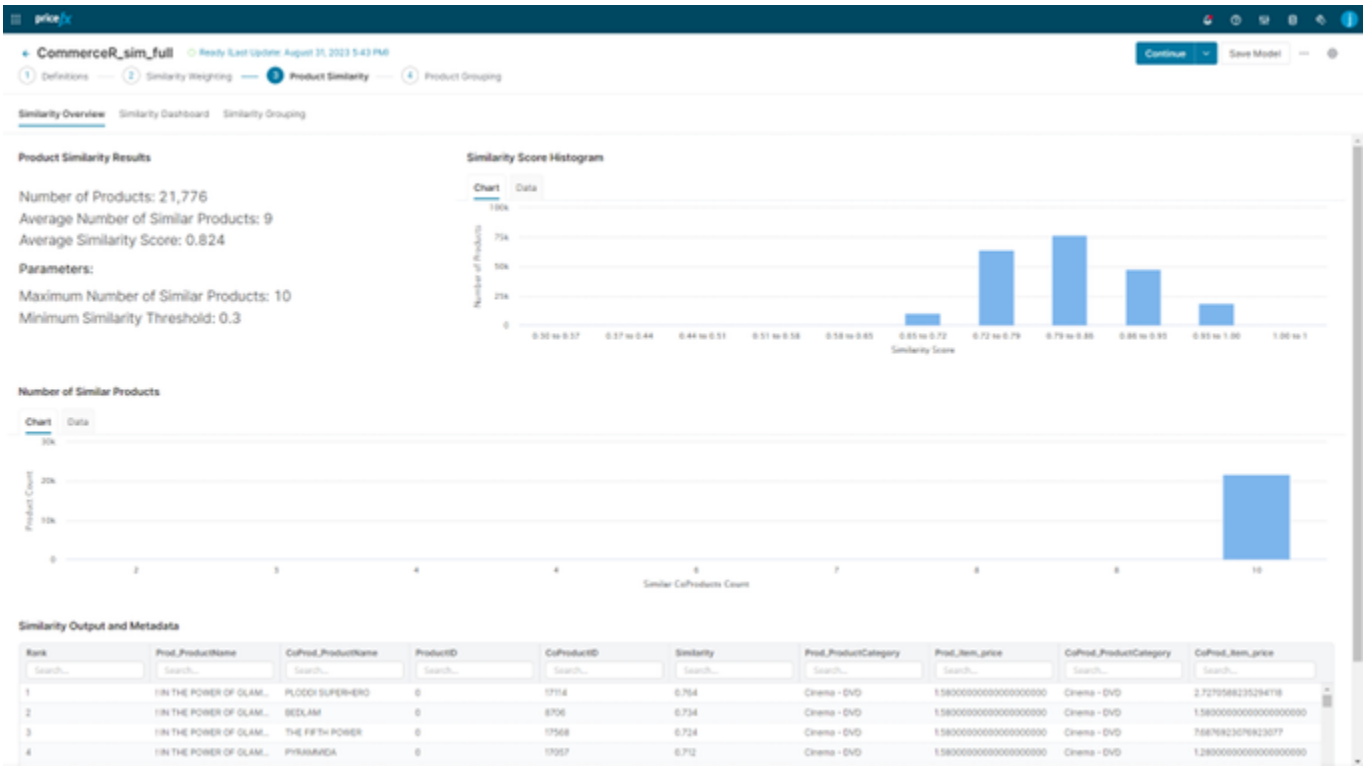


3. For a deeper dive, explore the data tables **Target Product Meta Data** and **Similar Co-Products Meta Data** below which properties of the selected product and its similar counterparts are shown. This allows for side-by-side comparison and validates the quality of the similarity computations.

Recommendations

- **Iterative Approach** - Start with equal weights and adjust based on the visual feedback from the box plots and product explorations. It is often helpful to iterate on weightings and thresholds multiple times to achieve optimal results.
- **Domain Knowledge** - If available, involve a domain expert. They can provide insights which attributes are more critical in determining product similarity in your specific business context.
- **Limit Extremes** - While it is possible, avoid setting any attribute weight to an absolute zero unless you are certain it holds no relevance. Even minor influences can sometimes provide valuable nuances in similarity computations.

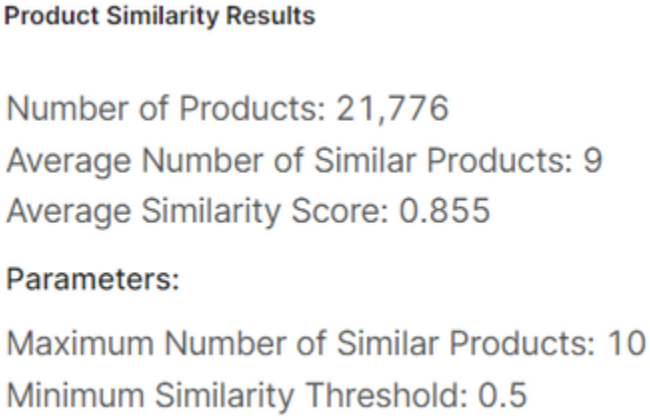
6. Get Overview on Similarities (Product Similarity Step)



This step presents the results of the similarity computations in a comprehensive dashboard designed to provide users with a holistic view of the results, visualization aids, and relevant metrics that help in assessing the efficacy of the similarity computations.

Product Similarity Results

This section presents high-level metrics which give a snapshot of the similarity results.



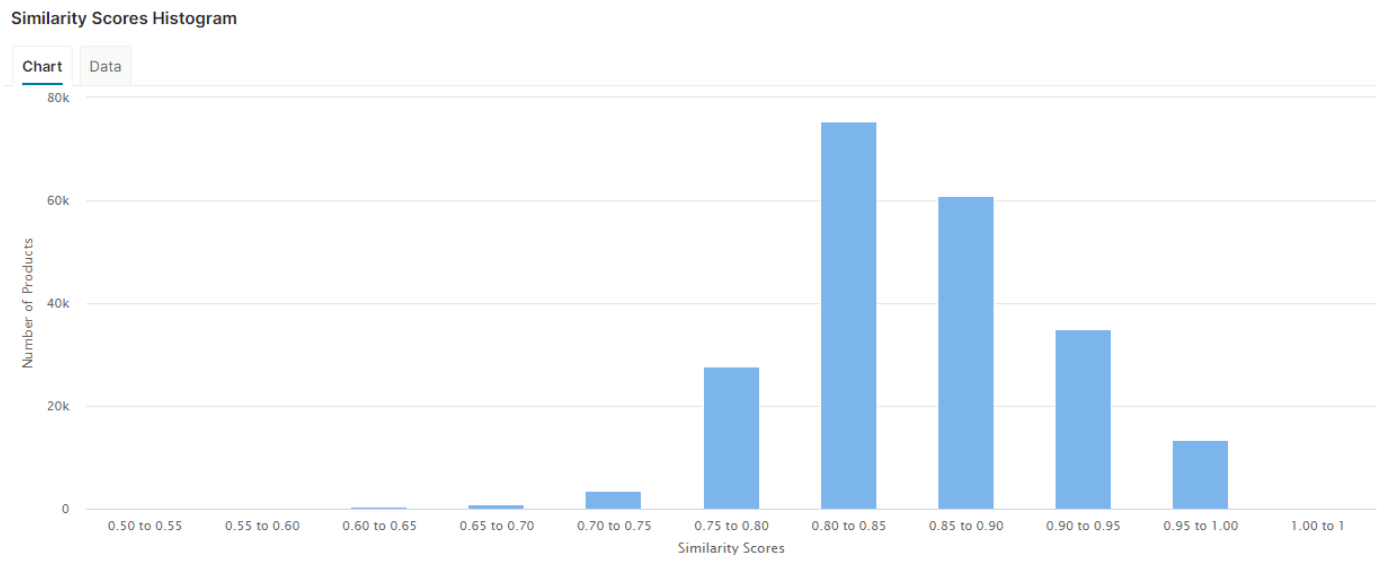
Product Similarity Results:

- **Number of Products** - Total count of unique products that were part of the similarity computation.
- **Average Number of Similar Products** - Across all products, this metric presents the mean count of products that were deemed similar based on the user-defined threshold.
- **Average Similarity Score** - Represents the average weighted similarity score across all product pairs that surpassed the threshold.

Parameters is a refresher section that reminds users about the parameters they set, ensuring transparency and easy revisits for iterations.

- **Maximum Number of Similar Products** - This denotes the upper limit on the count of similar products retained for each product post-similarity computation.
- **Minimum Similarity Threshold** - Indicates the cut-off similarity score below which product pairs were excluded from the results.

Similarity Scores Histogram



This bar chart provides insights into how the weighted average similarities are spread across product pairs.

Bar chart axes:

- **Horizontal Axis** - Shows ranges of similarity scores.
- **Vertical Axis** - Shows number of product pairs that fall within each similarity score range.

By looking at this chart, users can understand the concentration and distribution of similarity scores, which can be pivotal when deciding to revisit weights or thresholds.

Number of Similar Products

This histogram showcases how many products, on average, each product is similar to.

Histogram axes:

- **Horizontal Axis** - Shows bins representing count of similar products.
- **Vertical Axis** - Shows number of products that have a certain count of similar products.

It helps users visualize if most products have a lot of similar counterparts or if only a few products dominate the similarity landscape.

Similarity Output and Metadata

A detailed data table showcasing the pairs of products (a line is a relationship) that meet the similarity criteria.

Columns in the table:

- **ProductID** - Identifier for the first product in the pair.

- **CoProductID** – Identifier for the second product in the pair.
- **Similarity** – Computed weighted average similarity score (WAS) for this product pair.
- **Rank** – For a given ProductID, this is the rank of this relationship.
- ***Others*** – Available attributes for Product and CoProduct.

By exploring this table, users can delve deep into individual product similarities and get help in validation or further exploration.

7. Explore Similarity per Product (Product Similarity Step)

It is time to narrow down from the broad perspective of all products to the intricate details of a single chosen product. This dashboard is centered around understanding the similarity landscape for one product and observing its relationships in the dataset.

First, select a product by its Product ID from the dropdown menu on the left.

Available Charts and Summaries

Product Overview

Here you will get a detailed snapshot of the selected product's similarity status within the dataset.

- **Product ID** – Unique identifier for the chosen product.
- **Product Name** – The name or descriptor of the product.
- **Number of Similar Products** – The count of products that were deemed similar based on the user-defined threshold.
- **Average Similarity Score** – The mean similarity score of the chosen product against its similar products.
- **Max Similarity Score** – The highest similarity score achieved by the product against any of its counterparts.
- **Min Similarity Score** – The lowest similarity score of the chosen product against its similar products.

Parameters

This section is a quick reminder of the user-defined parameters that influenced the similarity results.

- **Maximum Number of Similar Products** – The upper limit set on the count of similar products retained for each product.
- **Minimum Similarity Threshold** – The user-defined cut-off similarity score.

Similarity Histogram

This visual representation provides a clear view of how the similarity scores for the selected product are distributed against their counterparts.

- **Horizontal Axis** - Shows ranges of similarity scores.
- **Vertical Axis** - Shows number of products (similar to the chosen product) that fall within each similarity score range.

Similarity Graph

This graph displays the network of relationships around the selected product.

- **Nodes**:
  - Dark Green Node - Represents the chosen product.
  - Light Green Nodes - Denote the products that are similar to the chosen product.
- **Edges** (or connections):
  - Connect the chosen product to each similar product.
  - Existing edges between green nodes represent connections between similar products, indicating they are also considered similar to each other.

This graph helps in understanding the local density of relationships, which can be helpful for certain analyses, such as network analysis and product grouping. The more one is intricated relatively to other graphs, the more the overall similarity in this graph is meaningful.

Product Data

This is a detailed table showcasing the various attributes and properties of the selected product. Depending on the dataset, this can include columns like:

- Product ID
- Product Name
- Category
- Price
- Description
- And more...

Similar Products Data

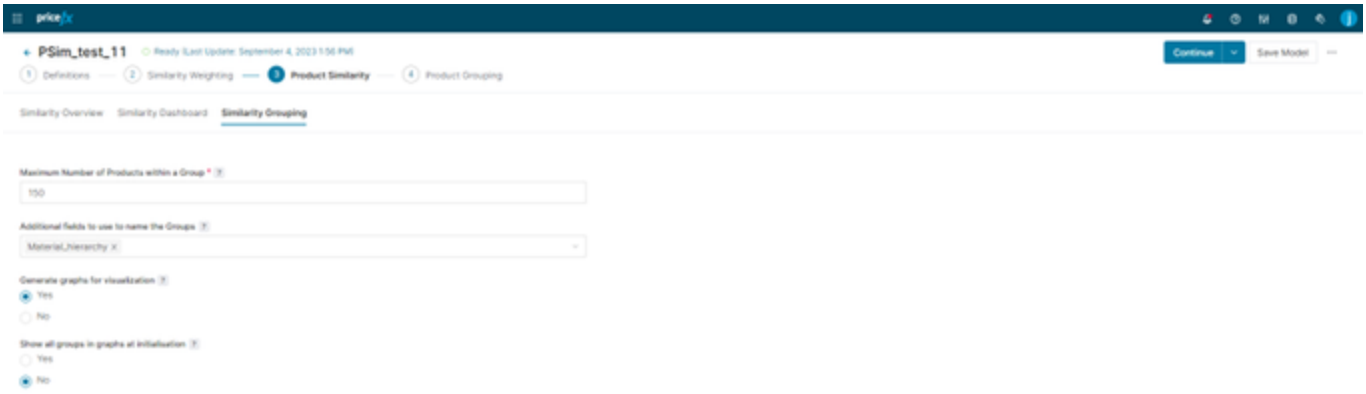
This is a tabulated view detailing the attributes and properties of all products that are similar to the chosen product. This table provides side-by-side comparison capabilities, making it easier to validate and analyze the similarity results.

Columns in the table (example):

- Product ID
- Product Name
- Category
- Price
- Similarity Score (against the chosen product)
- And more...

8. Configure Similarity Grouping (Product Similarity Step)

The last tab of the Product Similarity step is about transitioning from identifying similar products to creating tangible groups of these products. The configuration in this step ensures the resulting groups are meaningful, manageable, and aligned with business needs.



Maximum Number of Products within a Group

This parameter controls the size of each similarity group. Depending on the business use case, users might prefer smaller, tightly-knit groups or larger, broader clusters.

- **Default Value** is 150 products. This offers a balance between granularity and comprehensibility. The choice can be influenced by the business context or the total number of products in the dataset.

Additional Fields for Naming Groups

Group names are automatically generated based on frequency of words in text attribute defined under Product Name in the Definition step. You can further enrich that text with additional fields to make these names more meaningful, turning them typically into product categories. The dropdown menu contains text fields that were selected in step 1. You can select one or multiple fields from the dropdown menu. The values from these fields will be used to derive a name or label for each similarity group. For instance, if Category and Brand fields are chosen, a group might be named "Electronics - Samsung".

Please be careful to select fields that are unique per product, otherwise the naming will fail.

Generate Graphs for Visualization

Graphical representations can help in visualizing the formed groups and understanding the relationships within and between them. However, with extensive data, graph generation might be resource-intensive and time consuming.

- **Toggle Option (Yes/No)** - Users decide whether they want the system to generate visual graphs for the similarity groups.
  - **Yes** - The system will create and display graphical representations of groups.
  - **No** - The system will skip this visualization and computation time for this step will be lower, especially for large datasets.

Show All Groups in Graphs at Initialization

Once graphs are generated, users can decide whether they want to view all groups right at the start or explore them one by one. Showing all groups simultaneously can be overwhelming with big data, but it provides a comprehensive overview.

- **Toggle Option (Yes/No)**:

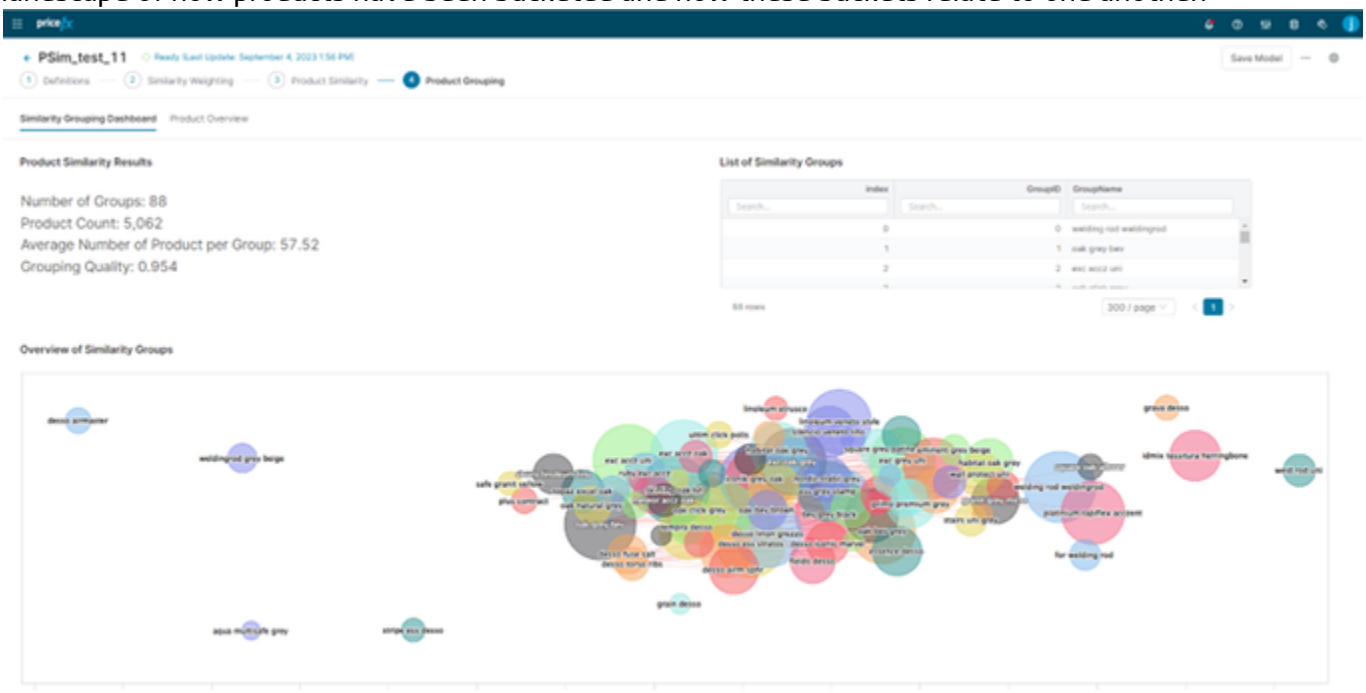
- **Yes** - All similarity groups will be displayed on the graph as soon as it is generated.
- **No** - The graph will be initialized in a condensed view, and users can choose to expand and explore specific groups as needed.

Once everything is set up, click **Continue** to trigger the grouping process.

9. Explore Similarity at Group level (Product Grouping Step)

This last step "Product Grouping" provides outputs of the previous steps and a dashboard to review product similarity groups.

The first dashboard takes you deeper into the heart of the grouped products, illustrating the macro landscape of how products have been bucketed and how these buckets relate to one another.



Product Similarity Results

This portlet provides a high-level summary of the results post-grouping.

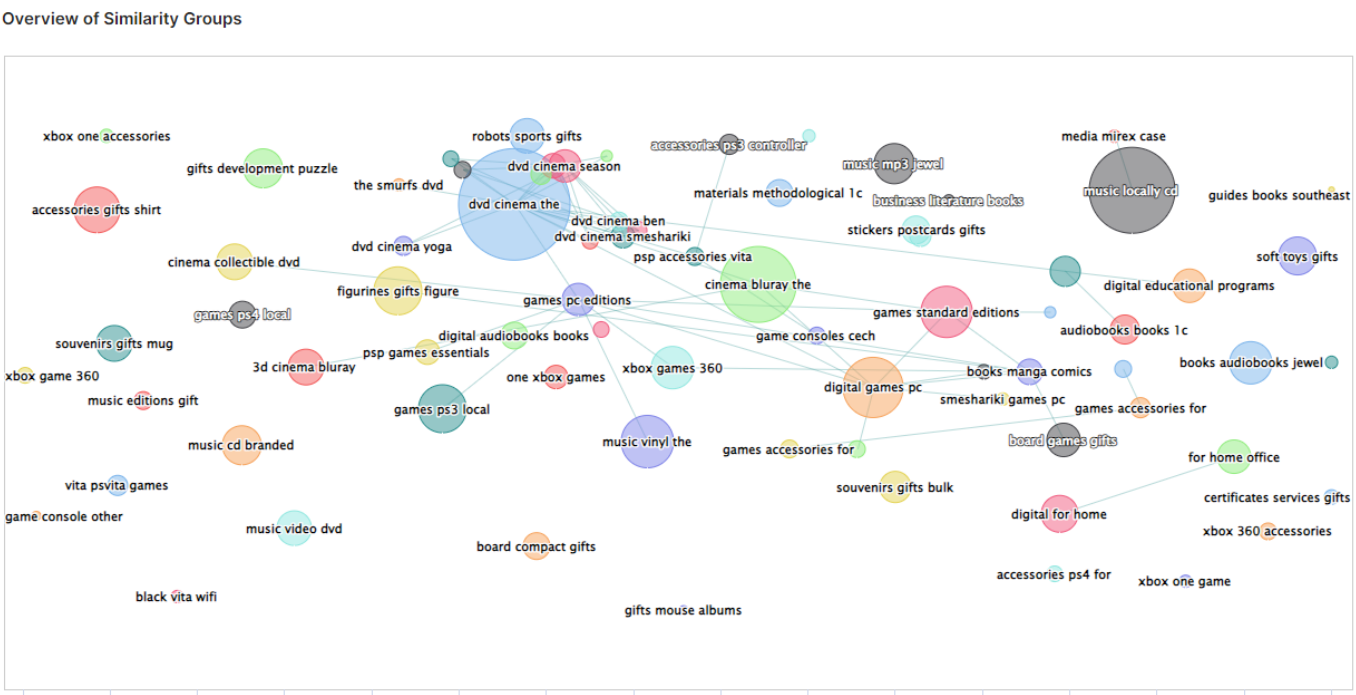
- **Number of Groups** - Total count of similarity groups created.
- **Product Count** - The total number of products that have been grouped.
- **Average Number of Product per Group** - Provides the average size of each group. It is calculated by dividing the total Product Count by Number of Groups.
- **Grouping Quality** - This value theoretically ranges from 0 to 1 and indicates the efficiency or accuracy of the grouping process (aka modularity). It is based on intra-group similarity (higher is better) and helps you in adjusting the settings. This can also be compared between models.

List of Similarity Groups

This is a tabular representation, scrollable if there are many groups.

- **GroupID** - Unique identifier assigned to each group. This ensures that even if groups have similar names, they can be differentiated.
- **GroupName** - The name or label of the group. This is formed based on the parameters set in Step 8, potentially utilizing additional fields for a descriptive name.

Overview of Similarity Groups



This is a graph representation that shows the interconnected landscape of product groups.

- **Nodes** - Each node represents a similarity group.
  - **Size** - Proportional to the number of products in the group.
  - **Label** - The name of the group printed on top of or below the node.
  - **Color Coding** - It is random.
- **Edges** - Connections between nodes (groups).
  - This shows that there are products in the two connected groups that are considered similar.
  - The thickness of the edge represents the number of such connections. Thicker lines mean more products between the two groups are similar.
- **Tooltip** - Hovering over a node provides additional details.
  - **Cardinality** - Indicates the number of products in the group.
  - **GroupID** - Unique identifier of the group.

10. Explore Products in Their Groups (Product Grouping Step)

This "similarity journey" finishes with a colorful firework style graph. It allows you to deep dive into the intricate web of product relationships and how these relationships translate into groups.

Group Selection

On the left side in Display Products you will find:

- **Similarity Groups** – Contains a list of all group names. Users can multi-select the groups they are interested in.
- **Apply Settings** – After making selection, users need to click this button. The dashboard will then refresh, displaying only products from the chosen groups.

Available Widgets

Product Similarity Results

This widget provides an overview and sets the stage for the exploration that will follow.

- **Displayed Products** – The total number of products that are currently displayed on the graph, based on the user's group selection.
- **Displayed Groups** – The number of unique groups that the displayed products belong to.
- **Unique Groups Names** – The number of unique names of the groups currently displayed. This may differ from **Displayed Groups** if at least two groups received the same name (that can happen if used textual product descriptors have a weak vocabulary).
- **Total Number of Groups** – This remains a constant, representing the total groups formed, helping to contextualize the current view in relation to the entire dataset.
- **Total Number of Products** – Another constant, representing the total products in the dataset. This gives the user a sense of scale and perspective.

Product Details

An essential table that provides the user with a detailed breakdown of individual products and their affiliations.

- **ProductID** – Unique identifier for each product.
- **ProductName** – If set, this provides the name or label of the product, helping with immediate recognition.
- **Coordinates in the Graph** – X and Y coordinates indicating the product's location in the visual graph. This is particularly useful if users wish to correlate table entries with their visual representations. Coordinates are precomputed, so more data points can be displayed.
- **GroupID** – The unique identifier of the group the product has been assigned to.
- **GroupName** – The name or label of the associated group.

For users' convenience, features like searching, sorting, and filtering can be integrated within this table.

Products and Groups Graph

The main representation of products and their connections.

- **Nodes** – Each node represents an individual product.
  - **Color Coding** – The color of a node indicates the group it belongs to. This helps in visually clustering similar products.
- **Edges** – Thin lines connecting products, representing the similarity between two products.
- **Tooltip** – Hovering over a product node gives a concise overview.
  - **ProductName** – The name of the product.
  - **ProductID** – Its unique identifier.
  - **GroupName** – The name of the group it belongs to.
  - **GroupID** – Unique identifier of the group.

Zooming, panning, and selecting multiple nodes will provide a deeper information about product and groups interconnections.

⚠ This chart requires a lot of resources to be displayed and may fail when too many products are displayed, especially if there are more than 20 000 products to display.

After completing step 10, you will have a comprehensive view of the product landscape, understanding both group-level and individual product-level relationships. This holistic insight can drive better decision-making, foster innovation, and unearth hidden patterns.

Best Practices

- **Regularly update data** – For best results, ensure your product and transaction tables are updated frequently.
- **Monitor weights** – Over time, the importance of certain similarity metrics might change. Periodically review and adjust these.
- **Feedback loop** – After grouping, review product placements in groups. If mismatches are found, adjust model settings and re-run.
- **Regular data audits** – Given the complexity and the multifaceted nature of the data being analyzed, it is recommended to periodically audit your data for consistency and correctness. For instance:
  - Ensure the unique identifiers remain unique and consistent across updates.
  - Validate that text attributes still make sense and remain relevant.
  - For categorical attributes, ensure that the categories are still valid and that no new ones need to be added.
  - Re-assess the validity and relevance of numerical attributes.
  - Periodically review the set price delta threshold for its relevance.

- **Data Quality and Cleansing** – Before starting with the product similarity analysis, always ensure that your data is clean. This includes:
  - Removing duplicates.
  - Handling missing values.
  - Correcting any inconsistencies or inaccuracies in the data.
    High-quality data will lead to more accurate and reliable similarity results.

Troubleshooting

- **Missing data** – Ensure all uploaded tables have complete data. Missing values can impact the quality of results.
- **Performance issues** – For large datasets, the analysis might take longer. Consider filtering the scope or optimizing your data for faster results.
- **Incorrect grouping** – If groups do not appear accurate, consider adjusting the similarity model settings or the grouping threshold.

Conclusion

Product Similarity is a comprehensive tool designed to streamline the process of identifying and grouping similar products in your business. By following this guide and understanding each step, you can maximize the benefits of this platform for your business needs.

Admin User Reference (Product Similarity)

Data Requirements (Product Similarity)

When using Product Similarity Accelerator, it is crucial to understand the type of data needed to make the most of the module's capabilities. Both mandatory and optional fields play an essential role in producing accurate and meaningful results.

Data can be loaded from either a Data Source or a Datamart.

⚠ Product Master, Product Extensions or Company Parameter tables are *not* supported; all product information should be loaded in a Data Source.

1. Mandatory Fields

These fields are necessary for the basic functioning of the system. Without them, Product Similarity cannot perform its core operations.

For Product Data

1. **Product ID** – A unique identifier for each product. It ensures that each product is treated as a distinct entity.
2. Some attributes, at least one of these 3 types:
   a. **Textual Attributes**
   b. **Categorical Attributes**

c. **Numerical Attributes**

For Transactional Data (Optional)

Transactional data is optional, however if selected, the following fields are required:

1. **Product Identifier within Transactional Data** - This should match the `ProductID` from the product data, so both tables can be joined.
2. Some attributes, at least one of these 3 types:
   a. **Textual Attributes**
   b. **Categorical Attributes**
   c. **Numerical Attributes**

2. Optional Fields

These fields, while not strictly mandatory, significantly enhance the system's outputs, offering richer insights to define product similarities. Most of the business value of this Product Similarity accelerator comes from the right product attributes for your business, which could be really diverse depending on the industry, so take a moment and list what makes a product specific for your own business. Here are some examples:

1. **Brand** - The manufacturer or brand name associated with the product.
2. **Size/Dimensions** - The physical size, measurements, or dimensions of the product.
3. **Packaging** - Type of packaging or any information about how the product is packaged for shipping and storage, such as bottles, bags, boxes... of different sizes.
4. **Weight** - The weight of the product.
5. **Color** - The color for the product..
6. **Material** - The materials used to make the product, especially important for clothing, furniture, and electronics.
7. **Power** or any other value that demonstrates the capabilities of the products, e.g. cooling capacities for AC.
8. **Specific Features** - Capabilities of the product, such as a camera's megapixels or a smartphone's operating system.
9. **Specifications** - Detailed technical specifications, such as processor speed, storage capacity, resolution, etc., depending on the product type.
10. **Power Source** - For electronics or appliances, information about the power source or energy requirements.
11. **Certifications** - Any industry-specific certifications or compliance with safety standards.
12. **Country of Origin** - Where the product is manufactured or produced.

For Product Data

1. **Product Name** - Highly recommended, a textual name or label of the product. It aids user recognition and can be utilized in text-based similarity computations. A product description can also be an alternative. ⚠️ The length is limited to 255 characters.
2. **Textual Attributes** - These can be additional textual attributes providing more context about the product, such as a long description (again, with the limitation of 255 characters).
3. **Categorical Attributes** - Fields such as hierarchical descriptors or product categories can provide valuable context for grouping products.
4. **Numerical Attributes** - Numerical specifications, such as size, power or unit price. For aggregation, data that can be either summed up or averaged.

For Transactional Data

1. **Criteria for Filters** – Users might want to limit the scope of analysis to specific time frames, e.g., the last two years.
2. **Text Attributes** – Can include product names or other descriptions that provide more context for each transaction, with the limitation of 255 characters.
3. **Categorical Attributes** – Similar to product data, this could be hierarchical descriptors or transaction categories.
4. **Numerical Attributes** – Data that can provide context for each transaction, such as transaction amount, quantity, or any other relevant metric.

By fulfilling the mandatory data requirements and supplementing them with the optional fields, you can maximize the value of Product Similarity Accelerator. It is advisable to provide as many relevant fields as possible to ensure nuanced, accurate, and comprehensive results.

Language Support

For text attributes, only the following **15 languages are supported**: Arabic, Chinese, Dutch, English, French, German, Italian, Korean, Polish, Portuguese, Russian, Spanish, Turkish.

Installation (Product Similarity)

The Clustering Accelerator deploys the Clustering Model Class and the related logics.

Prerequisites

Before you start the installation of the Accelerator, ensure you have the according Product Data Source and potentially a Transactions Datamart or Data Source.

For some important points about it, see Data Requirements (Product Similarity).

Deployment

1. Access PlatformManager at https://platform.pricefx.com/ and log in with your account or using O365.
2. Go to **Marketplace** > **Accelerator Packages**.
3. Find **Optimization - Product Similarity**.
4. Select your Target Partition from the drop-down menu.
5. Set up Datamart mapping.
   - Some rules apply in the mapping:
     - The numerical values must be extended.
     - The percentage values must be values between 0 and 1 (i.e. a margin rate is defined by margin/revenue).
6. Click **Continue** and wait until the deployment is complete, that is:
   - *ProductSimilarity* model class is deployed.
   - All 22 required logics and libraries are deployed and added to the partition.
   - *PythonEngine* is configured to operate on the partition.

Technical User Reference (Product Similarity)

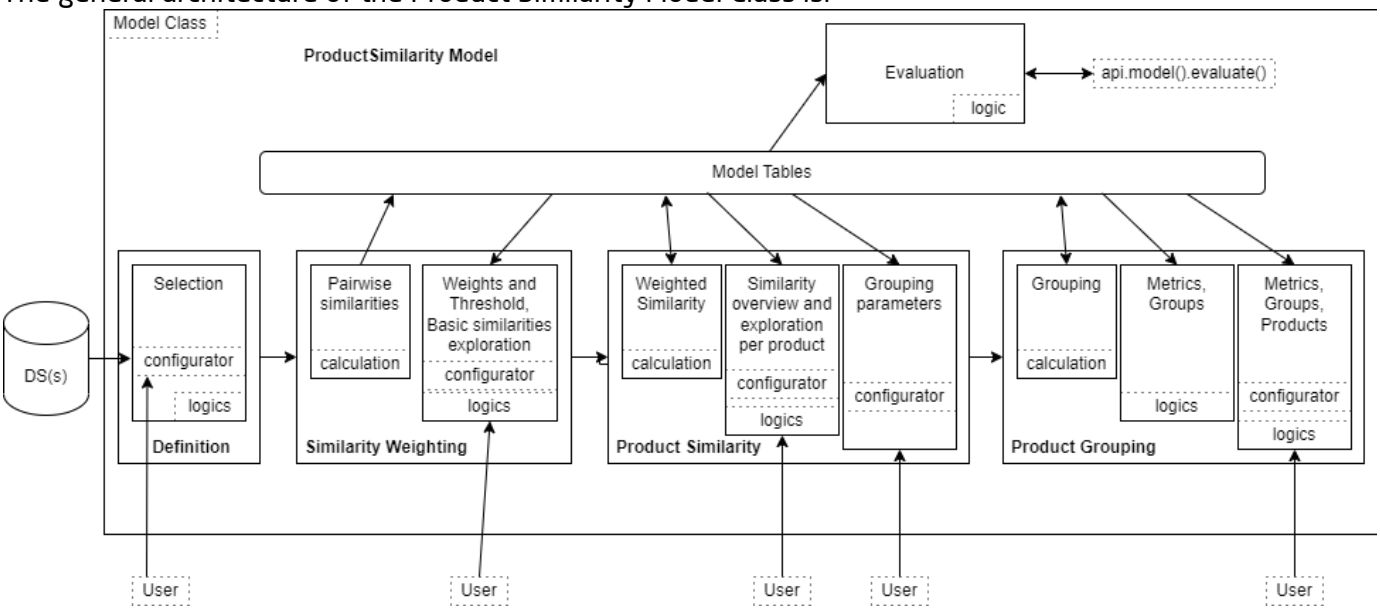
This section details the ModelClass and logics that the Product Similarity Accelerator deploys. For each step, its aim, outputs, and the main reasons to modify the logics are explained.

In this section:

Product Similarity Model Class

The Product Similarity Model Class organizes a list of logics to create the model architecture. It is a JSON file that refers to some logics and it is transformed into an optimized UI in the Pricefx platform.

The general architecture of the Product Similarity Model Class is:



It defines four steps:

- **Definition** – Sets the scope of the products tables and of transactions, and sets parameters for similarity exploration model.
- **Similarity Weighting** – Runs similarity measures and screens for more similar products, then lets the user set the weights and threshold for finest comparison of products.
- **Product Similarity** – Looks at the outputs of the similarity analysis, the similar products of any product and finally lets the user configure the grouping.
- **Product Grouping** – Looks at the groups and products in the groups.

Library

The logic is **ProdSimMC_Lib**.

- Aim of the logic

  ProdSimMC_Lib is used in nearly all the other logics deployed by the accelerator and defines a set of functions needed specifically for this accelerator, but also some constants used to easily change the user interface wording. There are the following elements:

  - `Parameters` - Contains a function to check the type of the columns when exporting some tables.
  - `Utils` - Constants definition.
  - `Labels`- Groups static fields used for naming variables, tables...
  - `Definitions` - Sets of tools dedicated to the Definition step.
  - `DataDefinitions` - Sets of tools dedicated to interaction with data and user settings.
  - `Configurators` - Groups the methods to deal with formatting user inputs.
  - `ConfigurationUtils` - Groups the methods to deal with initialization of user inputs.

  It is accessed via the calls on `libs.ProdSimMC_Lib.XXX` in the code.

- Common reasons to modify the logic

  If there is another kind of input to deal with.

Definition Step

There is no calculation logic in this step, and there are three tabs with related dashboard and evaluation logics: **ProdSimMC_Prod_Data_Def** and **ProdSimMC_Trans_Data_Def** and **ProdSimMC_Model_Def_Eval**.

- Aim of the logics

  These logics provide the user inputs to define at least a source of product data to map it, and optionally to define a source of transactional data (plus mapping) and to define what kind of text transformer to use, as well as the maximum number of similar products to keep in the following analysis.

- Outputs of the evaluation

  A table of the filtered product data and optionally, a table of the filtered transactional data that will be used for the similarity analysis.

- Common reasons to modify the logics

  - Some other mappings are needed or would be retrieved.
  - Some customized metrics that will require specific developments.
  - To define pre-set filters.
  - To add a chart to better understand the data. (Caution: it can take long, as the data are not yet stored in the model.)

Similarity Weighting Step

Contains one calculation sequence that chains 4 logics **ProdSimMC_LoadData_Calc**, **ProdSimMC_TextTransformers_Calc**, **ProdSimMC_ApproxNearestNeighbors_Calc**, and **ProdSimMC_CoProductMetaData_Calc** that are executed when accessing this step. The dashboard is split in two panels, one for user inputs, the other for evaluation.

Calculation: Data Aggregation

The logic is **ProdSimMC_LoadData_Calc**.

- Aim of the logic
  Aggregates the selected products and transactions data at the levels of ProductID defined by the user in the Definition step.
- Outputs of the calculation
  - Model tables of the aggregated data.
- Common reasons to modify the logic
  - To add a chart to better understand the data.

Calculation: Text Transformation

The logic is **ProdSimMC_TextTransformers_Calc**.

- Aim of the logic
  Transform the set of textual descriptors of each product in a vector of numerical values.
- Outputs of the calculation
  - Model tables of the transformed data.
- Common reasons to modify the logic
  - To add a chart to better understand the data.
  - To combine different transformers.

Calculation: Raw Similarity

The logic is **ProdSimMC_ApproxNearestNeighbors_Calc**.

- Aim of the logic
  Selects among all possible pairs of products a subset of pairs to compare. The comparison is based on the three kind of products' attributes Textual, Numerical, Categorical.
- Outputs of the calculation
  - Model tables of unweighted distances between products.
- Common reasons to modify the logic
  - To subset the distances in more atomic ones based on a better/specific business knowledge.

Calculation: CoProductMetaData

The logic is **ProdSimMC_CoProductMetaData_Calc**.

- Aim of the logic
  Provides for each product a subset of the best similar candidates.
- Outputs of the calculation
  - Model tables of the product pairs.

Setup Panel

The logic is **ProdSimMC_SimWeight_Def_Eval** and uses **ProdSimMC_SimWeight_Def_Configurator**.

- Aim of the logic
  It collects user inputs for:

- Weighting Textual Similarity
- Weighting Categorical Similarity
- Weighting Numerical Similarity
- Setting the similarity threshold to keep (or not) a pair of products
- Selecting one product at a time to visually evaluate its similarities

Outputs of the evaluation
User inputs.

Common reasons to modify the logic
- To differentiate threshold per similarity

EvaluationPanel

The logic is **ProdSimMC_SimWeight_Def_Eval**.

Aim of the logic
Exposes to the user some information about the final scope of the product similarity analysis:

- Products relationships data that can help in defining settings in the left panel
- Statistics about distribution of similarity measures
- Composition of the similarity based relationship of the select product
- Data about the selected product
- Data about products that have been considered similar to the selected product

Outputs of the evaluation
Overview before/after the threshold setting.

Displaying Box plot diagrams.

Displaing similarities for most similar products.

Common reasons to modify the logic
There could be different parameters to customize the scope of the product similarity analysis and their impact should be exposed here.

Product Similarity Step

Starts with one calculation logic **ProdSimMC_WavgSimilarity_Calc** that is executed when accessing this step which splits in three tabs: Similarity Overview, Similarity Dashboard, and Similarity Grouping. First, this calculation subsets the products' pairs that fulfill the minimum similarity criterion and saves them in a model table. Then, some other model tables are prepared to have the data ready for display in the dashboard's histograms and tables.

Similarity Overview

The logics is **ProdSimMC_OverviewSim_DashBoard**.

Aim of the logics
This dashboard provides insights about the computed similarities, based on the thresholds in some meaningful portlets. An overview gives macroscopic values: the number of products, average number of similar products, average score. Then some histograms give the user an idea of the distribution of the similarity score and of the number of neighbors a product may have.

Outputs of the evaluation
As for any other evaluation logic, there is no real output, it displays a dashboard.

- Common reasons to modify the logics
  To display other charts or to provide meaningful information in a different way.

Similarity Dashboard

The logic is **ProdSimMC_Similarity_DashBoard** which proposes an interactive dashboard for exploration of one product's similarities.

- Aim of the logic
  In the configuration panel on the left, the user selects a product by its **ProductID**. All available information about this product and how it is connected with its neighbors are displayed on the right.
- Outputs of the evaluation
  As with any other evaluation logic, there is no real output, it displays a dashboard.
- Common reasons to modify the logic
  To display other charts or to provide meaningful information in a different way.

Similarity Grouping

The logic is **ProdSimMC_Community_Parameters**.

- Aim of the logic
  This configurator lets the user set grouping parameters for the next step, including some constraints on the size of groups, the way groups are named and finally triggers (or not) some graph display that can cause trouble in case of oversized datasets.
- Outputs of the evaluation
  Internal parameters saved for next step.
- Common reasons to modify the logic
  If there is a new parameter for the grouping that the user needs to set.

Product Grouping Step

Contains one calculation logic named **ProdSimMC_WavgCommunity_Calc**.

- Aim of the logic
  It turns the similarity model table into a network of products upon which it is possible to make some clustering based on network topology. After this grouping process a naming process tries to give each group a meaningful name using contained product's name and selected descriptors.
- Outputs of the evaluation
  A set of model tables used by the dashboards.
- Common reasons to modify the logics
  To export other tables and thus provide more information in the dashboards.

Similarity Grouping Dashboard

The logic is **ProdSimMC_Communities_DashBoard**.

- Aim of the logic
  This dashboard provides the macroscopic view of Similarity Groups, their name and a summary of their content.
- Outputs of the evaluation
  As in any other evaluation logic, there is no real output, it displays a dashboard.

- Common reasons to modify the logic
  To display other charts or to provide meaningful information in a different way.

Product Overview

The logic is **ProdSimMC_ProductCom_DashBoard**.

- Aim of the logic
  This interactive dashboard provides a view on the content of the similarity groups displaying direct links between products. The user selects the name of the similarity groups they want to display.
- Outputs of the evaluation
  As in any other evaluation logic, there is no real output, it displays a dashboard.
- Common reasons to modify the logic
  To display other charts or to provide meaningful information in a different way.

Evaluations

The model has one evaluation: **ProdSimMC_ModelEvaluation_Eval**. That allows you to retrieve for one product or a list of products all the raw similarities that have been computed for it/them. For more details about model evaluations see https://pricefx.atlassian.net/wiki/spaces/KB/pages/3851026646/Query+Optimization+Engine+Results#Using-the-Evaluator.

Product Similarity Metrics (Product Similarity)

Product Similarity Accelerator leverages advanced metrics to identify similarities between products. Understanding these metrics will bring you a clear view on how product similarities are computed.

1. Sentence Transformers & Cosine Similarity (Textual Attributes)

Sentence Transformers

Concept: Sentence transformers map textual attributes (like product names or descriptions) into a dense vector representations. This captures the semantic essence of sentences or texts as the transformer has been previously trained to map such meaning. That way synonyms will be represented by similar vectors, which will provide a high similarity score between two similar products, such as Battery and Accumulator.

Use: To translate text into vectors and turn comparisons into mathematical operations. If several fields are selected, they are concatenated.

Cosine Similarity

Concept: It measures the cosine of the angle between two non-zero vectors.

Cosine Similarity(A,B)=AB / AB
where A and B are vectors.

Use: After transforming textual descriptors into vectors, cosine similarity is a kind of distance between two product descriptions, providing the similarity for textual attributes.

2. Hamming Similarity (Categorical Attributes)

Hamming Similarity

Concept: This metric quantifies the difference between two products defined by a set of categories, by checking the number of differences between the sets of values. If there are 4 categorical fields and 3 are similar between the two products, then the similarity will be 3/4 = 0.75.

Use: For categorical attributes, the Hamming similarity offers a way to determine the similarity of two product categories or hierarchies.

3. Mahalanobis Similarity (Numerical Attributes)

Mahalanobis Similarity

Concept: This metric provides the distance between numerical values after standardizing values for mean and standard deviation of each attributes, in order to compare numerical values on the same scale.

$DM(x) = \sqrt{((x-\mu)^T S^{-1}(x-\mu))}$
where x is a vector, μ is the mean of the distribution, and $S^{-1}(x-\mu)$ is the inverse of the covariance matrix.

Use: It is applied to numerical attributes, ensuring the same scale for each attribute to come up with the similarity for numerical attributes.

5. Weighted Average Similarity

Concept: In order to combine the textual, numerical, and categorical similarity score, a weighted average score is computed for each pair of products using the following formula:

$$\text{Weighted Average Similarity} = \frac{\text{wt} \times \text{St} + \text{wt} \times \text{Sc} + \text{wn} \times \text{Sn}}{\text{wt} + \text{wc} + \text{wn}}$$

Where:

- *wt* is the weight for textual attributes.
- *St* is the similarity score for textual attributes.
- *wc* is the weight for categorical attributes.
- *Sc* is the similarity score for categorical attributes.
- *wn* is the weight for numerical attributes.
- *Sn* is the similarity score for numerical attributes.

Use: This method offers users the flexibility to prioritize certain descriptor types, ensuring the composite similarity aligns with specific contexts or preferences.

6. Graph Construction & Community Detection

Concept: The final composite similarities lay the groundwork for a graph where products are nodes, and similarities define the edges between them. Originating from the graph theory and network science, the Leiden algorithm, used here, is renowned for detecting communities within complex networks. It optimizes modularity, a grouping quality metric, and ensures higher quality partitions compared to many other methods.

Use: In the scope of Product Similarity Accelerator, once the product graph is built, the Leiden algorithm identifies groups of similar products. These clusters represent products that are more similar to each other than to products outside their group, thereby defining 'similarity groups' of products.

Thus, users can extract meaningful clusters of similar products, offering deeper insights and aiding in data-driven decision-making processes.