Content Analyzer Plug-in for GAX

Text Preprocessing

3/14/2025

Text Preprocessing

As part of scheduling training, you can use the **Text preprocessing** pane (to the right of the **Model Training Options** pane) to remove extraneous text from the Training Messages of a Training Data Object. You create filters (patterns) that search for text and perform various deletion operations. This can be helpful when the emails that you want to use for training contain significant amounts of text that is:

- Irrelevant or misleading for classification purposes, and

- Identifiable by a regular expression.

Text Preprocessing

1. Click the plus-sign icon to create a new rule.

2. Types of rule:

   - **DELETE AFTER**—Search for a match to the pattern, then delete all text after and including the matching text.

   - **DELETE BEFORE**—Search for a match to the pattern, then delete all text before and including the matching text.

   - **DELETE ALL IF FIND**—Search for a match to the pattern, then delete the entire Training Message

that includes the matching text.

- **DELETE ALL IF NOT FIND**—Search for a match to the pattern, then delete the entire Training Message if it does not include the matching text.

- **DELETE PATTERN**—Search for a match to the pattern, then delete only the text that matches the pattern.

3. Test the pattern. Enter text, the result appears. If you modify the rule, you'll have to enter the text again to see the result from the modified rule.
   For the two DELETE ALL types, if you test the pattern and it finds a match, the **Output** window is empty. In actual use if there is a match, the entire Training Message is deleted from the Training Data Object.