Structured Analytics – Repeated Content Identification Setup Basics

This guide contains a workflow with basic information for setting up repeated content identification.

For more detailed information, refer to the documentation.

Repeated content identification setup

The setup for running language identification is comprised of two components:

1. Saved search
2. Structured analytics set

1. Saved Search Setup

Use the following conditions and fields to create the saved search used for email threading. You don't need to set a sort order on this search.

Search Name

There is no recommendation for the saved search name. Follow your team's normal protocol for naming searches.

Conditions

The condition for this search can be the same as the Conceptual Index search if it is different than the conditions noted below.

- Extracted text size is greater than 0 kb
- Extracted text size is less than 30,000 kb

Note: For workspaces with millions of documents, we recommend that you consider a sampling workflow. For more information, see Sampling for Repeated Content

Fields

Any fields are acceptable.

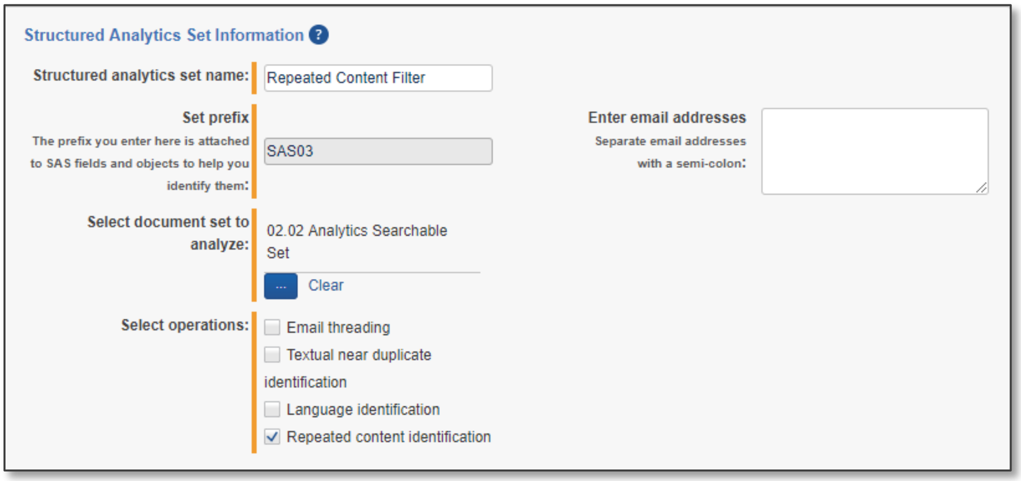
2. Structured Analytics Set

Here are the steps and choices for creating a structured analytics set.

Structured Analytics Set Information

- **Structured analytics set name** - enter a name for the structured analytics set.

- **Set prefix** - keep the default prefix or add your own prefix. Shorter prefixes (even just two characters, e.g. “LI”) take up less space in your views.
- **Select document set to analyze** - select the saved search you created above.
- **Select operations** - select Repeated content identification.

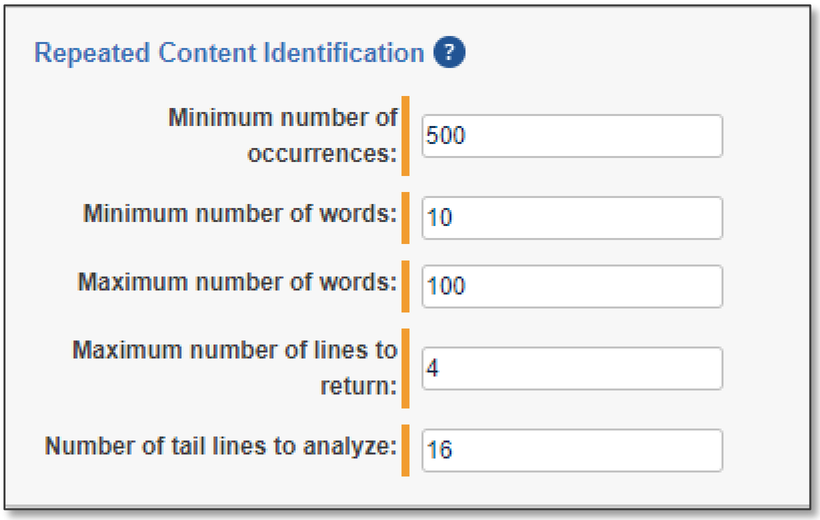


Structured Analytics Set Information ?

Structured analytics set name:	Repeated Content Filter
Set prefix The prefix you enter here is attached to SAS fields and objects to help you identify them:	SAS03
Enter email addresses Separate email addresses with a semi-colon:	
Select document set to analyze:	02.02 Analytics Searchable Set ... Clear
Select operations:	□ Email threading □ Textual near duplicate identification □ Language identification ☑ Repeated content identification

Repeated Content Identification

- **Minimum number of occurrences** - the minimum number of times a phrase must appear to be considered repeat content. We typically set this to .005 times the number of documents in your saved search.
- **Minimum number of words** - leave as default.
- **Maximum number of words** - leave as default.
- **Maximum number of lines to return** - leave as default.
- **Number of tail lines to analyze** - leave as default.



Repeated Content Identification ?

Minimum number of occurrences:	500
Minimum number of words:	10
Maximum number of words:	100
Maximum number of lines to return:	4
Number of tail lines to analyze:	16

Advanced Settings

- Choose the appropriate Analytics server.