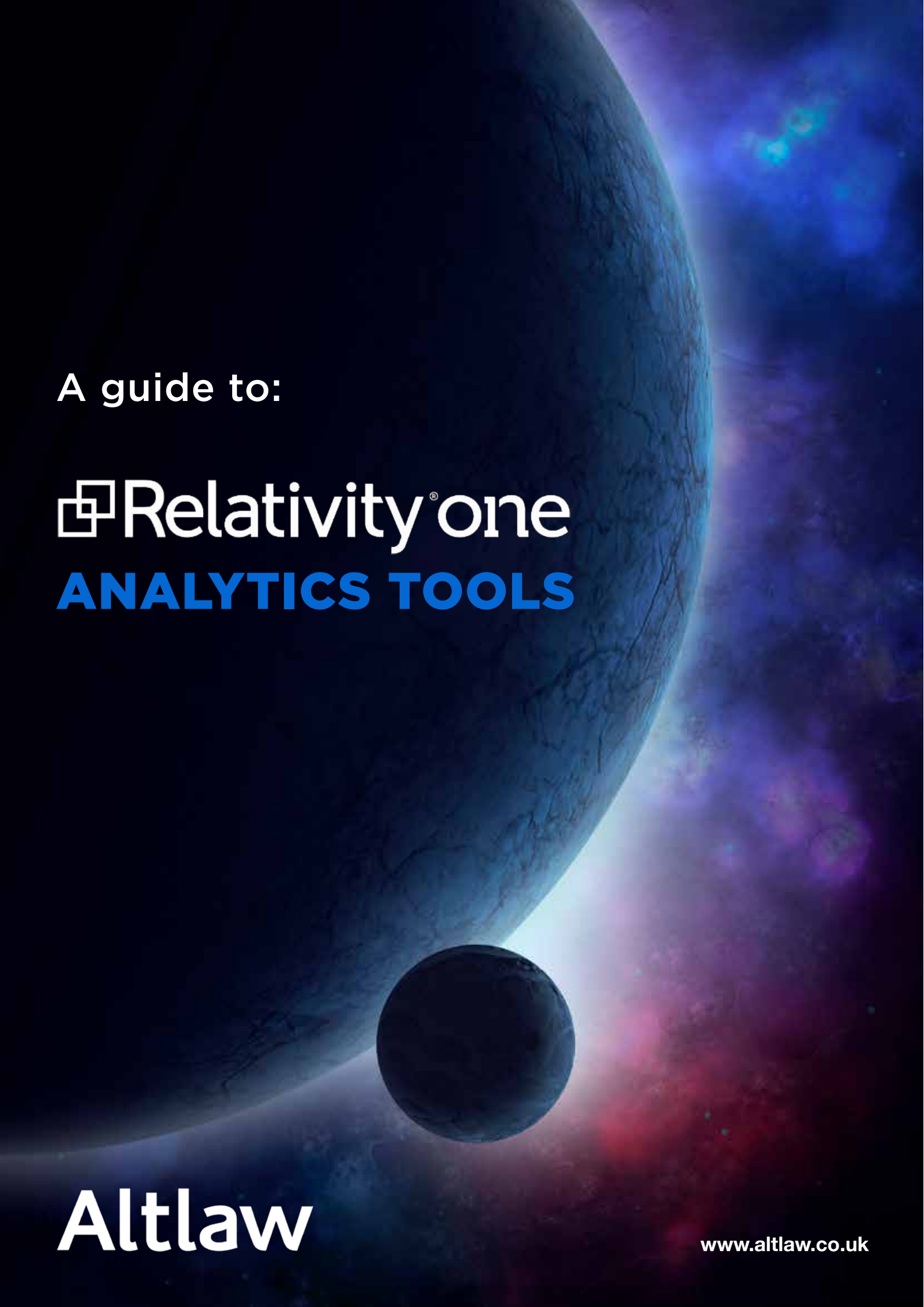


A guide to:

Relativity®one

ANALYTICS TOOLS

Altlaw

www.altlaw.co.uk

INTRODUCTION

This publication is a brief description of some of the analytics tools available within RelativityOne.

There is information on the different reasons for using these tools, their weaknesses and suitable use cases.

If you would like to discuss any of these tools further, or would like a conversation whether any of these tools may be useful for your case please contact your Altlaw Project Manager.

Analytics Tools

This document covers the following Analytic tools:

- Email Threading
- Clustering
- Categorisation
- Active Learning
- Repeated Content Filtering
- Textual Near Duplicate Identification
- Language Identification
- Keyword Expansion
- Find Similar Documents

If you would like to discuss any of these tools further, or would like a conversation whether any of these tools may be useful for your case please contact your Altlaw Project Manager.

EMAIL THREADING

- **What is it?**

  Email Threading gathers all forwards, replies reply-all messages and attachments from an email chain and groups them together for ease of review.

- **Why should you use it?**

  - Can cut down on number of documents to review

  - Prevents reading duplicative content

  - Uses email headers and body content, not hash values, so can identify duplicate emails even where email communication fields display email addresses differently i.e. fully qualified email address/friendly name/Exchange connection strings

  - Improves quality of review by showing the entirety of the conversation

  - Allows for the full email chain to be reviewed in sequential order, at once and not have individual emails within the chain mixed in with the rest of the data set

  - Scanned hard copy emails can be threaded

  - Identifies gaps/missing emails.

- **Weaknesses**

  - Can result in lengthy emails to review

  - Can result in disclosing full email chains when only one email is relevant

  - Can lead to large redaction exercise.

- **When Suitable**

  - When reviewing substantial amounts of email data

  - When data set contains email data from a number of different custodian mailboxes.

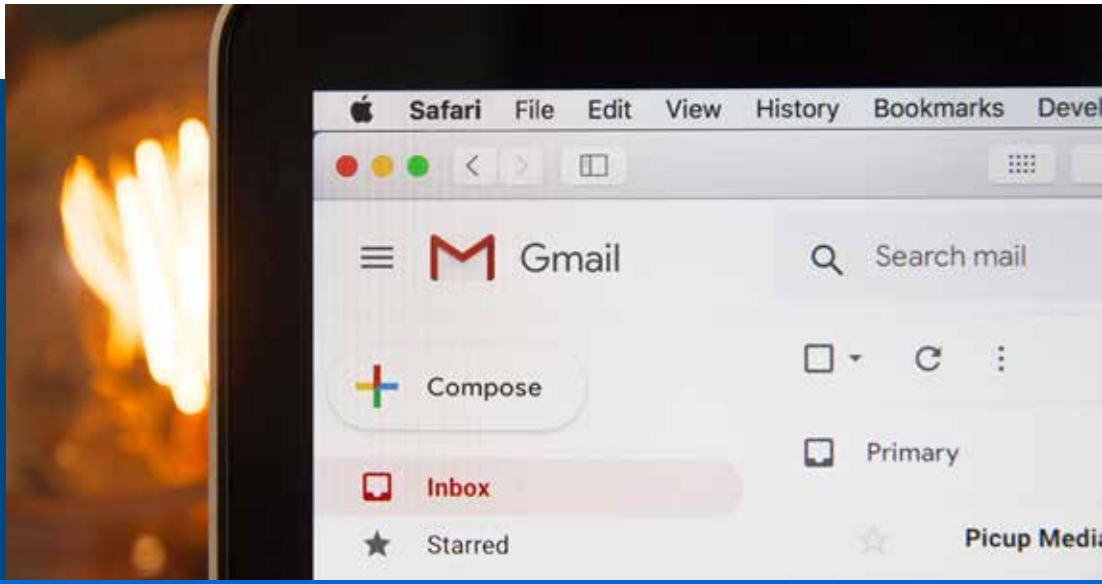
- **When unlikely to be suitable**

  - For standalone electronic documents

  - When reviewing non-email data

  - When coding for privilege this may lead to a large redaction exercise.

Email Threading gathers all forwards, replies reply-all messages and attachments from an email chain and groups them together for ease of review.



CLUSTERING

- **What is it?**

  Clustering groups conceptually similar documents together and places them into logical groups called clusters.

- **Why should you use it?**

  - Doesn't require user input or example documents in order to be applied
  - Allows you to deep dive into your data prior to review
  - Creates easy to navigate clusters named by their conceptual content
  - Can prioritise review by focusing on the most relevant clusters
  - Identifying irrelevant topics and quickly disregard not relevant documents which can speed up the review
  - Can reduce the risk of errors and improve coding consistency by assigning single clusters to the same reviewer
  - Help in creating more accurate searching criteria by identifying synonyms which may have been missed by relying solely on keyword searches
  - Allows you to gain a high level overview of themes discussed within your collection of documents.

- **Weaknesses**

  - May require some manual clean up to yield better clusters
  - Will break the chronology of the review queue to review by cluster
  - High population of foreign languages may need to be clustered separately.

- **When Suitable**

  - When presented with an unfamiliar data set
  - Prior to disclosure to check for coding consistency and ensure documents are not missed
  - During a managed review to QC the review teams work.

- **When unlikely to be suitable**

  - If data set is made up of very long documents which discuss a large number of different concepts
  - When data set is made up primarily of images or media files
  - If reviewing document chronologically is crucial.

Clustering groups conceptually similar documents together and places them into logical groups called clusters.

CATEGORISATION

- **What is it?**

  Identifies and groups conceptually similar documents together which can then be applied across your data set.

- **Why should you use it?**

  - Allows documents which deal with multiple concepts to be classified accordingly and be designated into multiple categories. A rank is then assigned showing how conceptually similar to each of the assigned categories a document is

  - Once examples have been submitted, it can be used to quickly sort documents into key issues, or identify hot documents which can be used to prioritise your review.

- **Weaknesses**

  - Requires example data/user input

  - Works best when the categories of interest have been identified

  - Example data must be focused entirely on a single concept, with at least two detailed paragraphs of meaningful text, free from distracting text such as repeated text, headers and footers

  - Any single document can only belong to a maximum of 5 categories.

- **When Suitable**

  - When particular issues or categories of interest have been identified

  - At least one example document focused on the specific conceptual topic has been identified for each category

  - When receiving a substantial data set i.e. a received disclosure that needs to be coded for issue, after having already coded for issue on your own dataset.

- **When unlikely to be suitable**

  - When categorising a dataset where a document may belong to more than 5 categories, i.e. if coding for a long list of key issues.

Identifies and groups conceptually similar documents together which can then be applied across your data set.

ACTIVE LEARNING

Active Learning is a technology-assisted review tool which predicts which documents are most likely to be relevant allowing your data to be organised quickly. There are two methods of review:

Prioritised Review

- **What is it?**

  Serves the documents ranked highest that the system believes are most likely to be relevant.

- **Why should you use it?**

  - Can speed up review and avoid costly review time
  - Allows you to quickly locate and review the most relevant documents
  - Elusion testing is a validation test which allows you to make an educated judgement on when to stop your review
  - Continuously learns from coding decisions and updates document ranks every 20 minutes ensuring the documents deemed likely to be most relevant are served up to the front of the review queue
  - Is language agnostic
  - Can be used to QC previously coded documents and find outliers and coding inconsistencies
  - Additional documents can be added once the review has begun and will be ranked when the classification next rebuilds
  - Provides clear visualisation showing review progress in real time
  - Minimal training only requires 5 documents coded with your positive choice, and 5 coded with your negative choice to rank documents
  - Can be run in combination with other analytic tools.

- **Weaknesses**

  - Will serve up random documents until training quota is met. Targeted searches may be required to find relevant documents to train the model if data set has low richness
  - User judgement must be made, with the assistance of statistical analysis, regarding when to stop the review
  - Settings cannot be changed once the review is started.

- **When Suitable**

  - When you want to review relevant documents and their family together
  - For data set with an expected lower richness level (not relevant documents)
  - Data sets >1000
  - When reviewing complete data sets
  - When reviewing filtered data sets
  - When you need to quickly review the most responsive documents.

- **When unlikely to be suitable**

  - For small volumes of data
  - Data sets made up of images or media files
  - Data sets made up of scanned hardcopy handwritten documents
  - Data sets with poor quality OCR
  - To locate privileged/not privileged documents.

ACTIVE LEARNING

Coverage Review

- **What is it?**

  Trains the model and quickly separates the documents into their positive and negative choice categories. Documents which are going to be the most impactful at training the model will be served up.

- **Why should you use it?**

  - To quickly separate documents into Relevant/Not Relevant categories

  - Is language agnostic.

- **Weaknesses**

  - Does not serve family documents together

  - Settings cannot be changed once the review is started.

- **When Suitable**

  - When a quick production is necessary

  - For a large project where not all relevant documents must be reviewed

  - For investigation and information mining.

- **When unlikely to be suitable**

  - For projects where all relevant documents must be reviewed

  - Data sets made up of images or media files

  - Data sets made up of scanned hardcopy handwritten documents

  - Data sets with poor quality OCR

  - If documents must be reviewed alongside their family.

Trains the model and quickly separates the documents into their positive and negative choice categories. Documents which are going to be the most impactful at training the model will be served up.

REPEATED CONTENT FILTERING

- **What is it?**

  Repeated Content Filtering identifies commonly occurring text within your dataset and then supresses this content from your Analytics.

- **Why should you use it?**

  - Identifies boiler plate text and commonly used footers which can be used to improve keyword searches

  - To improve the quality of an Analytics index and prevent boiler plate text and confidentiality footers overshadowing a documents authored content

  - Supresses matching text from the Analytics index it has been linked to, but does not alter the original document text

  - Can be used alongside regular expression to filter out commonly occurring pattered text such as URLs or bates stamps.

- **Weaknesses**

  - Cannot be directly applied to dtSearch or Search Term Reports

  - Requires some user configuration such as Number of Occurrences and Word Count.

- **When Suitable**

  - When running conceptual Analytics across your dataset

  - When running keyword searches which yield high false positive results due to matching terms found in email footers/ boiler plate text.

- **When unlikely to be suitable**

  - When not using Conceptual Analytic tools

  - Data sets made up primarily of images or media files

  - Data sets made up of scanned hardcopy handwritten documents

  - Data sets with poor quality OCR.

Repeated Content Filtering identifies commonly occurring text within your dataset and then supresses this content from your Analytics.

Altlaw

TEXTUAL NEAR DUPLICATE IDENTIFICATION

- **What is it?**

  Textual near duplicates identification analyses the extracted text of all documents and determines a percentage of similarity for each document compared to all others within the data set.

- **Why should you use it?**

  - Quickly identify textually similar documents within your data set to accelerate your review

  - Doesn't rely on hash values

  - A QC tool i.e. identifying near duplicate documents and compare Relevance, Privilege or Issue coding decisions.

- **Weaknesses**

  - Can be overly inclusive if percentage similarity is too low.

- **When suitable**

  - When hash values are not available

  - When your data set may contain two versions of the same document in different format, i.e. a native email, and a PDF copy

  - When metadata spoliation has occurred and hash values may not match.

- **When unlikely to be suitable**

  - Documents with a low word count

  - Data sets made up of images or media files

  - Data sets made up of scanned hardcopy handwritten documents

  - Data sets with poor quality OCR.

Textual near duplicates identification analyses the extracted text of all documents and determines a percentage of similarity for each document compared to all others within the data set.

Altlaw

Relativity one

LANGUAGE IDENTIFICATION

- **What is it?**

  Language identification examines the extracted text of each document to determine the primary language and up to two secondary languages present. This allows you to see how many languages are present in your collection, and the percentages of each language by document.

- **Why should you use it?**

  - Supports 173 languages (Full list available on request)
  - Considers all Unicode characters and understands the characters associated with each of the supported languages
  - Running language identification will not impact review time
  - Allows you to identify languages within your data set you may be unaware of
  - Allows foreign language documents to be isolated which can:
    - Produce better quality Analytics indexes
    - Enable separate foreign language review queue. Direct foreign language documents to foreign language reviewers

- **Weaknesses**

  - May give false positive results for example in emails if the body is in English but contains foreign language email footers.

- **When would it be suitable?**

  - All electronic document data sets with good quality OCR
  - When reviewing unfamiliar data sets
  - When running other Analytic tools which may benefit from split language indexes.

- **When is it unlikely to be suitable?**

  - Data sets made up of images or media files
  - Data sets made up of scanned hardcopy handwritten documents
  - Data sets with poor quality OCR.

Language identification examines the extracted text of each document to determine the primary language and up to two secondary languages present.

KEYWORD EXPANSION

- **What is it?**

  Keyword expansion allows a block of text, or term to be submitted and return a list of the more conceptually related terms within your data.

- **Why should you use it?**

  - Identify how a concept or term is expressed in a different language within your data set

  - Allows you to expand on a starting list of keywords and identify more relevant terms leading to more accurate searches

  - Identify synonyms or strongly related terms from your predefined keywords which you may not have considered

  - Provides a rank score of how closely related returned keywords are to the principle term.

- **Weaknesses**

  - Keyword expansion can only be run one word/phrase at a time.

- **When would it be suitable?**

  - When running keyword searches

  - When dataset contains documents from multiple languages

  - When trying to improve searching criteria

  - When trying to identify the different ways a concept has been relayed within the document set.

- **When is it unlikely to be suitable?**

  - Data sets made up of images or media files

  - Data sets made up of scanned hardcopy handwritten documents

  - Data sets with poor quality OCR.

Keyword expansion allows a block of text, or term to be submitted and return a list of the more conceptually related terms within your data.

FIND SIMILAR DOCUMENTS

- **What is it?**

  Find similar documents allows you to identify conceptually similar documents to the document you are viewing.

- **Why should you use it?**

  - Allows you to quickly find additional relevant documents which may have been missed

  - Can be used to QC coding choices and ensure review consistency prior to production.

- **Weaknesses**

  - Results require some manual QC

  - Can produce false positive results.

- **When would it be suitable?**

  - When hash values are not available

  - When your data set may contain two versions of the same document in different format, i.e. a native email, and a PDF copy

  - When metadata spoliation has occurred and hash values may not match

  - When trying to identify multiple versions of a document

  - When coding for privilege to ensure all privileged documents have been identified and coded correctly and redacted as necessary to prevent privileged information accidently being disclosed.

- **When is it unlikely to be suitable?**

  - Data sets made up of images or media files

  - Data sets made up of scanned hardcopy handwritten documents

  - Data sets with poor quality OCR

  - For large documents which discuss multiple topics

  - Documents made up primarily of numbers.

Find similar documents allows you to identify conceptually similar documents to the document you are viewing.