



Analytics Manual

v6.6

August 24, 2010

Contents

1	Overview	4
2	Latent Semantic Indexing.....	4
2.1	Concept Space.....	4
3	Concept Search.....	5
4	Search Index vs. Categorization Index	5
4.1	Search Index.....	6
4.2	Categorization Index.....	6
5	Executing a Concept Search	7
6	Related Terms.....	7
7	Document Classification.....	8
7.1	Dynamic Clustering.....	8
7.2	Categorization	9
7.2.1	Effective Example Documents	10
7.2.2	Categorization Settings	11
7.2.3	Primary Language Identification using Categorization	12
8	Optimization of Indexes.....	12
8.1	Training Documents	13
9	Index Creation and Server Information	13
9.1	Required Server Resources.....	14
9.2	Index Build	14
9.3	Working with an Index.....	15
9.4	Re-Indexing	15
10	Workflow Solutions	16
10.1	Keyword Expansion	16
10.2	Clustering.....	16
10.3	Categorization with primary language identification.....	16

10.4	Categorization	17
10.5	Concept Search	17
10.6	Similar Documents	17
11	Appendix A Primary Language Identification (PLI)	18
11.1	Importing Primary Language Identification Data	18
11.2	Categorization for Primary Language Identification	20

1 Overview

Relativity Analytics incorporates Content Analyst Advanced Analytics Engine as a way to leverage technology for accelerated document review. This document outlines the searching capabilities found in Analytics and provides information on the following:

- Latent Semantic Indexing
- Concept search
- Related terms
- Document classification
- Index creation and optimization

2 Latent Semantic Indexing

Relativity Analytics uses a proprietary indexing technology called Latent Semantic Indexing (LSI). LSI does not use ancillary linguistic references such as a dictionary or thesaurus to discover semantic knowledge. Instead, LSI leverages sophisticated mathematics to discover term correlations and conceptuality within documents.

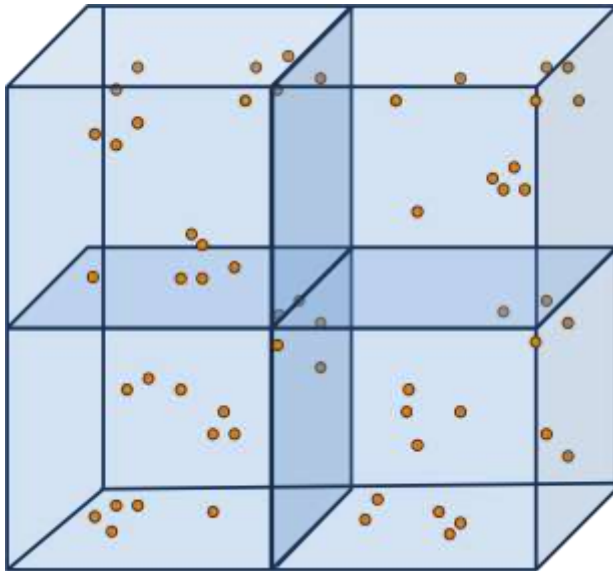
LSI is a wholly mathematical approach to indexing documents. LSI enables Relativity Analytics to learn language and ultimately the conceptuality of each document by first processing a set of data called a training set. The training set of documents may be the same as the set of documents that you want to index or categorize, it may be a subset, or it could be a completely different set of documents.

2.1 Concept Space

When Relativity builds an index, it first uses this training set of documents to build a mathematical model. This is called a concept space. The documents you are indexing or categorizing can be mapped into this concept space. While this mathematical concept space is many-dimensional, you can think of it in terms of a three-dimensional space such as a cube or a room.

The training set of documents enables the system to size this space as well as to create the algorithm to map searchable documents into this space. In our analogy of a three-dimensional room, documents that are closer together in this concept space are inherently more conceptually similar than documents that are further apart from each other.

The following illustration depicts a three-dimensional concept space into which documents have been mapped. Note that this mathematical mapping is potentially hundreds of dimensions, which cannot easily be displayed in a graphic.



Graphic Representation of documents mapped into a dimensional space

3 Concept Search

Concept search is very different from keyword or metadata search. A concept search performed in Relativity Analytics reveals conceptual matches between the query and the document.

A user can submit a query of any size—although a more thoroughly described concept is better than one or two terms—and receive resultant documents that contain the concept that the query expresses. The match is not based upon any specific term in the query or the document. The query and document may share terms, or they may not. The point is that they share conceptual meaning.

Concept search provides powerful benefits over keyword or metadata search. Concept search:

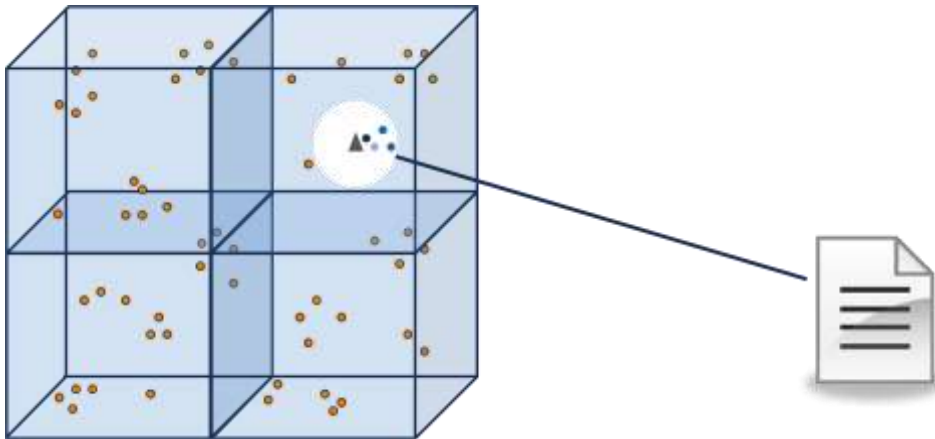
- Allows a user to find information without knowing precisely how to phrase the query.
- Circumvents the issues with language complexity. Regardless of what terminology a document incorporates, if it shares a conceptual relationship with the query, then it is returned with all the other resultant documents.
- Provides resultant documents with a focus on relevancy, not just shared terms as with keyword and metadata search.

4 Search Index vs. Categorization Index

When creating a new index, you need to decide whether you are going to create a Search index to perform queries against the entire document set being indexed or a Categorization index.

4.1 Search Index

One difference between the Search and Categorization index, in terms of the concept space, is that a Search index maps each document permanently into the concept space. In our three-dimensional analogy, each searchable document is a point in the room. When you submit a query against this index, the query is temporarily also mapped into the concept space, and the points closest to the query—each point representing an indexed document—are the most conceptually relevant to the query and are returned in the resultant list of documents.



Searching query mapped against the concept space. The triangle represent a search query and the shaded circle represents the area of documents that fall within the boundaries of the search. Here four documents from the searchable set are returned.

Because all the indexed documents are permanently mapped into the concept space, a Search index has a larger disk and RAM footprint and requires more RAM to load into memory.

When you create a Search index, Analytics permanently computes a position in the concept space for each document you designate as searchable. Searchable documents can be queried and clustered. A Search index, by its very nature of keeping track of the positions of all the searchable documents, can take longer to build and more than likely is larger in memory than a Categorization index.

4.2 Categorization Index

By contrast, a Categorization index maps only example documents permanently into the concept space. In our three-dimensional analogy, only example documents representing all your categories are the permanent points in the room. Therefore, when you go to run a dataset through Categorization, each document within that dataset is mapped temporarily into the concept space. The closest example document to the document being categorized creates the categorical designator. This requires some knowledge of the case and review of documents to determine Categories.

Because only the example documents are permanently mapped into the concept space, a Categorization has a much smaller RAM footprint and requires less RAM to load it into memory. Furthermore, because Categorization operates serially on each document being categorized and can quickly compute each document temporarily into the concept space, Categorization can process very large volumes of data within a reasonable time span. The only caveat is you must have some prior knowledge of your documents to know your example documents.

5 Executing a Concept Search

Every term known to Analytics Index has a position (what we call a “vector”) in the concept space. Furthermore, every searchable document also has a position in the concept space. An important aspect of the concept space is that vectors which are close together share a correlation or conceptual relationship. Increased distance indicates a decrease in correlation or shared conceptuality. When we speak only about documents, we can say that two documents that are close together share conceptuality, regardless of any specific shared terms.

Concept Searching, then, is the process by which a user specifies text explaining a single concept (what we will call the concept query) and then submits it to index for temporary mapping into the concept space. Analytics uses the same mapping logic to position the query into the concept space as it did the searchable documents.

Once the position of the query is established, Analytics locates documents that are close to it and returns those as conceptual matches. Keep in mind that the user can specify a threshold that widens or narrows the area around the query that Analytics engine inspects for conceptually matching documents.

The document that is closest to the query is returned with the highest conceptual score. This indicates distance from the query, not percentage of relevancy—a higher score means the document is closer to the query, thus it is more conceptually related).

6 Related Terms

Relativity Analytics can position any term, block of text, or document into its spatial index and return the closest documents. It can also return the closest terms. Doing this by submitting a single term provides you with a list of highly correlated terms synonyms or strongly related terms in your document set. When you submit a block of text or a document, you get a list of single terms that are strongly related to that document.

This type of term expansion allows you to get a sense for the different usage of language to express the same or similar concepts. In eDiscovery, you might start with the keyword list of the case and expand each one to see other highly correlated terms. Or, you might find parts of documents or complete documents for which you want to see individual terms that are highly

correlated. Analytics ability to show you related terms gives you a deeper semantic understanding of the terms and documents within an indexed dataset.

7 Document Classification

Document classification is a process by which analytics software inspects the conceptuality of each document within a set of electronic documents and places each document in the most appropriate category.

Relativity Analytics has two modes of document classification:

- Dynamic Clustering
- Categorization

Each of these modes performs document classification employing different methods, and each requires differing levels of user interaction in order to structure electronic documents hierarchically. An unsupervised method requires very little or no input from the user, while a supervised method always requires user input to classify documents.

7.1 Dynamic Clustering

Dynamic Clustering is Relativity's unsupervised mode of document classification. Dynamic Clustering does not require that the user provide any input for creating the organizational hierarchy or create any definitions of the categories.

Once Relativity has an active Searchable Analytics Index built, you can either submit conceptual queries or dynamically cluster the documents. Refer to the Relativity Admin Manual for the steps to create an index and cluster documents.

When you cluster an index, the analytics engine inspects the spatial positions of the documents within the conceptual index. Because closeness in this concept space indicates conceptual similarity regardless of specific keywords or terminology, Relativity Analytics initiates algorithms to effectively identify the most sensible groupings of documents. After Relativity Analytics creates the clusters, it runs a naming algorithm to label each node in the hierarchy appropriately so that the user understands the conceptual content of the clustered documents.

Dynamic Clustering is very appropriate to use when working with datasets about which little is known. Just selecting the document group and creating the cluster is a simple process. However, because Dynamic Clustering is unsupervised, there is no way to note what concepts are of particular interest to you. All documents in a dataset get clustered (or classified) somewhere once.

Dynamic Clustering does not perform conceptual culling of uninteresting documents. There is a group created for items without searchable text. In fact this is an easy method to find items that don't have enough searchable data. Dynamic Clustering can group documents based on custodians, search term results, date ranges or the entire database. While clustering doesn't require

much user input, a more focused approach like Categorization requires upfront user input.

7.2 Categorization

Categorization is Relativity Analytics-supervised mode of document classification. Whereas Dynamic Clustering can be a fairly automated, hands-off process, Categorization always requires thorough and up-front preparatory work. This preliminary work includes:

- Creating a taxonomy and then defining the desired conceptual content of each category through the submission of example documents to the Analytics engine.
- Example documents defined for the appropriate concepts which a document should have if it is to be deposited within the category.

While the preliminary work may sound prohibitive to end users who simply want to get down to reviewing documents, the trade-offs are compelling. Categorization exists to break users out of the habitual process of linear review.

By allowing the user to express interest in conceptual categories Analytics can provide the benefits of focused review by identifying all documents in a dataset that bear strong conceptual similarities. Studies have proven that assigning closely related documents in groups creates huge efficiencies in document review. Some partners have actualized a doubling or tripling in the number of documents reviewed per hour.

Unlike Dynamic Clustering, Categorization enables documents to be placed into multiple categories, if a conceptual match with more than one category exists. Most documents do deal with more than one concept or subject, so forcing a document to be classified according to its predominant topic may not reflect other important conceptual content within it.

Categorization avoids this by exposing a setting to the user allowing a document to go into more than one category. Categorization is best employed as a bulk-classification mechanism when the following conditions are present:

- You know the categories or issues of interest.
- You know how you want to title the categories.
- You have one or more focused example documents to represent the conceptual topicality of each category.
- You have one or more large sets of data that you want to categorize rapidly without any user input after setting up the category scheme.

Example documents should always be identified by the subject matter experts of any case. Typically, this would be one or more of the attorneys most familiar with the case and with the issues surrounding the case. Before identifying example documents, though, the subject matter expert should have a clear idea of what categories or issue tags they want to establish.

Remember, an example document conceptually defines a category, so you must first know what your categories are before you can find the most appropriate example documents. Keep in mind that a category does not necessarily need to be focused around a single concept. For example, a category might deal with fraud, but different example documents for the category might reflect different aspects of fraud, such as fraudulent marketing claims, fraudulent accounting, and fraudulent corporate communications. The key point, though, is that each example document should be focused on a single concept.

Often, the subject matter expert possesses hot or relevant documents to the case. These documents are ideal starting points. If the body of the relevant documents is substantial and covers all the intended categories, then you might be ready to begin categorization. However, this is more than likely not the typical situation, so further data mining is usually necessary to beef up the corpus of example documents. Dynamic Clustering, free-form Concept Search, and Find Similar all work well in tandem to help locate a sufficient number of example documents.

When performing Categorization, you can choose to have the example documents that define categories also provide semantic knowledge to the Analytics engine as training documents. Equally, you can even designate the documents you intend to categorize as training documents. This latter approach ensures that Analytics semantic knowledge encompasses all the meaningful verbiage within the documents to be categorized.

Categorization can be an iterative process, especially during the initial phases of a case. Therefore, you should understand how to analyze the categorization results, export those results, and refine your example documents for future categorization processes.

Categorization is useful at many stages in the EDRM workflow, including early case assessment, first pass review, issue review, and quality assurance.

7.2.1 Effective Example Documents

Example documents define the concepts that characterize a category. Without example documents, the system would have no way of knowing the requisite concepts that a document should possess if it belongs in a category.

Therefore, properly defining example documents is probably the most important step in setting up Categorization. The rule of “garbage in equals garbage out” very much applies to the act of defining example documents. As an overarching rule, an example document should possess these qualities:

- **A single conceptual focus**
 - The example document should explain a single concept relevant to the category. This is not to say that a category should only focus on one concept. Quite the opposite, which is the reason that multiple examples should always be present for any given category.

- **A fully described concept**
 - The example document should fully explain the single concept it is defining. Single terms, phrases, and sentences do not convey enough conceptual content for Relativity Analytics to learn anything meaningful from them. Strive for a fully developed paragraph or two, but usually no more (typically, most writers shift conceptual focus between paragraphs).
- **Clean example documents**
 - As with training documents, example documents also should be free of distracting text such as headers, footers, repeated text or garbage text such as OCR errors. When creating example documents, ensure that they are clean and free of this type of verbiage. Finding example documents usually begins with the project manager or the subject matter expert. Hot documents that are very relevant to the case are the best places to start. Excerpting the particularly interesting sections of hot documents is very effective. However, don't forget other Relativity Analytics functions including keyword search, concept search and Dynamic Clustering can be equally effective in locating sources of example documents.

Here is an example of how you might set up categorization example docs:

- **Cars**
 - Example Doc 1 focuses on internal combustion engines
 - Example Doc 2 focuses on transmissions
 - Example Doc 3 focuses on chassis
- **Planes**
 - Example Doc 4 focuses on jet propulsion engines
 - Example Doc 5 focuses on wings
 - Example Doc 6 focuses on cockpit controls

In this case, all six example documents would be permanently mapped into the concept space, and if the default threshold is retained, then each example document would have a hit sphere of 50.

During categorization, if the first document of a dataset—perhaps a document discussing wishbone suspension—is mapped into the concept space falling within Example Doc 3's hit sphere, Relativity Analytics notes that the document belongs to the Car category and then proceeds to categorize the next document. In this case, Example Doc 3 is the closest example document and is the one that caught the document.

7.2.2 Categorization Settings

Several settings affect Categorization, including the number of allowable categories per document and the threshold setting. At times, you might want to allow documents to be categorized into more than one category. This makes sense, because most documents involve many concepts, so a single document

may have a strong correlation with example documents from multiple categories.

The threshold setting for Categorization functions in exactly the same manner as the threshold with a concept search. The threshold defines a hit sphere around each example document in the concept space, so when a document being categorized falls within this area it is considered a conceptual match with the example document. For higher recall but less precision, decrease the threshold; for higher precision but less recall, increase the threshold.

7.2.3 Primary Language Identification using Categorization

Primary Language Identification (PLI) is a specialized application of Relativity Analytics user driven Categorization. Using a well-tested set of example documents and stop word list created by CAC, you can categorize any dataset and determine the predominant or primary language of each document. Currently, Analytics software supports approximately 20 languages.

One point to keep in mind is that PLI determines the primary language of the document. If a document contains multiple languages, the strongly predominant one determines category. Therefore, PLI is broad division of the documents into categories indicating the strong presence of any language.

PLI is ideal during the early phases of eDiscovery, especially Early Case Assessment (ECA) when you are trying to assess the costs of ongoing litigation. PLI provides valuable information about the multiple languages within a dataset, which allows you to make informed decisions about translational services and multi- or cross-lingual search strategies.

Please refer to Appendix A for instructions on setting up a index for language identification.

8 Optimization of Indexes

The results of Analytics searches is dependent upon the quality of the data. Providing bad or extraneous text to the Concept Search Index will pollute your results. Applying filters to the data as it is loaded to the search Index will clean up this polluted data for index building and search retrieval. Filters do not affect the original data and only apply to the index providing a alternate version of the data for index purposes.

Word repetition is how concept searching builds the concept space. Sensing words' correlations as they repeat in documents builds relationships that are expressed in conceptual searching.

Email addresses repeat at the top of documents and might provide improper correlations with the content of a document. Removing the email header information or repeating information is an excellent method of improving results. Here are a list of the filters used by Content Analyst software to remove extraneous data from the Analytics text.

- **Email Filter** – The email filter will remove the header information from the text. This filter essentially looks at the To, From, Subject and Date. This is generally recommended to ensure that the authored content is not overshadowed by headers in the conceptual space.
- **Go Words Filter** – OCR isn't always a perfectly arranged groups of words. Sometime images or hand writing can create odd strings of characters to form. If these items repeat they might become associated into the index. The Go Words filter makes the system use a dictionary to determine if the characters truly make a word.
- **Regular Expression Filter** – Regular expressions are a sort of programmatic way of finding text that meets the criteria provided in the expression written. Repetitive content in footers is an excellent example of when regular expressions can be used to remove content that repeats across documents. This repetition of content might bring documents together conceptually without merit.

8.1 Training Documents

Whenever you build an index you must provide the system with training documents. Input of Training documents is the system learning language and the correlations between terms and ultimately conceptuality. It is the mass of training documents that formulate the mapping scheme of all documents into the concept space. You are directly affecting the ultimate results of your categorization when you specify training documents.

You can use the dataset of documents you are going to categorize as the set of training documents, and in many instances this approach is highly desirable. However, with larger datasets, you would not necessarily want to use the entire set of documents as training, as a larger training set requires more server resources such as RAM memory.

9 Index Creation and Server Information

As stated in the introduction, every search and analytics engine needs to discover the text for all the data intended to be queried through a process called indexing. Indexing is one of the key areas where different technologies can be applied to perform essentially the same task. It is also the key area of differentiation in terms of speed and accuracy of the analytics engine.

Relativity Analytics implement proprietary indexing technology to discover and index structured and unstructured data. A purely mathematical approach to indexing text such as CA's involves sophisticated linear algebraic algorithms.

Relativity Analytics inspects all the meaningful terms within a document and uses this holistic inspection to give the document a position within a spatial index. The benefits of LSI's mathematical approach include the following:

- Relativity Analytics learns term correlations (interrelationships) and conceptuality based on the documents being indexed. Therefore, it always is “up-to-date” in its linguistic understanding.
- Relativity Analytics indexes are always resident in memory when being worked with. Therefore, response time is exceedingly fast.
- Relativity Analytics is inherently language agnostic. Therefore, it can index most languages and accommodate searches in those same languages without additional training.

9.1 Required Server Resources

To perform the sophisticated mathematics required to index documents requires substantial server resources. Server memory is crucial to building an Analytics index. The more memory your server has, the larger the datasets that can be indexed without significant memory paging. Furthermore, increased memory speeds up index build times. Analytics also depends on CPU and I/O resources at various stages of the build. Ensuring that your server has multiple processors and fast I/O throughput also creates efficiencies in the build process.

Finally, Relativity recommends installing Analytics on a 64-bit server with a 64-bit operating system for production environments. Another thing to keep in mind is that an Analytics index is ultimately stored on hard disk, though it is loaded into memory when you want to query across the documents.

While the destination of the built Relativity Analytics indexes is configurable, storing these indexes on a network storage device introduces another potential bottleneck of network bandwidth and speed. Our own tests show that building an index is most efficient when the dataset being indexed is local to the Relativity Analytics server and also when the built index is stored locally.

If that is not possible, then you should consider the network bottleneck when estimating index build times. Several factors affect the aforementioned resource consumption. These factors include the following:

- Number of total documents in the dataset being indexed.
- Number of unique terms across all the documents in the dataset being indexed.
- Total mean document size (as measured in unique terms).
- Number of configured dimensions for the index.
- Amount of metadata associated with the index.
- Number of characters in each document’s itemId.
- Configured amount of documents used as training.

9.2 Index Build

Relativity Analytics indexes are spatial, and in order to build a many-dimensional semantic or concept space, Analytics needs to perform some sophisticated mathematics. Analytics starts by using training documents to understand all unique terms in the corpus of documents and all term correlations.

At these stages, Relativity Analytics is essentially performing sophisticated mathematics to build the spatial index and computing the means by which to map searchable documents into the concept space.

After creating the concept space, Relativity Analytics needs to populate it. Depending on whether you are creating a Search or Categorization index, Relativity Analytics permanently positions each searchable document or example document into the concept space as well as builds a keyword index (if selected when you configured the index, but only for a Search index).

The process of adding searchable documents to the concept space can be relatively quick (for smaller datasets) in comparison to building the concept space or building the keyword index (if selected). Keep in mind that during this stage, Relativity Analytics creates a keyword index if you selected that option. In fact, each searchable document is added to both indexes before Relativity Analytics proceeds to the next document. Relativity Analytics displays Updating Searchable Items at this point.

As noted earlier, with large datasets, adding the searchable documents to the concept space requires less time than building the keyword index, though both activities occur during the Updating Searchable Items stage.

9.3 Working with an Index

Once an index is built, all files that make up that index—including an index-specific build log—are stored on disk. In essence, the bulk of the index includes Relativity Analytics computations for the locations in the concept space of all known terms and searchable documents.

To use an index, however, you have to enable it, which loads this data into RAM memory on the Relativity Analytics server. Of course, enabling a large number of indexes at the same time can consume much of the memory on the Relativity Analytics server, so usually you will only want to enable indexes that are actively being worked with (querying documents or classifying them).

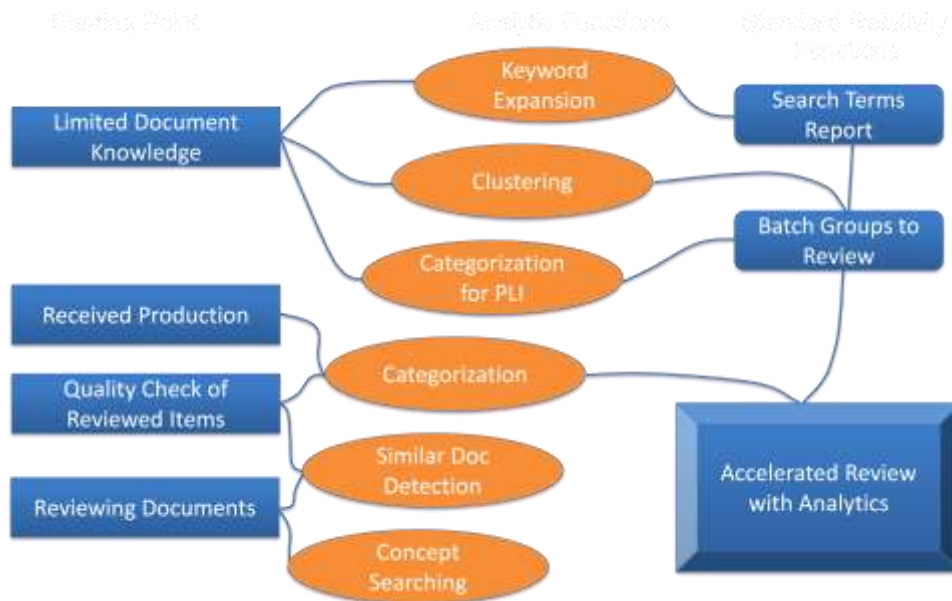
9.4 Re-Indexing

Re-indexing does not necessarily force Relativity Analytics to go through every stage of an index build. The general rule to remember is that if you do not affect the training documents—those documents which are used to learn term correlations and to create the mapping for searchable documents—then Relativity Analytics does not recompute term correlations and document positions for previously indexed documents. In essence, Relativity Analytics only adds any new searchable items to the concept space and also adds those new terms to the keyword index. This process is called folding.

Folding is a relatively quick process. However, in large operational environments, another option is to implement Relativity Analytics Live Searchable Updates feature, which allows searchable documents to be added to the index resident in memory.

This saves you from having to disable the index, rebuild the index, and enable it again (as indexes grow in size, the time to enable them also increases). At some point, you will need to rebuild the index, but in production environments where indexes are very dynamic, you might want to consider this option.

10 Workflow Solutions



Workflow Possibilities

10.1 Keyword Expansion

Keyword expansion allows you to see terms or keywords you might not have originally expected.

- Create Search Index
- Use Keyword Expansion to find all possible terms
- Create Search Terms Report of all terms
- Batch items based on terms

10.2 Clustering

Clustering groups or sorts documents based on common content. You can cluster all documents or smaller groups. After clusters are created batches can be created based on clusters.

- Create Search Index
- Create clusters
- Batch items based on clusters

10.3 Categorization with primary language identification

Using a standard set of documents with various languages and categorization you can identify documents of different languages.

- Load PLI data and identify data as PLI items
- Create Categorization Index
- Batch documents of other languages to reviewers who can translate or have non English documents translated for English reviewers before they begin review.

10.4 Categorization

Categorization takes documents already grouped or issue coded and compares them against another document set to apply established issues or categories to new documents based on conceptual content.

Examples of Categorization Workflow:

- Take initial group of documents deemed as privilege or responsive and categorize against rest of database to organize and prioritize review.
- Quality check documents flagged as privilege against rest of database to insure nothing was missed.
- Use currently reviewed documents against production from opposing counsel to prioritize the review of their documents.
- Randomly sample 10% of document database and use coding decisions on this group to categorize and prioritize review of all other documents.
- Use document coding from completed review to prioritize documents in similar new case.

Categorization Steps

- Apply issues or categories to set of documents
- Create Categorization index
- Prioritize review based on applied issues or categories

10.5 Concept Search

Concept searching is applying a block of text against the database to find documents of similar conceptual content. This can help prioritize or help find important documents.

- Create Search Index
- Use Search window to execute search

10.6 Similar Documents

Using Analytics Relativity can create a similar document window in the related items pane. These documents can be reviewed and possibly coded as a group.

- Create Search Index
- Select option to create similar documents in Search Index console

- Right click to find similar documents or view items in related items similar documents pane.

11 Appendix A Primary Language Identification (PLI)

Relativity Analytics can be used to identify documents of 18 different foreign languages. Knowing the languages of documents before a review can help with properly staffing and planning the review process. The following languages can be identified using PLI:

- Arabic
- Chinese
- English
- Finnish
- French
- German
- Hebrew
- Hindi
- Italian
- Japanese
- Korean
- Norwegian
- Portuguese
- Russian
- Spanish
- Swedish
- Turkish
- Vietnamese

11.1 Importing Primary Language Identification Data

Follow the steps below to implement PLI in a database. The following fields need to be created to utilize Primary Language Identification:

- **Field Name – Field Type**
- Language Category – Multiple Choice
- Language Example – Yes/No
- Language Category Rank – Decimal Field

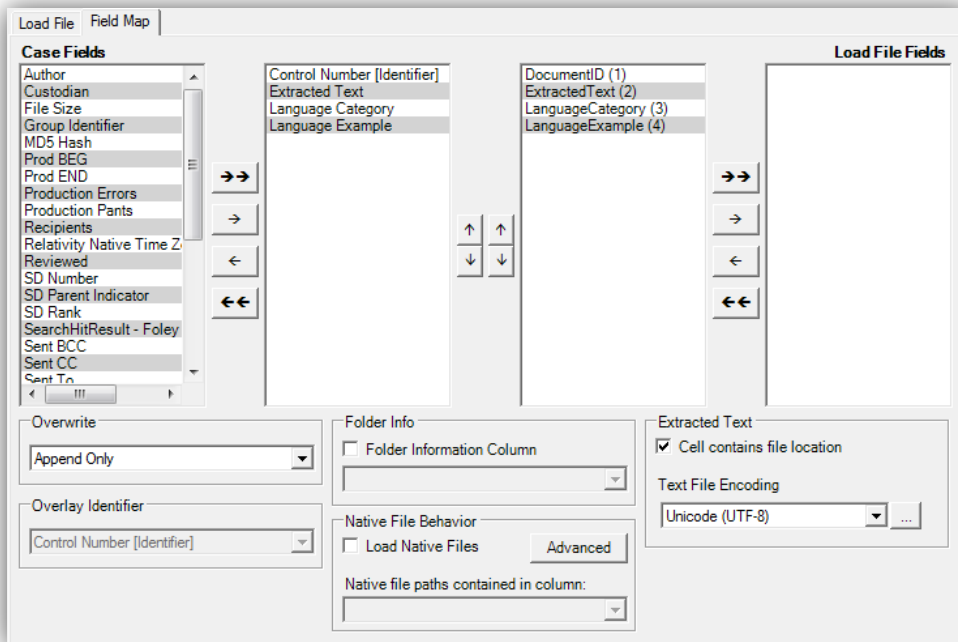
In order to perform PLI you will need to obtain the Primary Language Identification.rar file. This file is available on Customer Portal. It is recommended that you create a separate folder that is secured on your database for the sample foreign language files.

1. Unzip the Relativity Analytics Primary Language Identification.rar file
2. Open the Relativity Desktop Client (RDC), select the desired workspace

3. Right-click on the newly created Relativity PLI folder, select Import, select Load File
 - a. If no PLI folder was created then proceed by selecting Tools| Import| Document Load File.
4. Select the PLI_Import.csv as the load file by clicking on the ellipses in the Load File tab and navigating to the extraction location.
5. Map the fields in the load file to the appropriate workspace fields as seen below
 - a. **PLEASE NOTE:** The PLI data contains UNICODE characters. Please ensure that the Extracted Text field is Unicode enabled.
 - i. If Unicode is set to **No**, it is not recommended to enable this setting on databases larger than 50,000 records during peak hours.
6. Ensure that “Append Only” has been designated in the Overwrite section
7. Ensure that the “Cell contains file location” check box is checked, and the Text File Encoding is set to UTF-8 in the Extracted Text section
8. Import file



Changing your extracted text field to Unicode will increase your database size.



Import Dialog Box

11.2 Categorization for Primary Language Identification

After you have finished importing this data into your database create a saved search that returns the new foreign language documents and only the extracted text field for those documents. Name this search Relativity PLI Training Source.

PLEASE NOTE: If you configured a Long Text field above for Step 5 use this field in place of Extracted Text for this search only. Create another saved search that returns only extracted field for all documents you wish to compare against. Name this search Relativity PLI Searchable Source.



For an added efficiency you can exclude example documents from the searchable set. They will not be categorized or searched. Fewer documents to search saves the system time.

Next create a new Analytics Index from the Search Indexes tab. Setup the information as:

- Name: **Foreign Language Identification**
- Index Type: **Category**
- Dimensions and Number of Processes: **default**.
- Training Set: **Relativity PLI Training Source**
- Searchable Set: **Relativity PLI Searchable Source**
- Example Indicator Field: **Language Example**

- Category Field: **Language Category**
- Category Rank Field: **Language Category Rank**
- **Leave the rest of the choices as default**
- Concept Stop Words: **Remove the default stop words. Open the stopwords.txt file included with the PLI documents, select and copy entries in this file, paste this in the Concept Stop words section.**
- Click **Save**

Data Source	
Training Set:	Relativity PLI Training Source ▼
Searchable Set:	Relativity PLI Searchable Source ▼
Example Indicator Field:	Language Example ▼
Category Field:	Language Category ▼
Category Rank Field:	Language Category Rank ▼
Minimum Score:	50% ▼
Max Categories/document:	1 ▼
Auto Stop Feature Comp:	False ▼

Category Index settings

Utilize the Search Index Console to create the index. Click on the buttons in the following order as they become available. You might need to Refresh the Page or click off the page and return to make the buttons available.

- **Full Population**
- **Build Index**
- **Enable Queries**
- **Categorize All Documents**

Now you can check your results. Create a saved search to return all documents which displays DocumentID, Language Category and Rank. Filter for documents categorized as the desired language.