

CONCEPTUAL CATEGORIZATION

Understanding and Implementing Categorization

UNDERSTANDING CATEGORIZATION

Over the last several years legal document review has experienced a sea-change as it relates to data volumes and subsequent costs of attorney review. As we look to available technology to help bring calm to these troubled waters many have focused the spotlight on “Predictive” or “Suggestive” Coding. Regardless of the label you prefer, this technology uses machine learning capabilities to categorize documents in ways useful to the user. Whether this technology ultimately supplants much of human review or merely serves as a valuable review aid will depend on a lot of things outside of the technology itself. For now, however, it’s important to know about it and how you might appropriately implement it into your overall review strategy and workflow.

TCDI uses Content Analyst Analytical Technology (CAAT) as the back-end concept engine for document Categorization. This technology employs a supervised mode of classification which gets its power from the ability to learn nuanced aspects of a category through the inclusion of example documents. It enables you to classify documents rapidly based on the concepts contained in each. Categorization also gives you great flexibility in constructing and naming the categories of interest (the taxonomy), as well as determining what types of documents should be categorized into each.

The end result is a dramatic increase in the speed of document review through relevant organization.

WHEN TO USE CATEGORIZATION

Categorization is an excellent option when the following conditions are present:

- You have very large, unorganized datasets that you want to organize or prioritize based on document conceptuality.
- You anticipate rolling collections and database loads that need to be organized throughout the course of document review.
- You have subject matter experts that know or can define what the categories or issues of interest are.
- You know how you want to title or represent the categories.
- You can identify focused example documents to represent the conceptual topic(s) of each category.

CREATING CATEGORIES

Keep the following in mind as you are creating your taxonomy (category structure):

- A category does not necessarily need to be focused around a single concept - ask yourself, what are we really interested in?
- Ask yourself, how do the users/reviewers want to see these documents organized?
- Consider using Dynamic Clustering on a subset of data to see how dynamic taxonomies are created.
- Defer to subject matter experts about what the categories of interest are, as well as what to name each category.
- Provide common-sense names for your categories that fully describe what types of documents have been classified within them.



THE IMPORTANCE OF EXAMPLE DOCUMENTS

Categorization functions by having a set of pre-identified example documents that represent each category. Through these conceptual example documents, or exemplars, CAAT is able to derive the conceptual intent of each category. Based on this learned conceptuality, it then locates and tags all documents in a collection that are conceptual matches.

A conceptual match is categorized or tagged as belonging to its most appropriate category. Documents that do not match any category are categorized or tagged as Uncategorized.

The accuracy and relevancy of Categorization depends in large part on the effectiveness of the example documents.

Tips For Refining Examples:

- Have a subject matter expert provide any known “hot” documents for the case or project.
- Make sure any example document fully encapsulates the concept (avoid example documents with keywords, phrases, or bullet points statements only).
- With longer example documents, consider excerpting only the portions that are relevant to the concept of the category to which you are assigning it.
- If categorizing on multiple issues, consider having TCDI run a Self-Test to ensure that categories do not have a perceived overlap because of sample documents.

FINE-TUNING CATEGORIZATION RESULTS

Categorization is a process that can be iterative during the formation of categories and the testing of results. You can test your results by identifying and categorizing a subset of data that you or a subject matter expert knows well. Keep the following recommendations in mind while analyzing results and fine-tuning your Categorization structure:

- Be prepared to refine your sample set of documents based on initial Categorization testing. Ineffective results can be tracked back to the sample document that “caught” them and those sample documents can be removed from the training set.
- Results deemed highly relevant but containing low relevance scores should be considered for inclusion to the training set. This “user feedback” loop will increase the accuracy of subsequent Categorization results. As training sets are updated, also consider running occasional “Self Tests” to measure for category overlap.
- While it’s true refining the sample set will improve relevance within your categories, these efforts should be balanced based on your objective of using this technology within your overall review strategy.
- Get an understanding of any large drops in conceptual scores for documents categorized in any category. This might indicate a shift from highly relevant to less relevant documents.
- Consider allowing documents to be categorized in more than one category. Documents in multiple categories will contain different conceptual scores which can provide insight into the conceptual interrelationship between categories.
- Review the conceptual threshold settings to balance relevance and recall.

CATEGORIZATION RETURN ON INVESTMENT

Categorization demands some initial up-front work in terms of establishing the categories and identifying and refining example documents for each. However, the gains in efficiency when reviewing larger numbers of documents will be well worth it. Models have shown the use of Categorization can increase review efficiency by up to 80%. To learn more about how Categorization can be put to use in your next project, contact a TCDI representative.