



NEAR-DUPLICATE DETECTION

In addition to identifying exact-duplicate documents, substantial time and cost savings can be realized by identifying and grouping similar documents, known as "near-duplicates," and treating them in a consistent fashion. With near-duplicates making up as much as 30% or more of an electronic collection, identifying and grouping these similar documents increases speed and efficiency in the document analysis process, provides for a more consistent review of similar files, and reduces associated costs.

What is a Near-Duplicate?

A near-duplicate is a document that has a slight difference with respect

to other documents in your collection. Differences may include words, file types, and file formats, for example:

- > different text
- > different fonts
- > different metadata
- > the same document in Word and PDF formats

Near-duplicates occur most often in business correspondence, including emails and templates. Common near-duplicates include document drafts and revisions, forwarded e-mail messages (with or without added comments), and older versions of documents stored on backup tapes and other media. >>>

The screenshot displays a software interface for document review. The main window shows a table of documents sorted by 'Equiset' (similarity score). A callout box labeled 'Sort by Equiset' points to the 'Equiset' column header. Another callout, '2 documents in Equiset #4', points to two rows with an Equiset score of 4. A third callout, '3 documents in Equiset #7', points to three rows with an Equiset score of 7. A fourth callout, 'Degree of similarity to Pivot', points to the 'DocSimilarity' column. A fifth callout, 'Preview of selected document', points to a preview window showing a document titled 'Smith, Jones & Brown Publishing Author's Contract'.

StagedDocID	File Name	Equiset	PivotFlag	DocSimilarity	File Type	Title
00000003.00000.0000	ExtendedListofFields.pdf	2	Pivot		Adobe PDF	
00000004.00000.00000	FAQ.doc	3	Pivot		Word Document	
00000005.00000.00000	FileProcessor.doc	4		96	Word Document	Id
00000014.00025.000	FileProcessor.doc	4	Pivot		Word Document	Id
00000006.00000.00000	ICA installation.doc	5		94	Word Document	
00000014.00001.00003	ICA installation.doc	5	Pivot		Word Document	
00000000.00000.00000	ListofExtractedMetaFields.xls	6	Pivot		Excel Spreadsheet	
00000000.00000.00000	Search Term Overview.doc	7	Pivot		Word Document	Keyword Searches for Electron Data - TL
00000000.00000.00000	Search Term Overview.doc	7		92	Word Document	Keyword Searches for Electron Data - TL
00000000.00000.00000	Search Term Overview.doc	7		90	Word Document	Keyword Searches for Electron Data - TL
00000011.00000.000	TechLawSolutions Searching Parameters.doc	8	Pivot		Word Document	AND connector
00000012.00000.000	Tips for Drafting Search Parameters.doc	9	Pivot		Word Document	Tips for Drafting Search Parameters

NEAR-DUPLICATE DETECTION (CONTINUED)

Near-duplicates may be clearly related to another document when viewed side-by-side, but the metadata is distinct. These slight differences may be vitally important, and the risks of automatically treating them as a “true” duplicate can be significant, and subject to sanctions. Identifying and grouping near-duplicates prior to review permits your team to analyze the differences efficiently and apply consistent rationale to your review.

How Does TechLaw Solutions Identify Near-Duplicates?

Regardless of the original media, TechLaw Solutions’ experienced consultants work with you to determine the criteria for near-duplicate detection on a project-by-project basis. Textual similarities between documents are analyzed, and an advanced algorithm mathematically creates specialized groupings for documents in your collection. Based on your criteria, documents with a specified percentage of textual similarity are identified and grouped together for review. Once identified, “like” documents can be assigned to a single reviewer for efficient and consistent analysis.

Electronically Stored Information (ESI)

Identifying and suppressing exact-duplicates is a common ESI processing requirement, whether it is within or across custodians. Applying an MD5 hash algorithm creates a “digital fingerprint,” quickly identifying exact-duplicates. However, until recently, near-identical duplicates were left unorganized and unidentified unless a reviewer happened to recognize a similar document, or costly and sophisticated search strategies were utilized.

TechLaw Solutions offers a solution to the challenge of identifying near-duplicate documents. Extensive technical and project management experience, and specialized technology powered by Equivio assist in automating the near-duplicate identification process.

Hard Copy

While near-duplicate detection is typically considered for electronic files, it can also be used with digitized hard copy materials that have been converted to full-text, usually through OCR processing. Near-duplicates identified from hard copy or from electronically stored files can be grouped for simultaneous review providing enhanced consistency in considering “like” duplicates over an entire collection of data.

What are the Benefits of Using TechLaw Solutions’ Near-Duplicate Detection?

Near-duplicate identification enables reviewers to simultaneously evaluate similar documents. Along with the time and cost savings of organizing and reviewing collections by near-duplicate groups, the risk of different reviewers assigning inconsistent review decisions to similar documents is significantly reduced.

The return on investment can be substantial. Skillful document management requires fast, accurate, and consistent processing. The TechLaw Solutions’ near-duplicate detection capability accelerates your document review process and extends your team’s ability to master e-Discovery collections. ○