dtSearch Desktop/Network
Indexing and Search techniques

T207 – IDENTIFYING DUPLICATES, SELECTING & COPYING FILES

dtSearch Desktop/Network is a powerful search tool used by professionals for a wide variety of tasks. This tutorial aims to show you how to search using a "list of words", how to identify duplicates using the MD5 hash indexing option, how to sort and select search results, and how to copy selected files. These are all typical processes used in litigation, eDiscovery and forensic searching.

Course Prerequisites
dtSearch Desktop/Network 7.88 or later
T207 search query and test documents (see Appendix)

This training session covers several advanced topics of interest to those involved in litigation, e-forensics, eDiscovery and early case assessment.

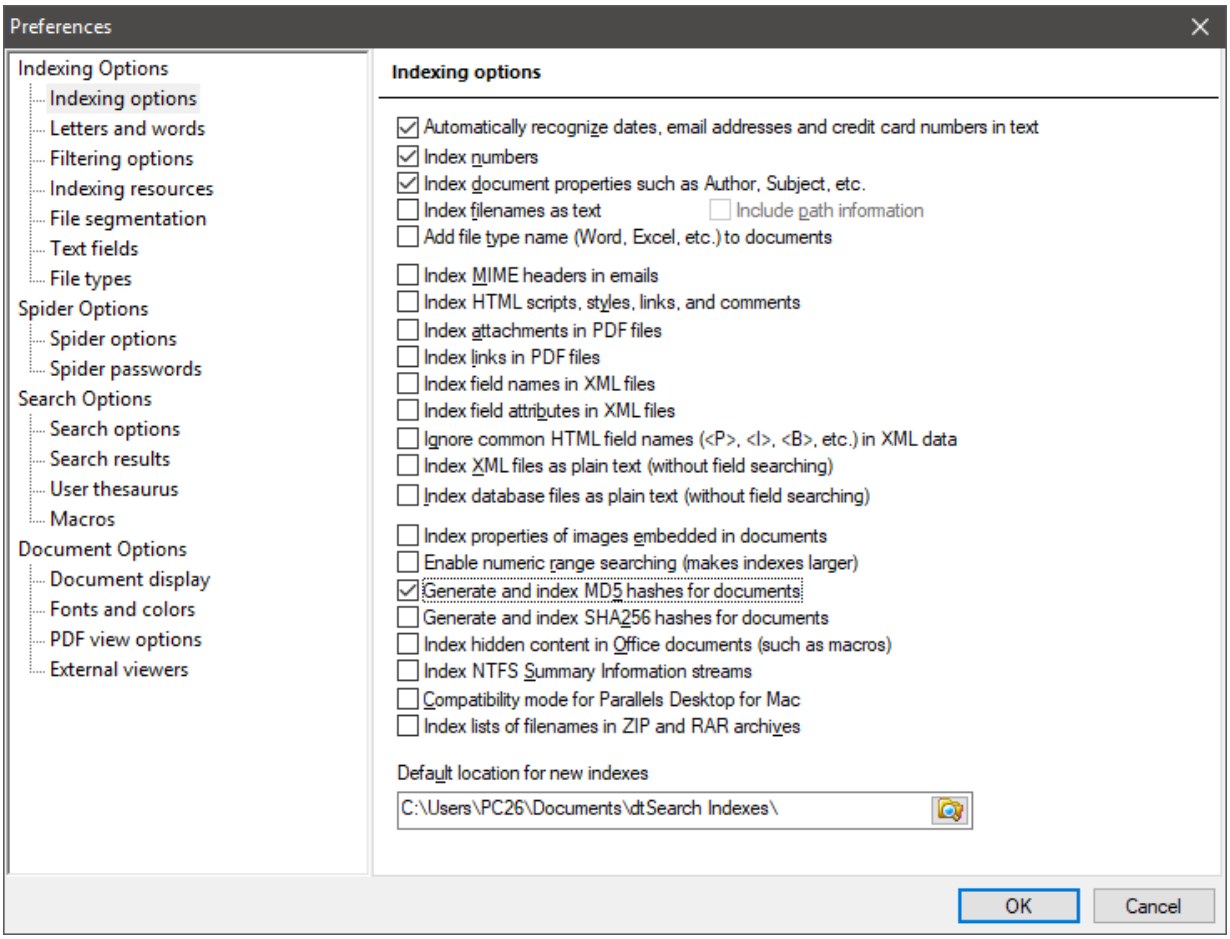
A common requirement in disputes that may require court involvement is “to meet and confer” to identify what E-data to collect for possible use in litigation. This may involve discussing where the data is stored, who created it and in what time period for example. A useful outcome could be agreeing a list of search terms.

This tutorial shows some techniques that might be useful but should not be taken as legal advice.

Before beginning the training session, all copies of dtSearch Desktop must have the same initial indexing and display setup. Access to the “List of words” and T207 test documents is also required. This can be carried out by each trainee as part of the session or by an instructor before the session starts (See Appendix).

Initial setup of dtSearch Desktop:

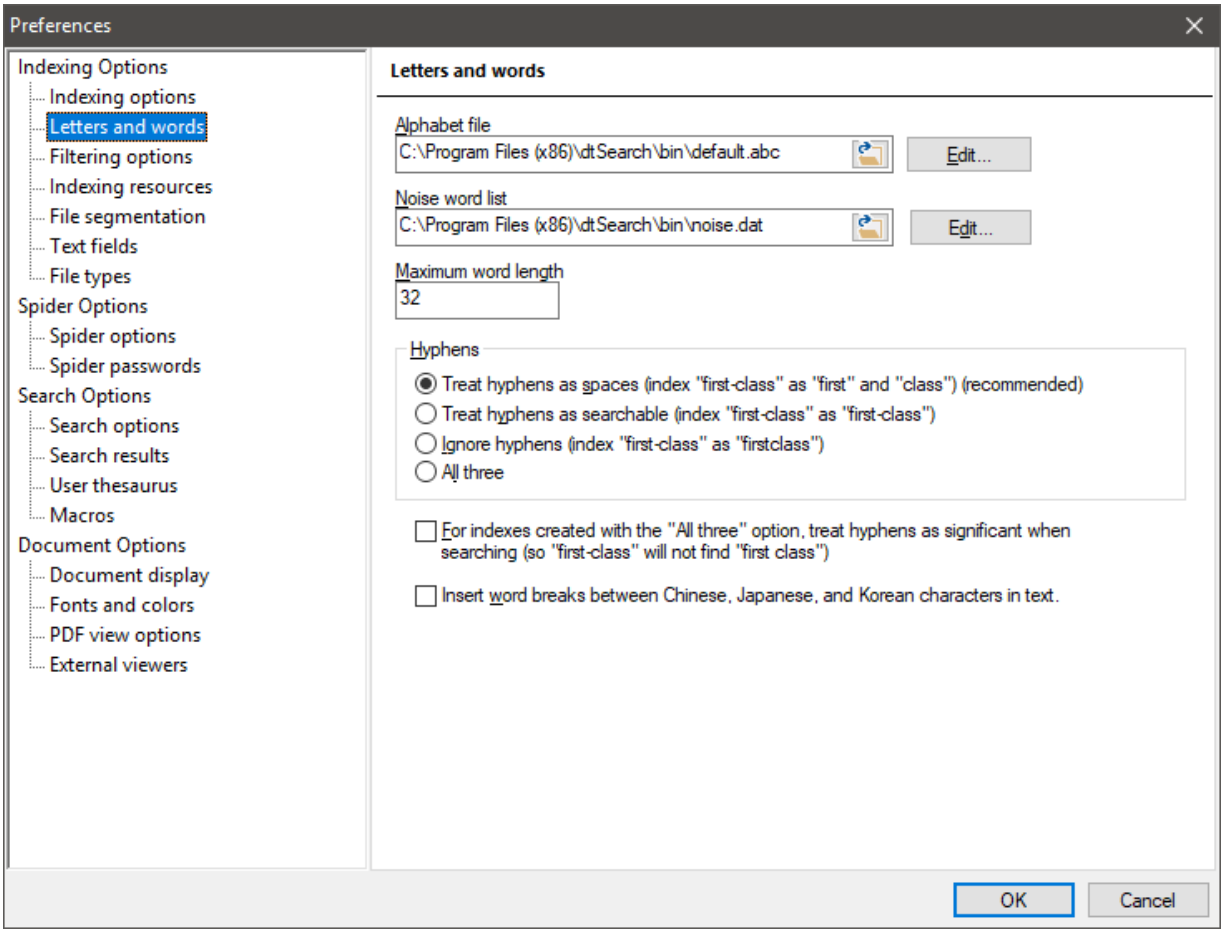
From the **Options** menu, choose **Preferences > Indexing Options.**
Select ‘**Generate and index MD5 hashes…**’. Other useful options for this type of search may be automatically recognising email addresses, index document properties to see who authored a file and index numbers in case sums of money etc are involved.



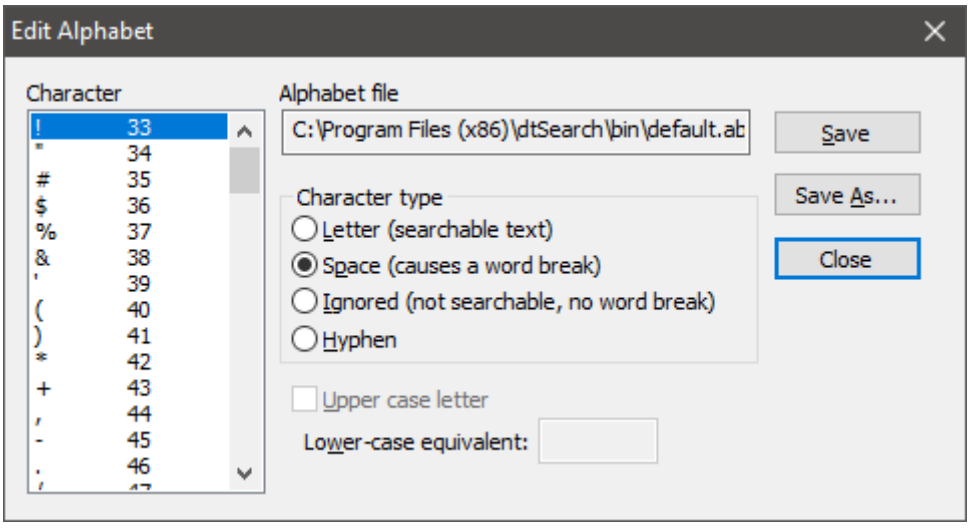
***TIP:** To use the keyboard instead of a mouse to navigate, use **Ctrl+Tab** or **Ctrl+Shift+Tab** to move down or back up in the left-hand panel. Use **Tab** or **Shift+Tab** to move down or up in the right-hand panel.*

Next choose **Letters and words.**

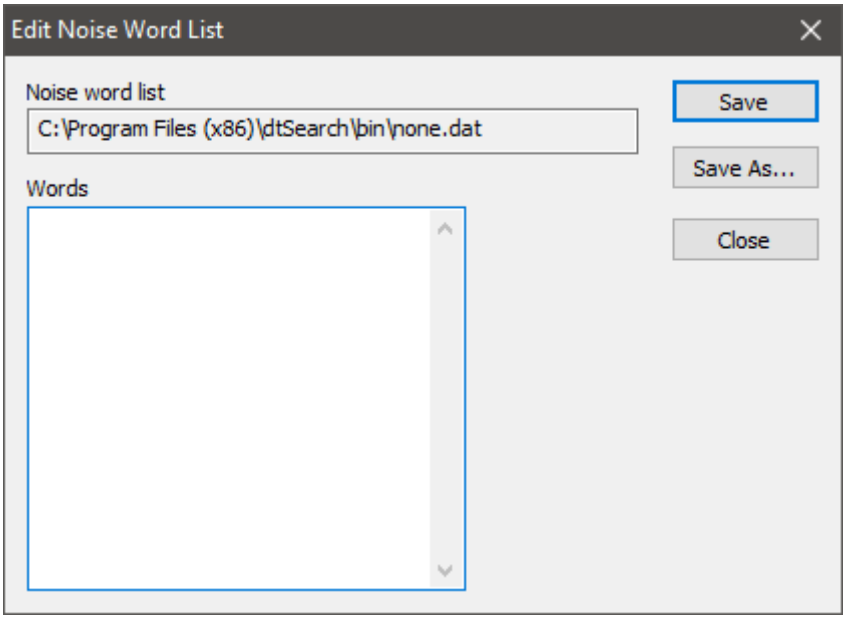
We need to make sure the Alphabet file has the factory default settings. Click on the **Alphabet file Edit…** button.



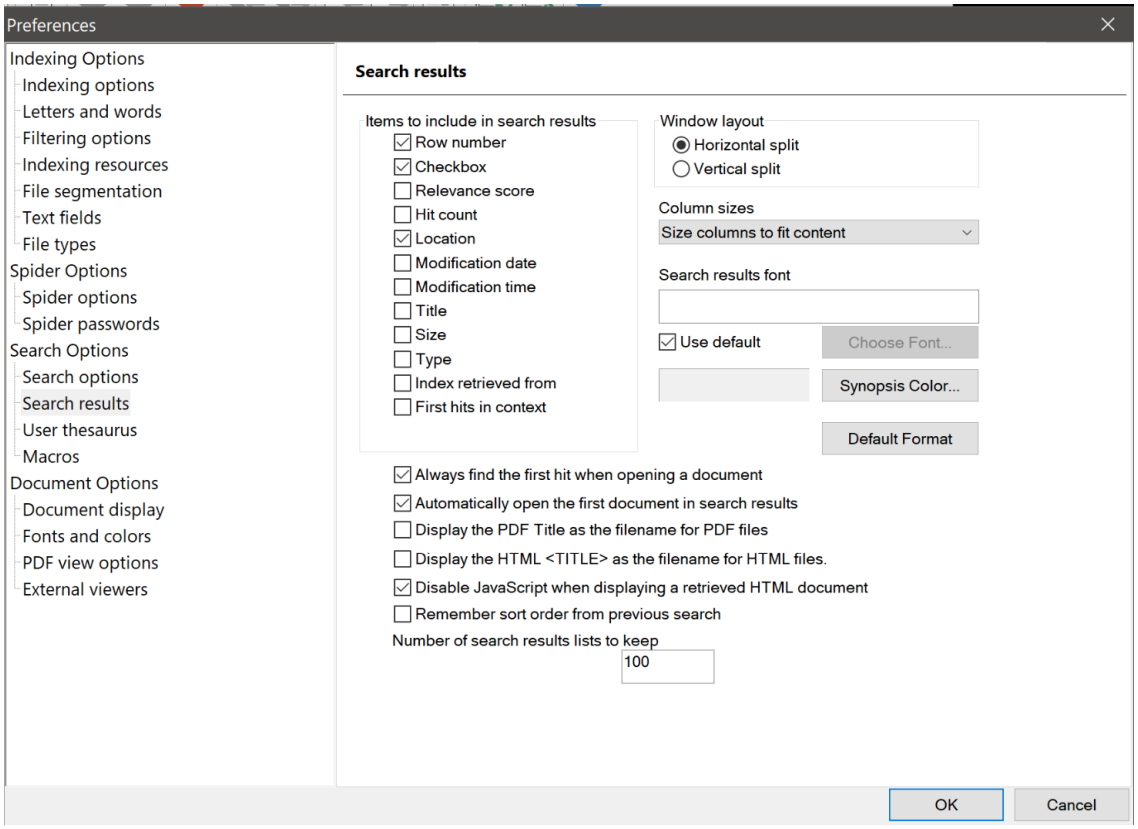
Make sure that all characters from 33 to 36 are set to **Space**. 37 to **Ignored,** 38 to 44 to **Space**, 45 to **Hyphen**, 46 and 47 to **Space**. If you make any changes press **Save** before closing the dialog.



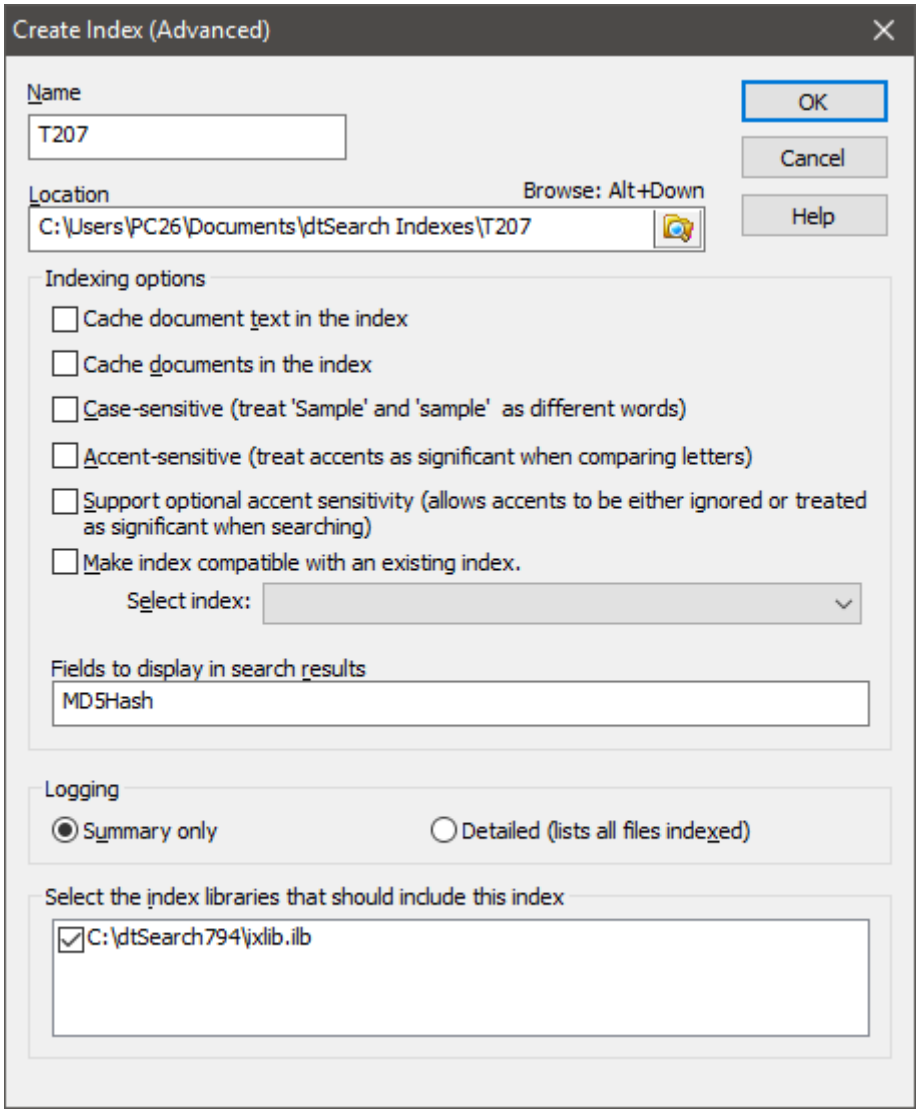
Click the **Edit…** button alongside the **Noise word list** textbox. For this session we need an empty noise word list. Create one by deleting all the words in the list, then press the **Save As...** button and save it with a file name of *none.dat*. Now **Close** the dialog.



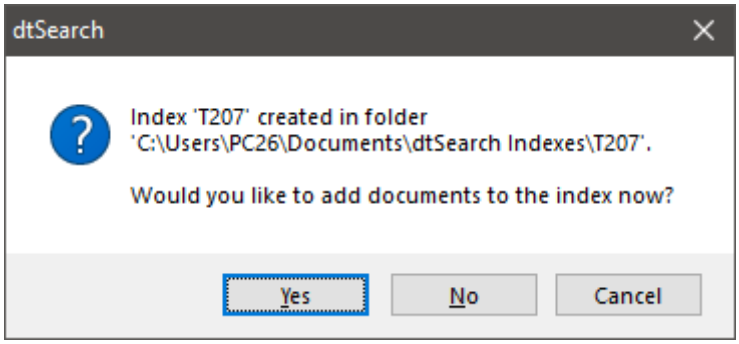
Finally set the **Search results** layout to include **Checkbox, Location** and **Row Number** as shown below and click OK.



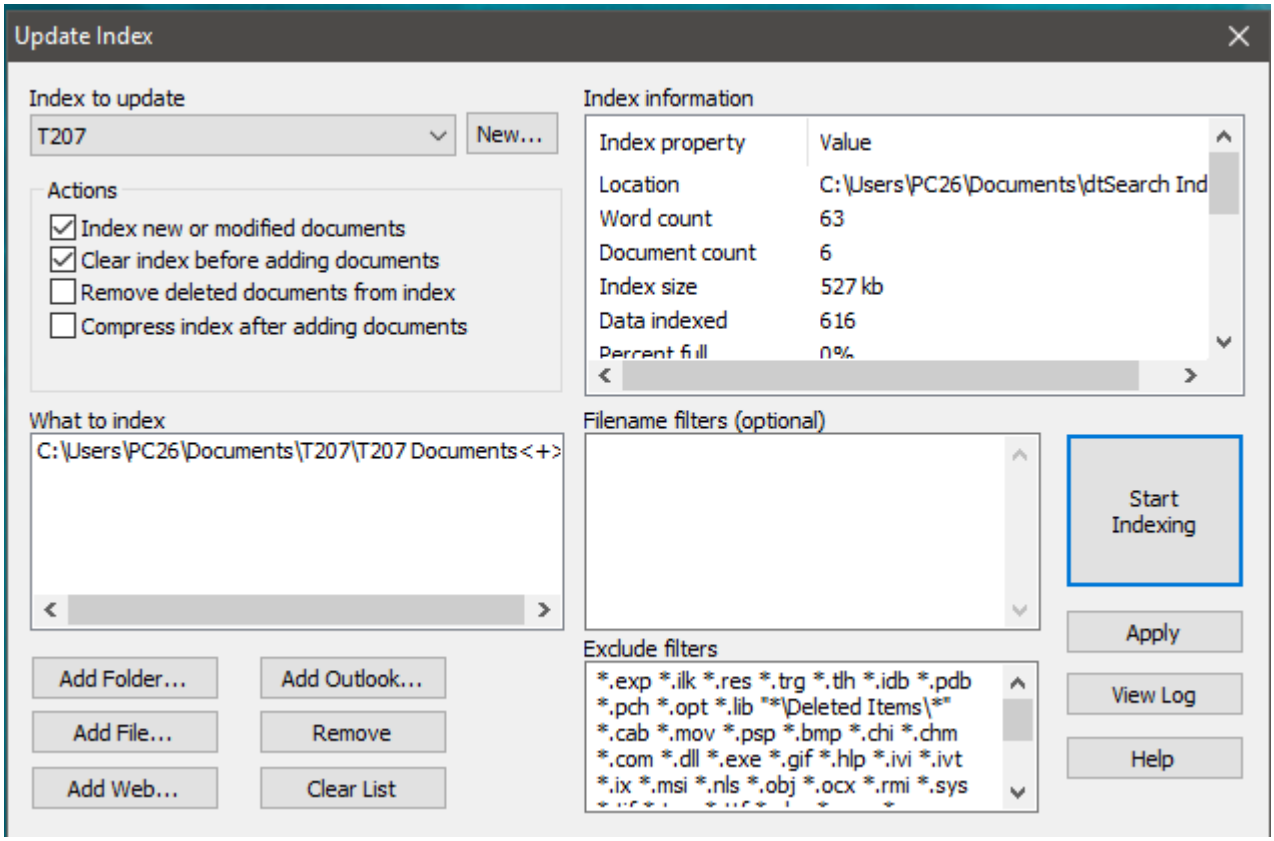
Now we are ready to create an index. From the **Index** menu select **Create index (Advanced) ...** Enter the name of the index as shown and enter *MD5Hash* exactly as shown into the *"Fields to display…"* text box and click **OK**.



Click **Yes** to add documents to your new index.



In the **Update Index** dialog that appears, press the **Add Folder...** button.
Browse to `Documents\T207\T207 Documents` (See Appendix) and click **OK**.
The **<+>** at the end of the folder path indicates that subfolders will be indexed; if it is not present, right click on the folder path and select **include subfolders**.

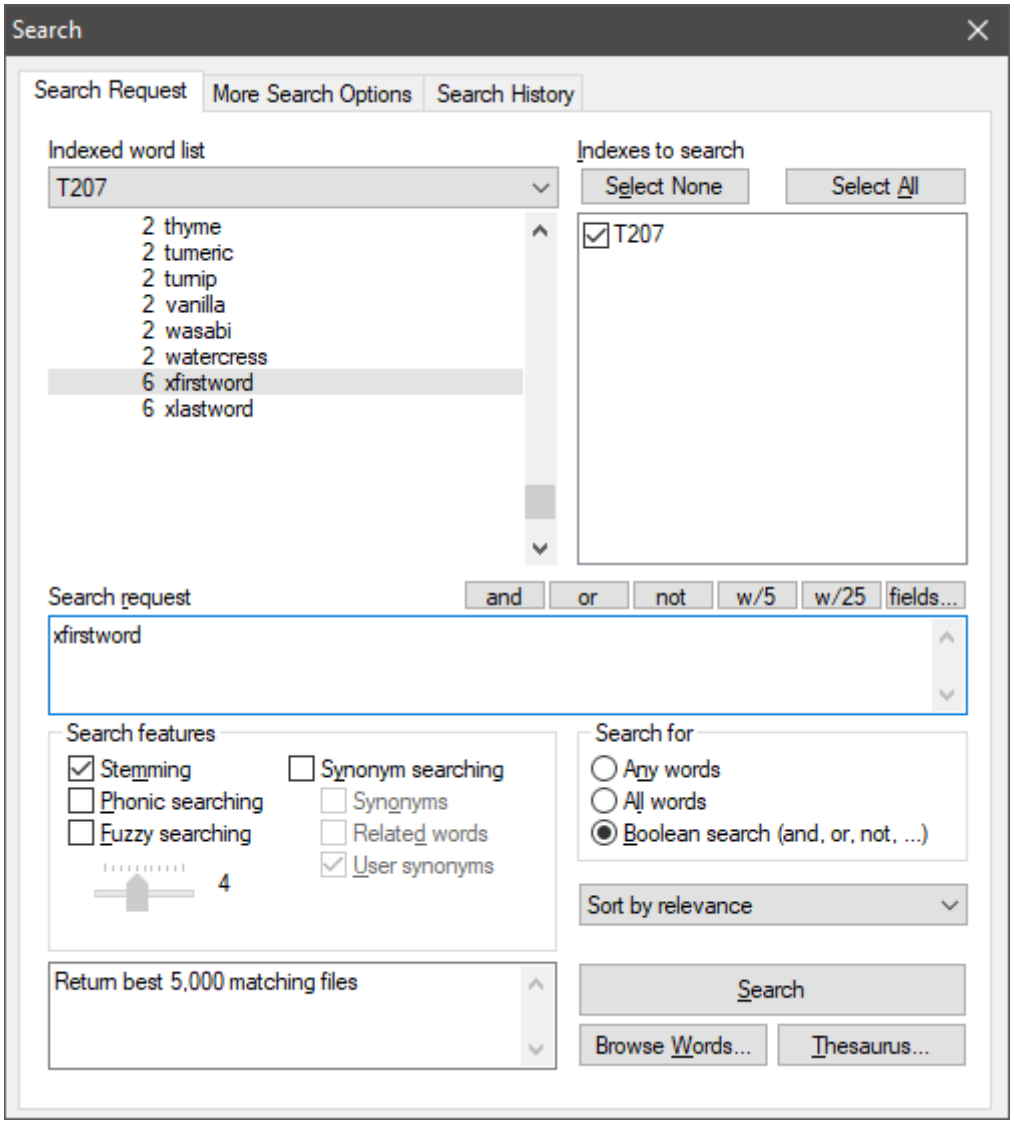


Click the **Start Indexing** button. When indexing is complete press the **Close** button.

We are now ready to start searching!

For a single keyword search, open the Search dialog. Press the **Select None** button to unselect any previously selected indexes, then select the T207 index. Under **Search features** select **Stemming**, and under **Search for** select **Boolean search**.

Enter a search query `xfirstword` and press **Enter**, this should return all the six documents.



The best way to search using a long Boolean query is to use the **Search for List of Words** function.

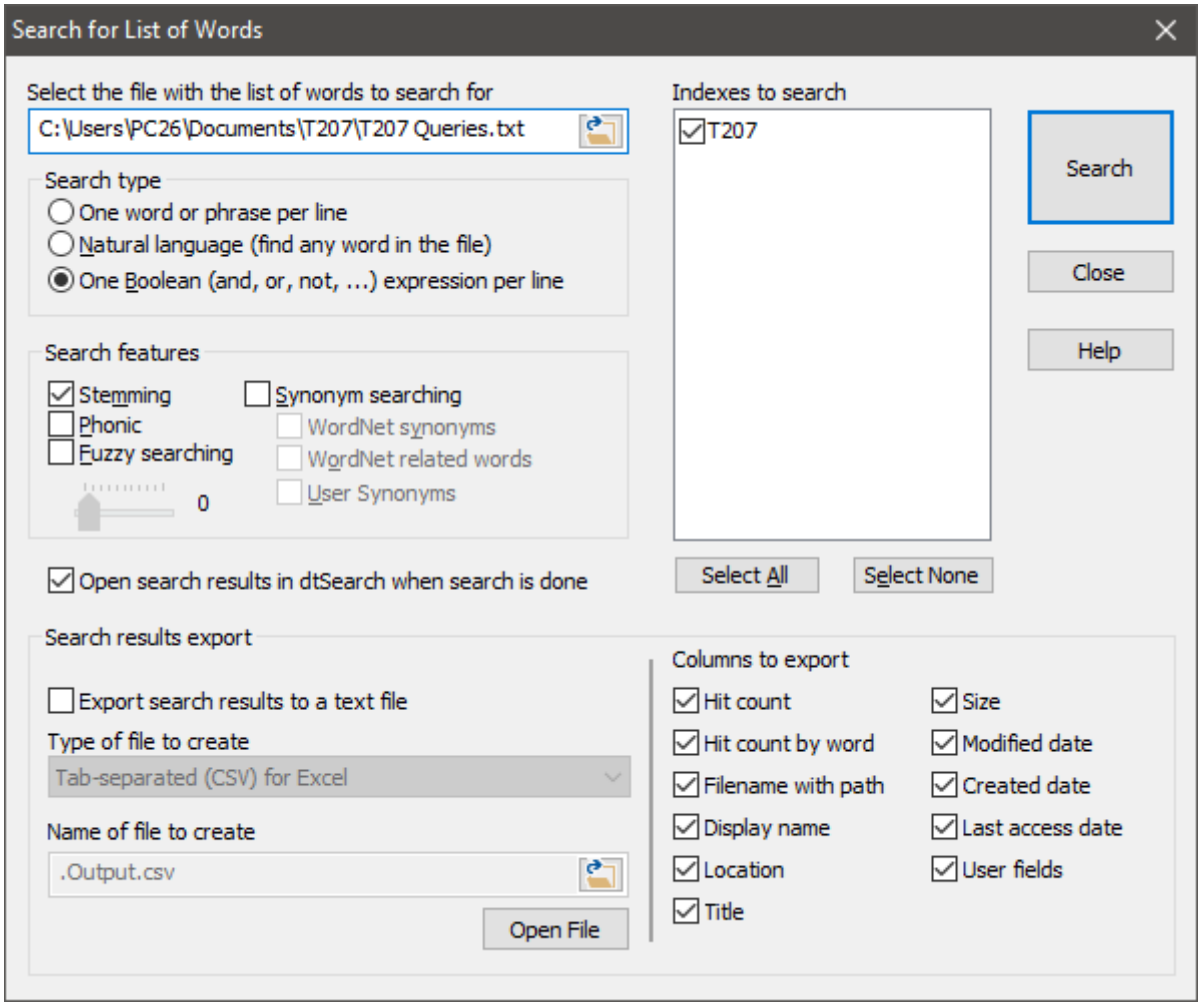
List all the queries in a text file, one query per line. For this exercise use the sample **T207 Queries.txt** file (see Appendix).

Select **Search > Search for List of Words...** menu (Ctrl+Shift+W), browse to `T207 Queries.txt`

Select **One Boolean expression per line**.

Select the **Stemming** and **Open search results in dtSearch when search is done** checkboxes.

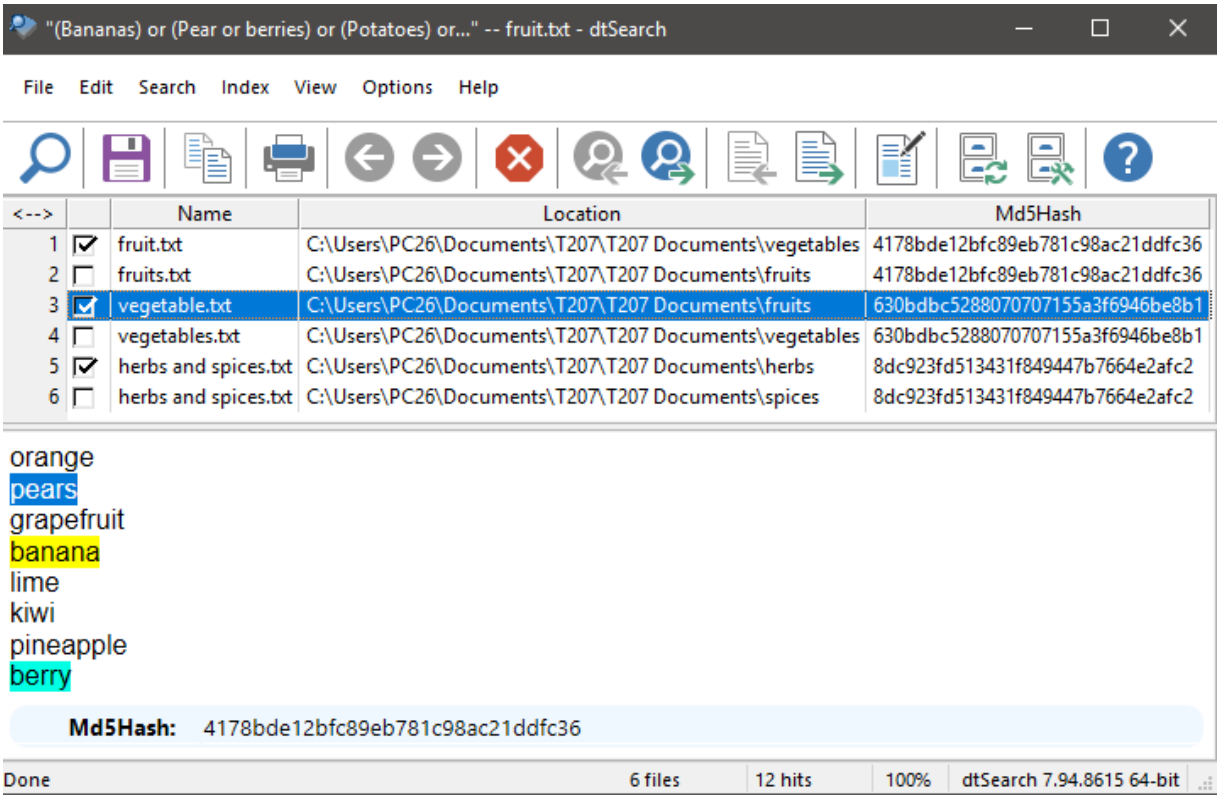
Select the **T207** index to search.



Now press the **Enter** key or click **Search**, and **Close** the dialog when the search has completed.

After the search, click on the **MD5Hash** column header. This will sort the results by the field **MD5Hash**, duplicate MD5 hashes will be displayed together.

The MD5 hash is based on the content of the file, duplicates are identified even if the files have been copied to another folder and had their filenames changed.

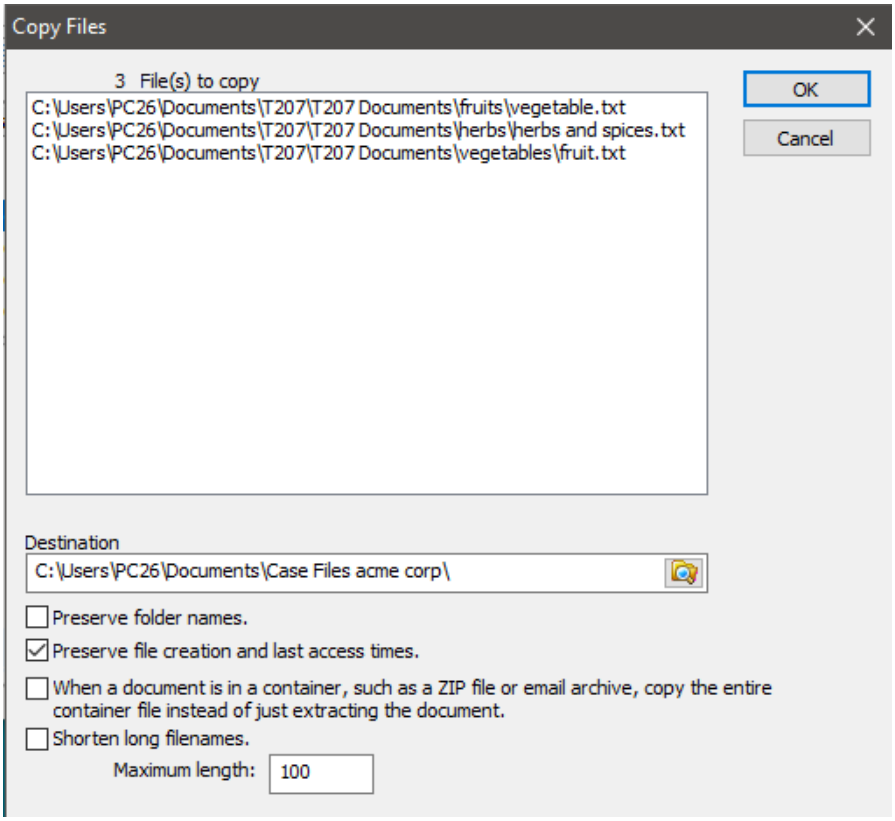


Click on the checkboxes (or **Tab**, **Up/Down arrow** to navigate to the row, then **Shift + Space**) to select the files to be copied into a separate folder.

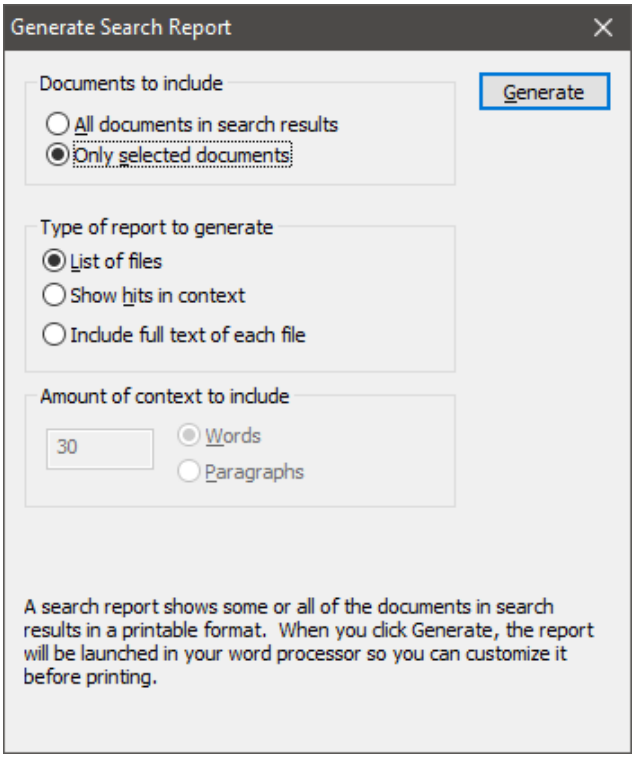
To select multiple rows to be copied, click on the first row you want to copy (or **Tab**, **Up/Down arrow** to navigate to the row) then press **Shift + Up/Down arrow** to select the rows and click (or **Shift+Space**) on the checkbox in the last row that you want to copy.

Select the **Edit > Copy File…** menu and choose the destination folder.
For investigative work there are options to preserve the folder names, file creation and last access times. There is also an option to shorten long filenames.

Finally, click **OK** to start the copy operation.



You can also generate a report from **File > Print Search Results…** with an option for **only selected**.



APPENDIX

T207 search query and test documents

Download T207.zip from this link: T207.zip

Unzip the file (right-click and select **Extract All...**) and put the extracted **T207** folder into the **Documents** folder on each student's PC.

Search for List of Words

These files are intended to illustrate the use of the **Search for List of Words** function and the effect of stemming on search results.

`T207 Queries.txt` contains a list of words or simple Boolean queries (Garlic and Peppers) on each line, and dtSearch Desktop expands this into a Boolean search expression:

```
(Bananas)or (Pear or berries) or(Potatoes)or(Garlic and Peppers)
```

This search query will return all six test documents, even though none of them contain the exact words Bananas, Pear, berries or Potatoes; this is because Stemming (and automatic conversion of words to lower case) ensures that a search for Potatoes will find potato, a search for Pear will find pears, a search for berries will find berry, and a search for Bananas will find banana.

Using the **Search for List of Words** function dtSearch will automatically add OR between each search term and will add parenthesis around search terms to avoid ambiguity.

The **Search for List of Words** function is less prone to error than entering a long Boolean query in the Search Dialog and ensures the same search query can be easily repeated by others.

For other tips see: https://www.dtsearch.com/images7/dtSearch-Tips-InsideCounsel.pdf

Screenshots

The screenshots used in the article are from dtSearch Desktop 7.94 running on Windows 10.
To make the title bars easier to distinguish the default white theme was changed. If you want to do this in Windows 10 go to **Settings>Personalization>Colours**, uncheck **Automatically pick an accent colour from my background**, check the **title bars** checkbox. Choose a custom colour such as 'storm'.

Accessibility

If you are running a training session for a group, it's important that all participants can see (and hear) projected screens or other material (e.g. PowerPoint slides) and those that need extra contrast or other assistive technologies are catered for. Changing the mouse pointer scheme to inverted extra-large and using the **Display pointer trails** option can be beneficial, these can be edited in Windows 10 from **Settings>Themes>Mouse pointer.**

For more information see: https://www.w3.org/WAI/teach-advocate/accessible-presentations/