

# dtSearch Desktop/Network Indexing and Search techniques

## **T206 – SEARCHING FOR NUMERIC PATTERNS**

dtSearch Desktop/Network is a powerful search tool used by professionals for a wide variety of tasks. This article aims to show you how to search for numeric patterns such as IPv4 or IPv6 addresses using wildcards or regular expressions in a one-time search or to find all such patterns in a collection of files.

### **Course Prerequisites**

dtSearch Desktop/Network 7.68 or later

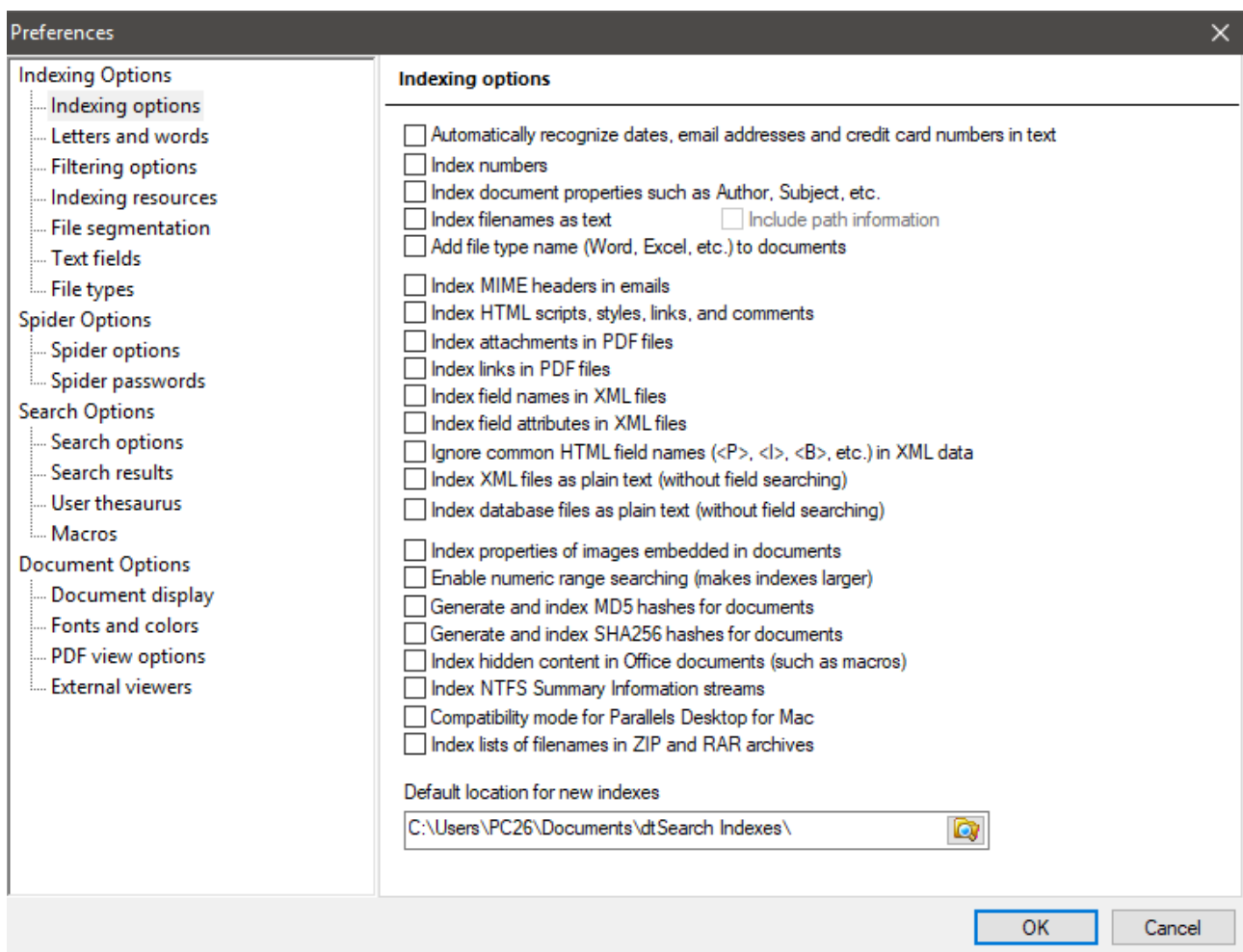
User Thesaurus Plus 1.2 (Optional – Includes IPv4 & IPv6 macros)

This training session covers advanced topics of interest to those that need to find documents containing numeric patterns such as IP addresses (IPv4 or IPv6). After completion you should be able to make searches to find a number or all instances of a numeric pattern using a macro

Before beginning the training session, all copies of dtSearch Desktop must have the same initial indexing and display setup. Access to the **T206** 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**

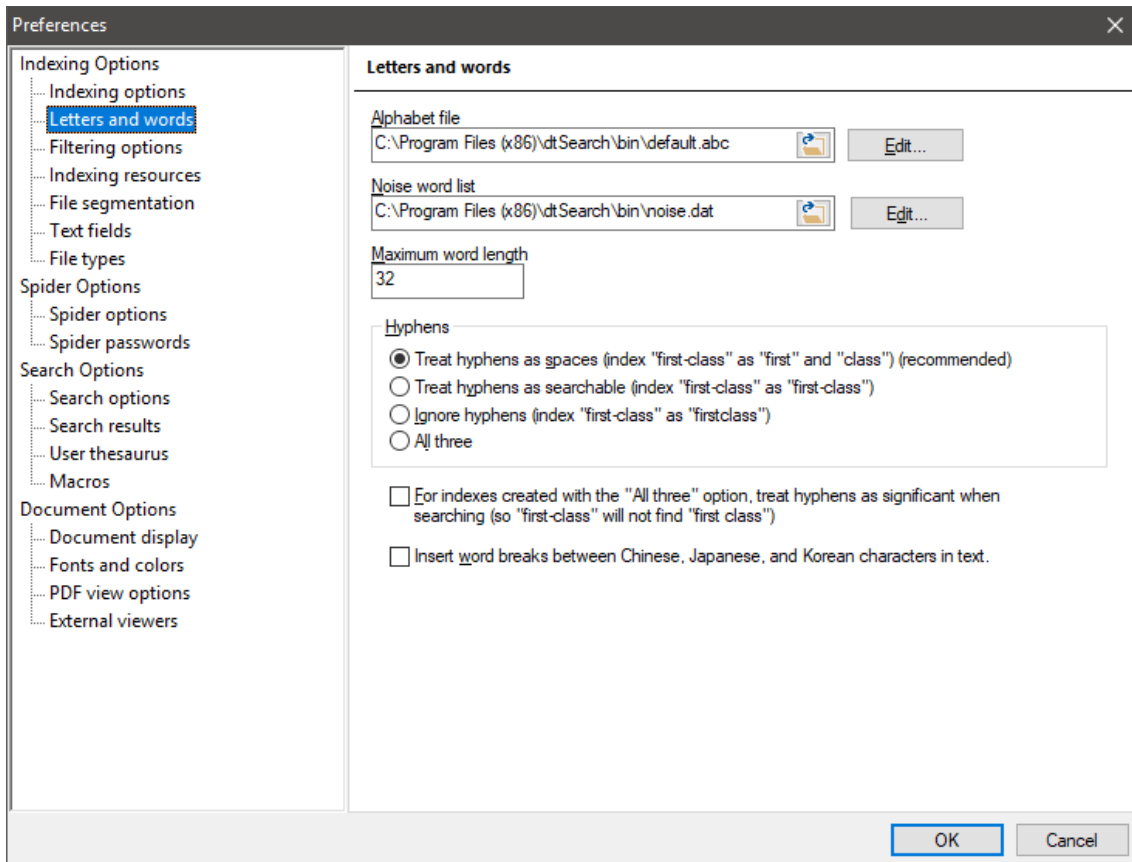


Make sure that all checkboxes are **not** selected.

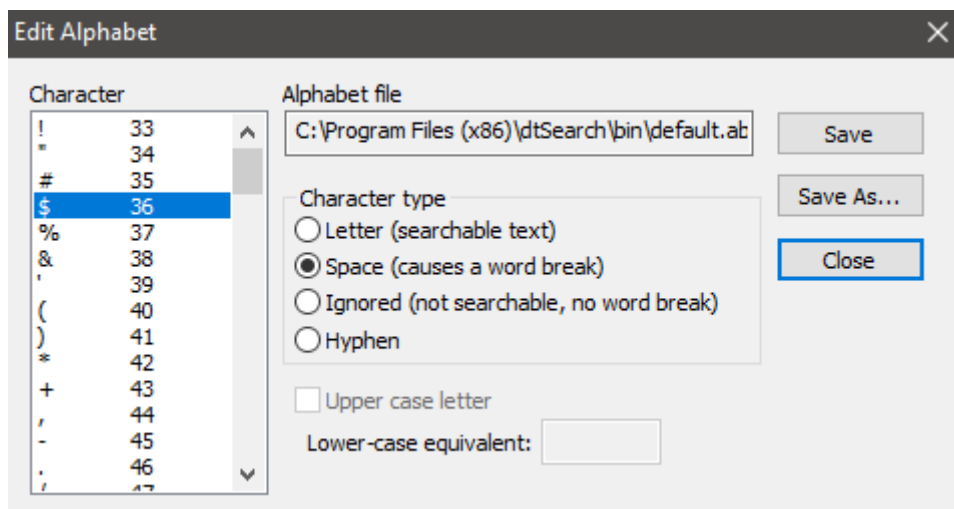
**TIP:** To use the keyboard instead of a mouse to navigate, use **Ctrl+Tab** or **Ctrl+Shift+Tab** to move down or back up in the left-hand panel. Use **Tab** or **Shift+Tab** to move down or up in the right-hand panel.

Next choose **Letters and words**

Click on the **Alphabet file Edit...** button.



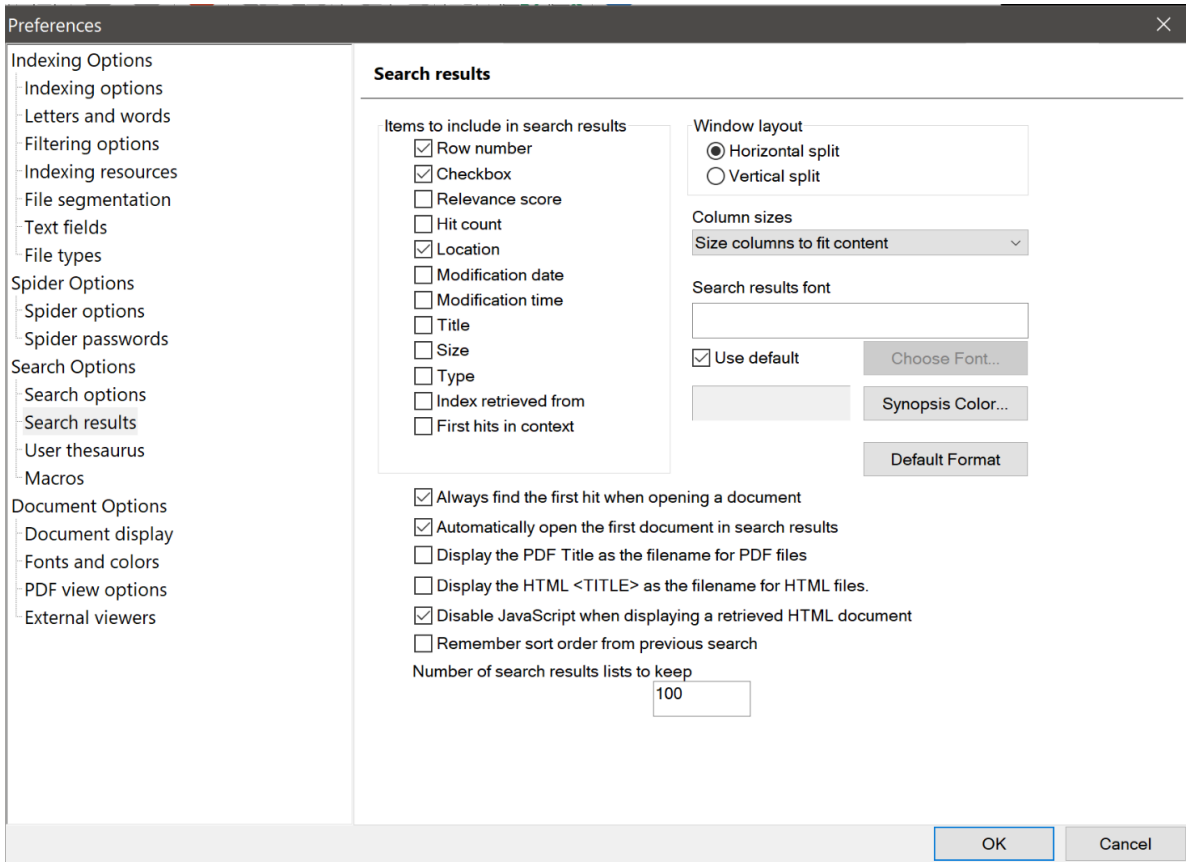
Make sure that all characters from 33 to 36 are set to **Space**. 37 to **Ignored**, 38 to 44 to **Space**, 45 to **Hyphen**, 46 and 47 to **Space**. If you make any changes press **Save** before closing the dialog.



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.

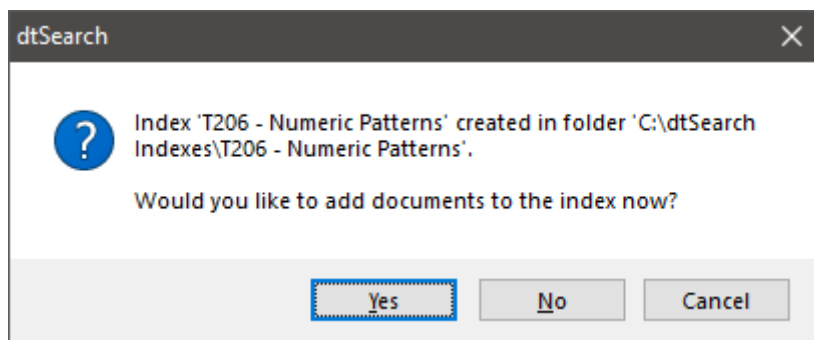
## T206 – SEARCHING FOR NUMERIC PATTERNS

Finally set the **Search results** layout to the settings shown below:



Now we are ready to create an Index. From the **Index** menu select **Create index...**  
Enter the name of the index as `T206 - Numeric Patterns` and click **OK**.

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



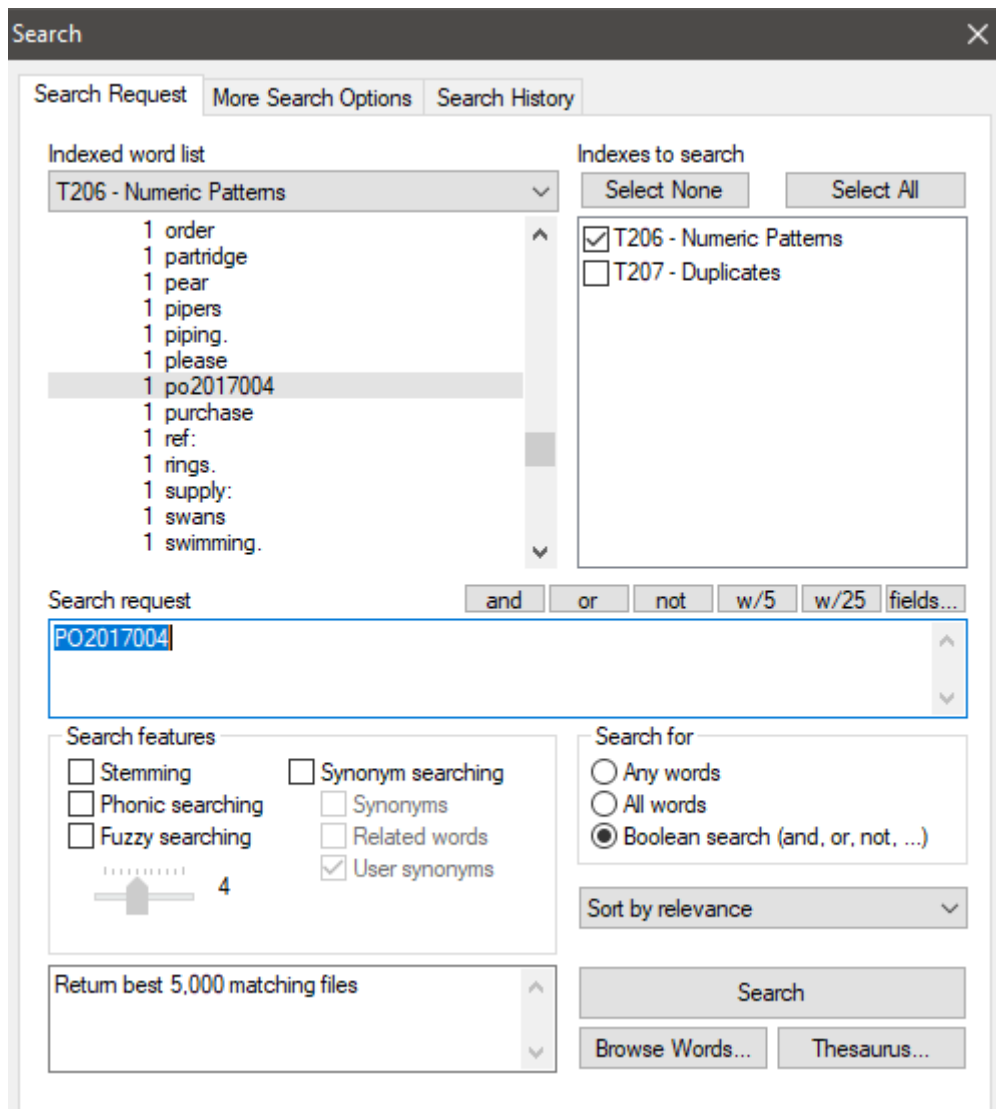
In the **Update Index** dialog that appears, press the **Add Folder...** button and browse to the **T206** test document folder (see **Appendix**).

The **Update Index** dialog will now re-appear. Click the **Start Indexing** button. When indexing is complete (less than 1 second) press the **Close** button.

**We are now ready to start searching!**

## T206 – SEARCHING FOR NUMERIC PATTERNS

For a single keyword search, open the Search dialog. Press the **Select None** button to unselect any previously selected indexes then select the T206 index. Make sure no **Search features** are enabled and that Boolean search is selected.



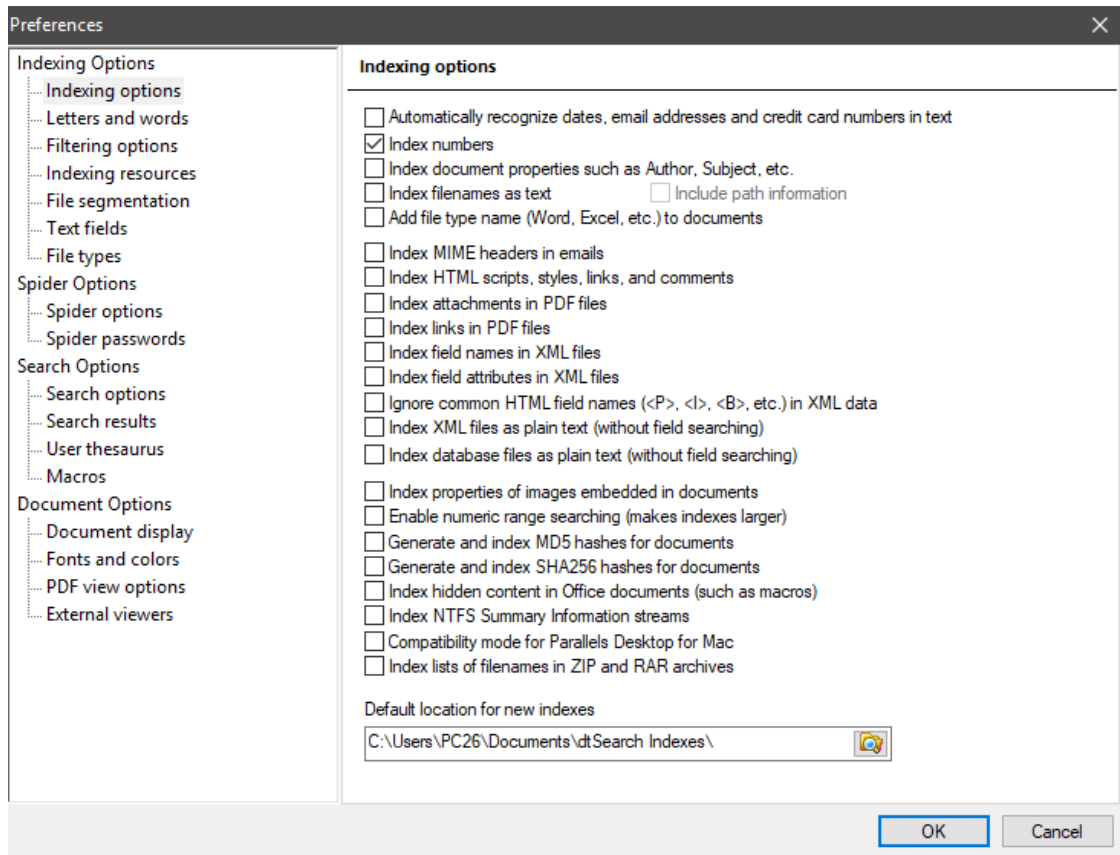
Search for **PO2017004** and the Example Purchase Order.txt should be retrieved. Now search for the telephone number **0208-590-8888**. You should get a 'no files retrieved' message, even though we know **0208-590-8888** exists in Example Purchase Order.txt.

dtSearch does not index currency signs or numbers by default. This is to make indexes smaller and reduce indexing time, however, a 'number' is any 'word' beginning with a digit. Because of this behaviour, **PO20170004** and similar strings that begin with a letter will be found, even though they contain mostly digits.

To find a term beginning with a digit, we need to edit the indexing options.

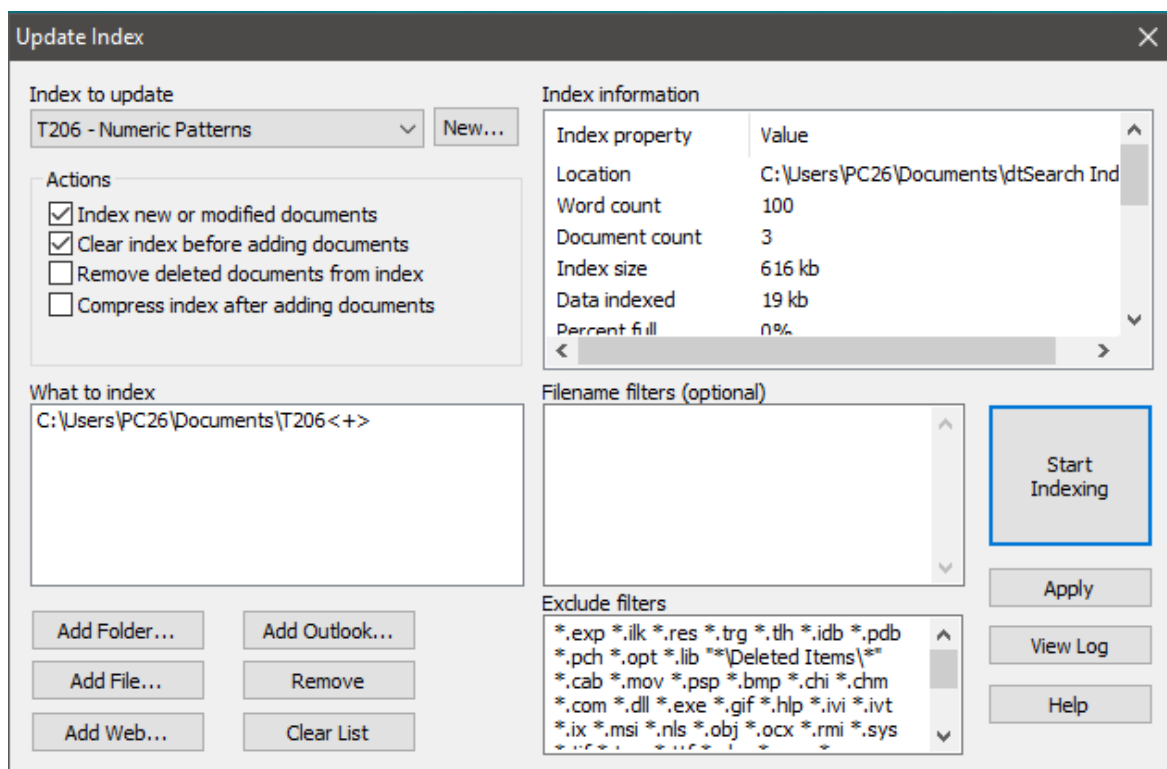
## T206 – SEARCHING FOR NUMERIC PATTERNS

Select **Options > Preferences > Indexing options**, check the **Index numbers** checkbox and click **OK**



This option will not affect an index that has already been built.

To see the benefit of this option, rebuild the index by selecting **Index > Update Index...** and enable the two indexing Actions as shown below.



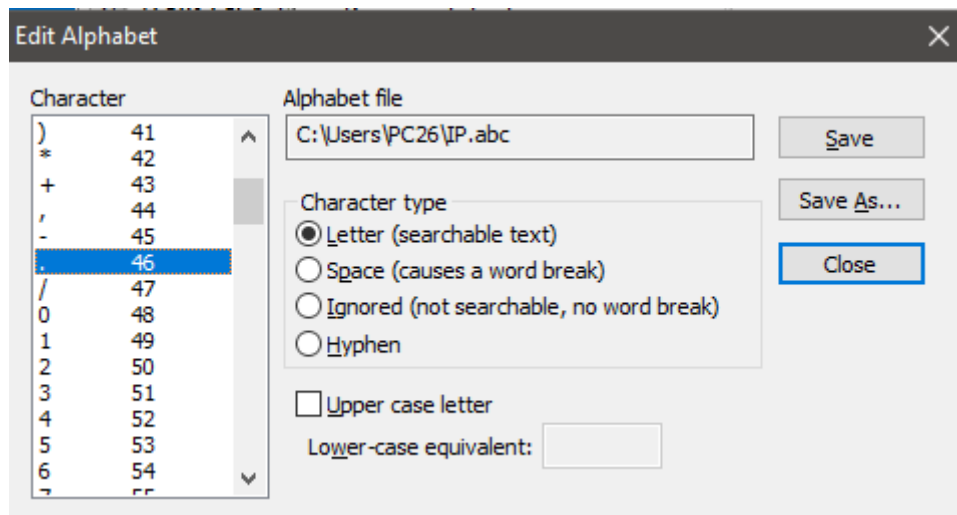
Search for **0208-590-8888** again. Example Purchase Order.txt should now be returned. Although this is a simple example, it highlights the importance of choosing relevant indexing options.

Most search software use **?** and **\*** as wildcards to represent a single character and multiple characters respectively. dtSearch has an additional single digit wildcard **=** that is useful for increasing the precision of your searches. **PO20170004** can be returned with a search query of **PO2017=====** or **PO=====** instead of the less precise **PO\*** or **PO????????** which could return **POPULATION**.

Another default behaviour of dtSearch is to treat commas and full stops (period) as a space so that words with trailing punctuation are indexed correctly, this has the undesired side-effect of causing an IPv4 address such as **198.51.100.255** to be indexed as four separate numbers.

To modify this behaviour, select **Options > Preferences > Letters and words** and click **Edit...** next to the Alphabet file textbox.

Find character 46 and select **Letter** as shown. This modified alphabet file will severely affect standard searching and should only be used for specialised numeric searches. Use **Save As...** and save this modified file as IP.abc. Now press **Close**.



Rebuild the index by selecting **Index > Update Index...** and enable the two indexing Actions as before. You will now be able to search for **198.51.100.255** as one term, which should return IPv4.txt.

Searching for an IPv6 address such as **2001:0db8:85a3:0000:0000:8a2e:0370:7334**, requires further changes. Edit the IP.abc alphabet file to make character 58 a letter and **Save** it.

You will also need to increase the maximum word length to 40 characters to accommodate the length of an IPv6 address, select **Options > Preferences > Letters and words** and change the default value in the **Maximum word length** field from 32 to 40. Click **OK** to save and close the dialog.

These changes alone will not be enough though, character **58 (:)** is a reserved character in dtSearch, to give terms a variable weighting in Boolean searches. To make this character searchable we need to modify the Windows Registry <http://support.dtsearch.com/faq/dts0131.htm>

CAUTION: The next section is intended for advanced users, administrators, and IT Professionals.

*Serious problems might occur if you modify the registry incorrectly, follow the steps carefully. For extra protection, you may want to back up the registry first so that you can restore the registry if a problem occurs.*

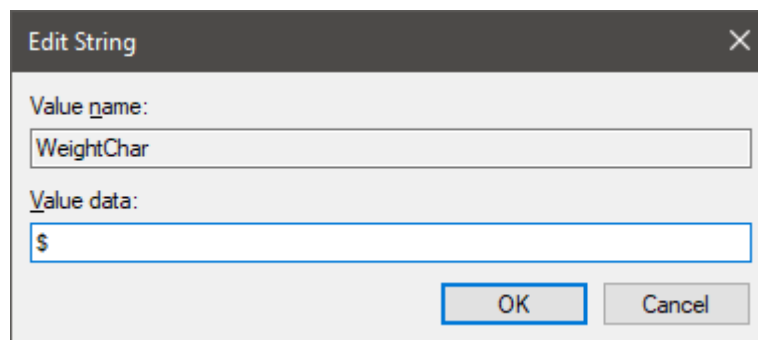
*How to back up and restore the registry in Windows 10 (or 8):*

(See <https://support.microsoft.com/en-us/help/322756>)

1. From the **Start** menu, type **regedit.exe** in the search box, and press **Enter**.  
(You may be prompted for an administrator password or for confirmation).
2. In Registry Editor, locate and click  
**[HKEY\_CURRENT\_USER\Software\dtSearch Corp.\dtSearch\Settings]**.
3. Click **File > Export**.
4. In the **Export Registry File** dialog box, select where to save the backup, then type a name in the **File name** field.
5. Click **Save**.

Exit dtSearch and run RegEdit (See (1) in text box above for Windows 8 to 10) or for Windows 7: Use the keyboard combination Windows key + R or right-click the Start Menu to open Run. In Run, enter “regedit” (without quotes) then Click **OK**.

Open the key **[HKEY\_CURRENT\_USER\Software\dtSearch Corp.\dtSearch\Settings]** to access your dtSearch registry entries. Find WeightChar, **Right Click > Modify...** and edit the **Value data:** field from : to \$



Click OK and close the Registry Editor.

After making these changes, open dtSearch and rebuild the index, make sure that the alphabet file is IP.abc. After the rebuild you will be able to search for a full IPv6 address such as **2001:0db8:85a3:0000:0000:8a2e:0370:7334**.



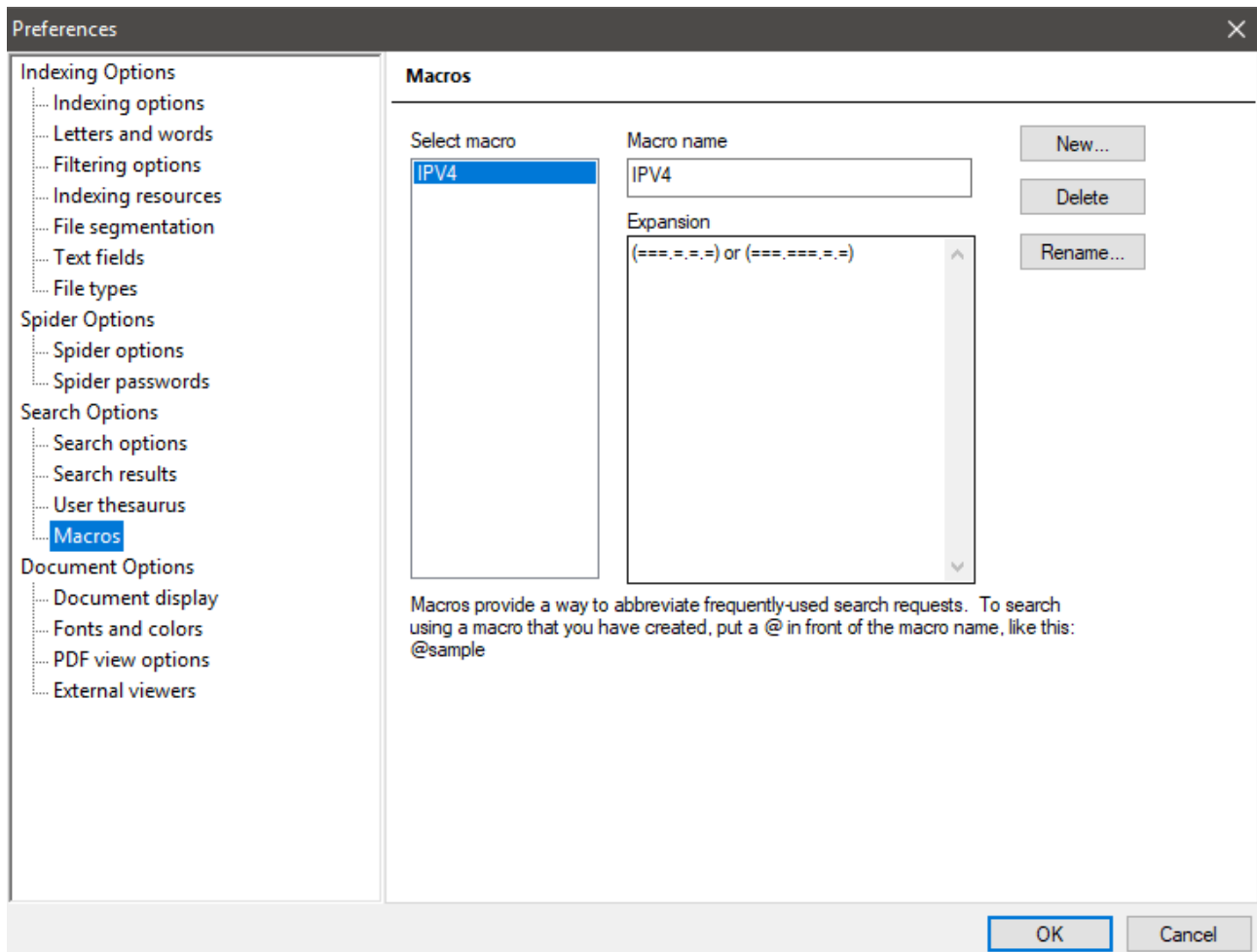
## Using Macros and Regular Expressions

If you need to find all the IPv4 or IPv6 addresses in a collection of files, you could use the simple wildcard approach, but you would need to allow for all the different formats of addresses. For example, just to find IPv4 addresses 127.0.0.1 and 192.155.0.1 would require a search term of `(===.=.= OR ===.===.=.)`.

### Macros

To avoid having to type in a long search query each time it is best to save the search query as a macro.

Select **Options > Preferences > Search Options > Macros** click **New...** and enter a name for your macro IPV4, enter the search query `(===.=.= OR ===.===.=.)` in the expansion text box, and press OK.



### Regular Expressions

To allow for all possible IP address formats would require a much longer search query.

An alternative to using a long wildcard search query is to use a Regular Expression.

See [https://support.dtsearch.com/webhelp/dtsearch/regular\\_.htm](https://support.dtsearch.com/webhelp/dtsearch/regular_.htm) for details of Regular Expression syntax in dtSearch.

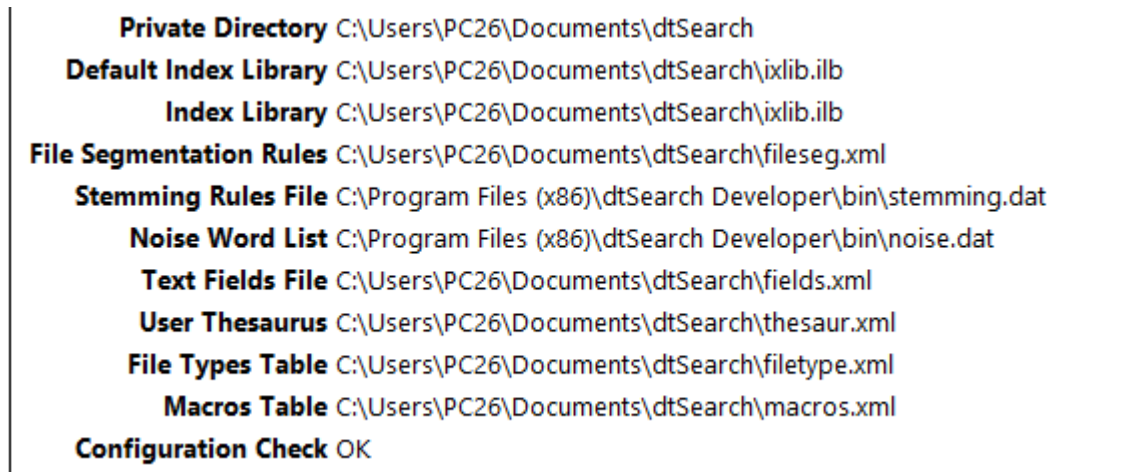
If you are not familiar with Regular Expressions, you may find it easier to use the example Macros supplied with **User Thesaurus Plus** available from <https://www.dtsearch.co.uk>

## User Thesaurus Plus sample macros

User Thesaurus Plus 1.2 is supplied with IPv4, IPv6 and Bitcoin address sample macro files.

## Initial set-up for use with dtSearch Desktop or Network

Open dtSearch Desktop, select **Help > About dtSearch...** and scroll to the **dtSearch Configuration** section.



Select the path displayed after **User Thesaurus** and **Copy** it (right-click > copy or Ctrl+C).

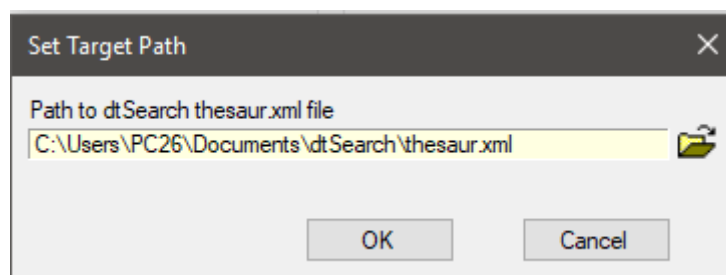
If you have edited the macro file in dtSearch Desktop you can import it into User Thesaurus Plus so that you can more easily edit it in the future. Click on **File > Import file** and select the file.

You will be prompted to rename the file for identification in User Thesaurus Plus. Press OK to copy the file into the User Thesaurus Plus data folder.

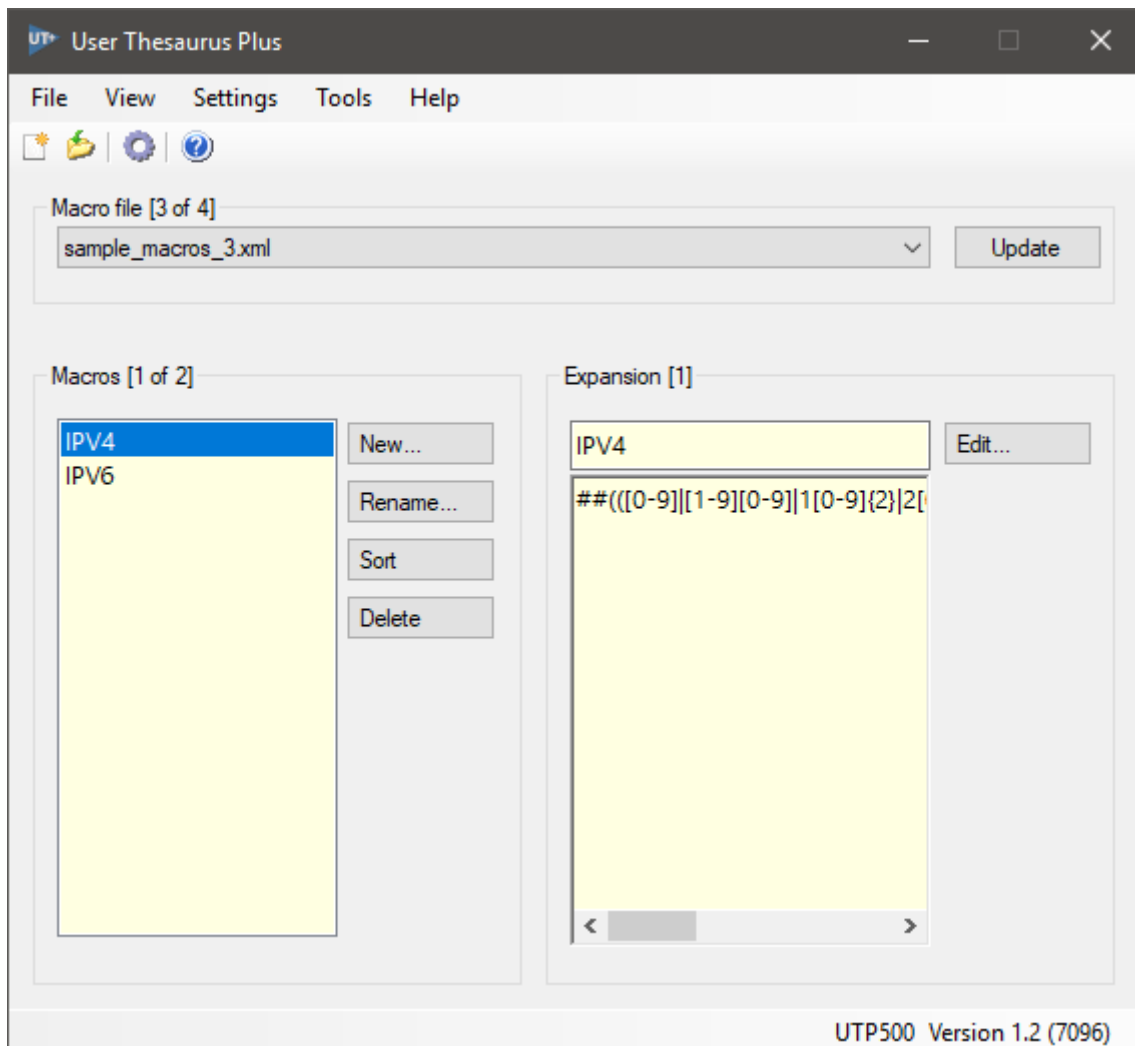
Open User Thesaurus Plus, go to **Settings > Target path**, and **paste** in the file path.

You will be prompted to update the Target file. **Warning:** Selecting **Yes** will **overwrite** your current thesaur.xml file.

\*NOTE: You may need to change user permissions so that you have read/write access to the macro and user thesaurus files. On Windows 7 it is advisable to save the file in the User/AppData folder or Documents folder.



Select **View** and choose **Macro**. Select the sample\_macros\_3.xml file from the drop-down list and click **Update** to make the changes effective in dtSearch. dtSearch does not have to be closed for the change to take effect.



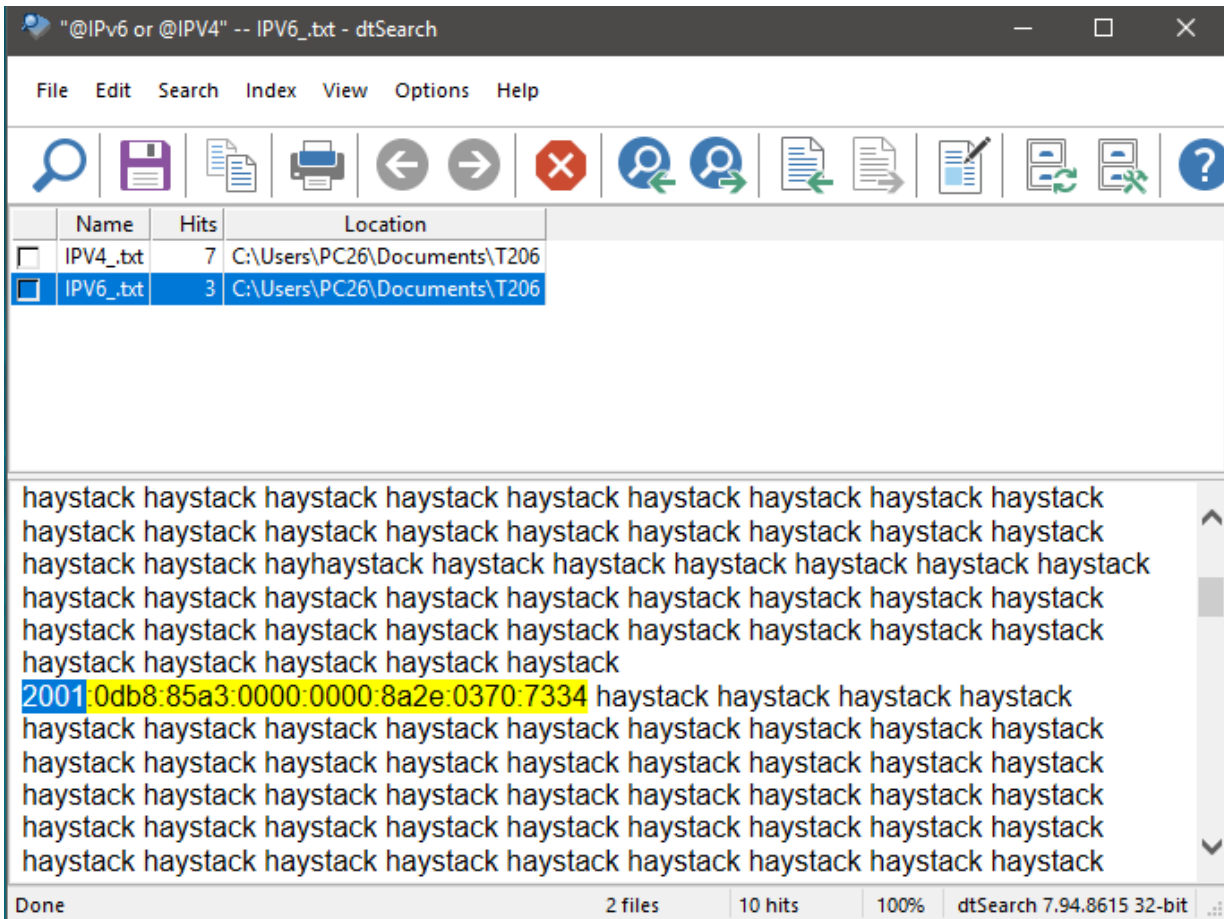
When not in use, User Thesaurus Plus can be closed or minimised to the taskbar. All changes are saved automatically.

User Thesaurus Plus lets you create multiple thesaurus and macro files so that you can choose from many smaller files for highly specific search purposes. User Thesaurus Plus also has enhanced editing features - sort synonym groups; copy and paste from the web and let the built-in filtering remove invalid characters and formatting; stop-word highlighting and removal; merge files

User Thesaurus Plus also includes an alphabet file editor for all currency signs.

**Searching for IP addresses using the sample macros**

In dtSearch Desktop open the search dialog, select the T206 index and enter @IPv4 to search for all IPv4 addresses, enter @IPv6 to search for all IPv6 addresses, or enter @IPv4 or @IPv6 to find both types. The search results should show seven IPv4 addresses hit-highlighted or three IPv6 addresses hit-highlighted.



## APPENDIX

### T206 test documents

Download T206.zip from this link: [T206.zip](#)

Unzip the file (right-click and select **Extract All...**). Put the extracted T206 folder into the **Documents** folder on each student's PC.

The IPV4\_.txt and IPV6\_.txt files contain several IP addresses mixed in with regular text.

The Example Purchase Order.txt file is meant to demonstrate how typical trade documents such as invoices, purchase orders, commercial invoices or credit notes can be found by a reference number which typically start with one or more letters can be found even if indexing of numbers is turned off to save indexing space. (<http://tfig.unece.org/contents/trade-documents.htm>)

### NOTES:

**dtSearch Desktop v7.85 to v7.94:** ensure that in **Options|Preferences|Fonts and Colors** the **multiple colors** hit-highlighting option is **not** selected, it may prevent hit-highlighting with some macros.

**User Thesaurus Plus** assumes that the macro and thesaurus files are in the same folder. User thesaurus Plus 1.1 includes an IPv4 macro only, version 1.2 includes IPv4, IPv6 and Bitcoin address macros. (The macro.xml and thesaur.xml files must allow read/write access).

**Macros** for Regular Expressions in dtSearch need to start with **##**

### 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/>

In dtSearch Desktop the font size in the Search Dialog can be increased from the **Options|Search Dialog Font Size...** menu. The Search Results can also be increased in size by holding down the **Ctrl** button and scrolling the mouse. The Zoom % is displayed in the status bar alongside the dtSearch version number.