dtSearch Desktop/Network
Indexing and Search techniques

T201 - SEARCHING FOR MONEY

dtSearch Desktop/Network is a powerful search tool used by professionals for a wide variety of tasks, this short course aims to show you how to use the User Thesaurus Plus Add-on product to improve the precision and recall of some tasks which often prove tricky even for experienced search professionals.

Course Requisites
dtSearch Desktop/Network 7.68 or later
User Thesaurus Plus 1.1
Internet access

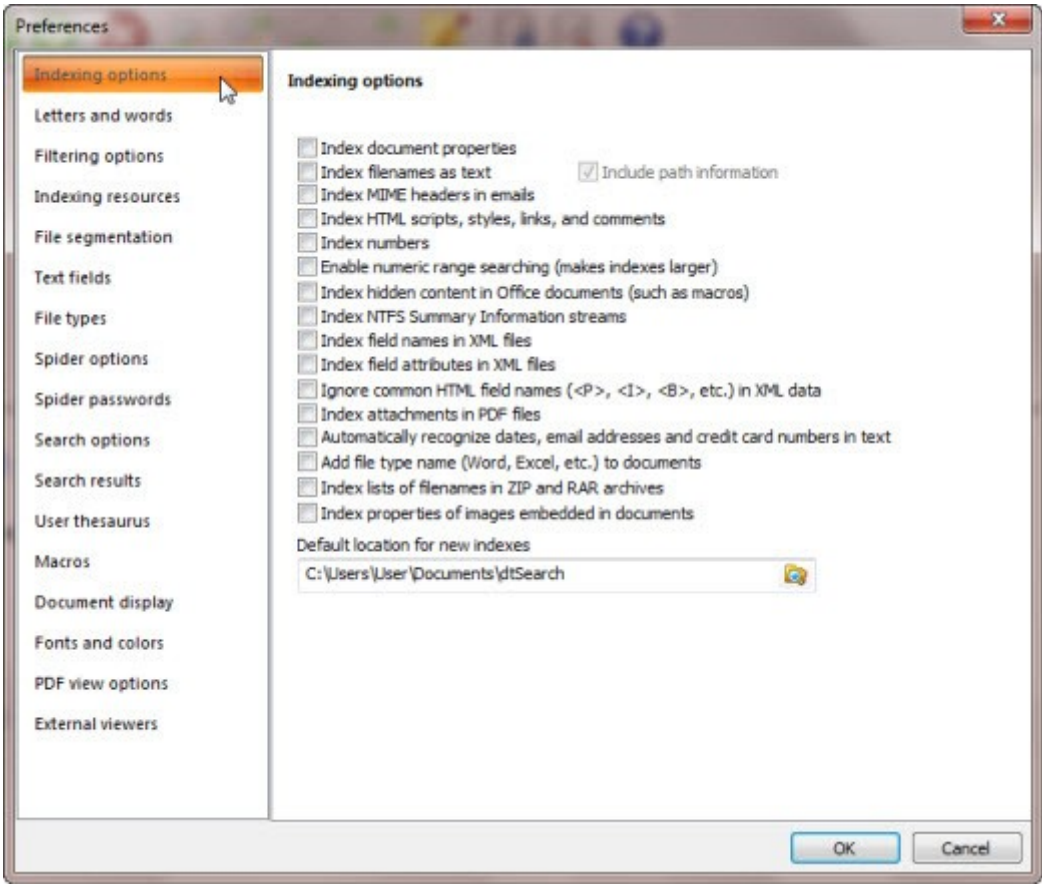
T201 - SEARCHING FOR MONEY

This training course covers several advanced topics of interest to those that need to find documents containing references to sums of money; after completion you should be able to make searches to find all instances of a sum of money in a specific currency.

Initial set-up of dtSearch Desktop:

From the **Options** menu, choose **Preferences > Indexing Options**

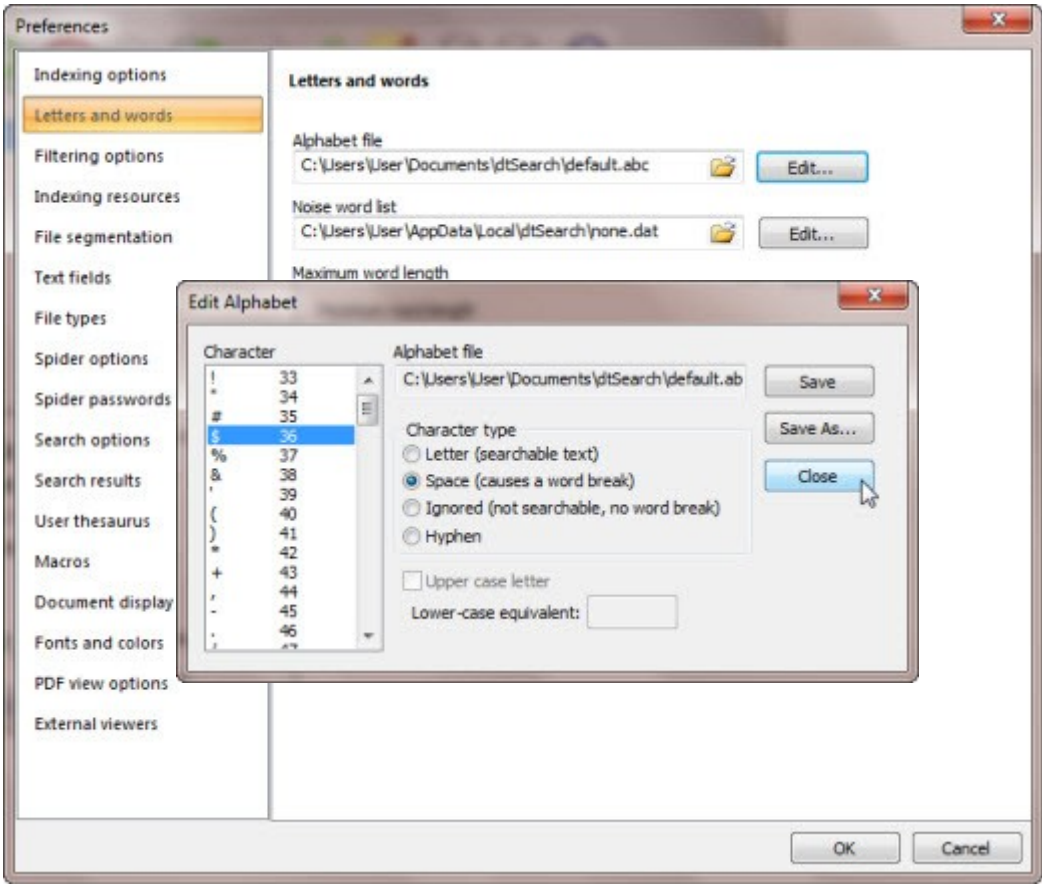
Make sure all check-boxes are **not** selected.



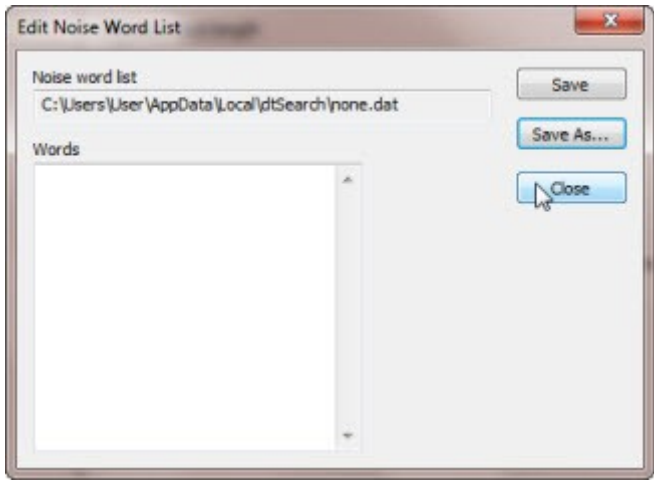
TIP. *To use the keyboard instead of a mouse to navigate, use* ***Ctrl+Tab*** *or* ***Ctrl+Shift+Tab*** *to move down or back up in the left hand panel. Use the Tab key or Shift+Tab to move down or up in the right hand panel.*

Next choose **Letters and words**

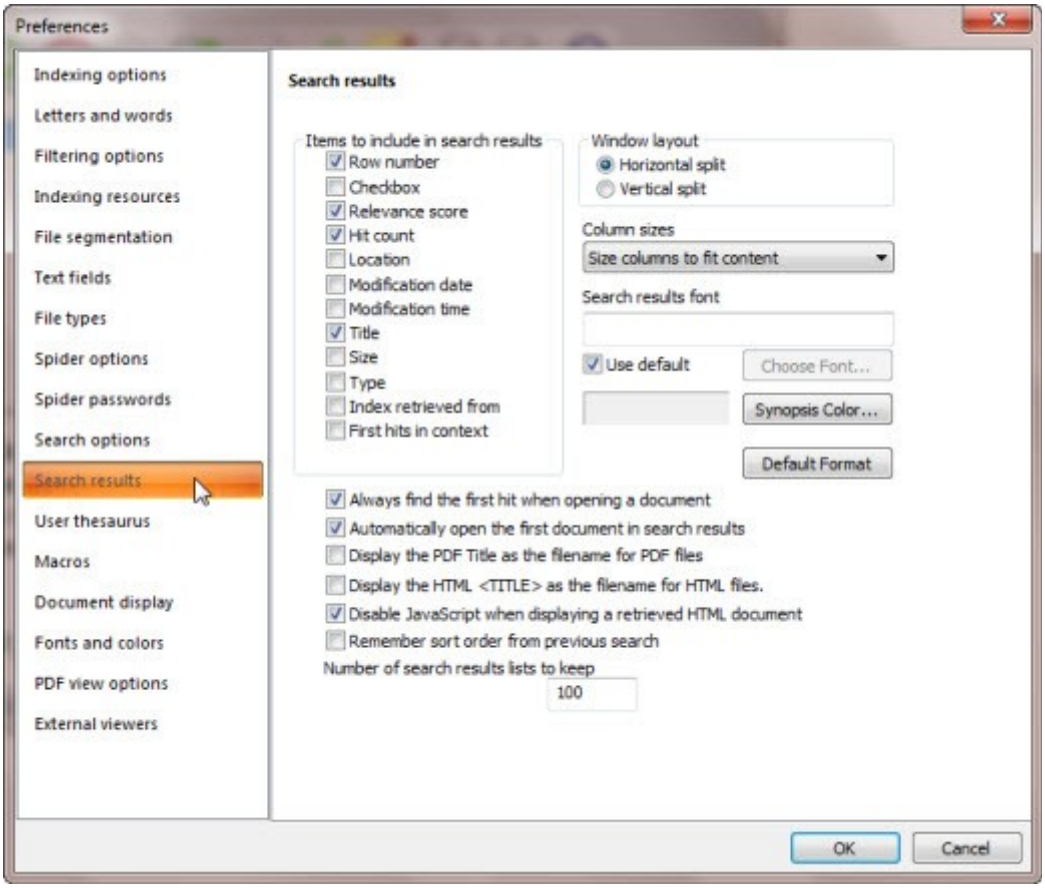
We need to make sure the Alphabet file has the factory default settings, click on the Alphabet file Edit button. Make sure that the $ sign (ASCII 36 decimal) and all other characters from 33 to 47 are set to **Space.** If you make any changes click on the **Save button** before closing the dialog.



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



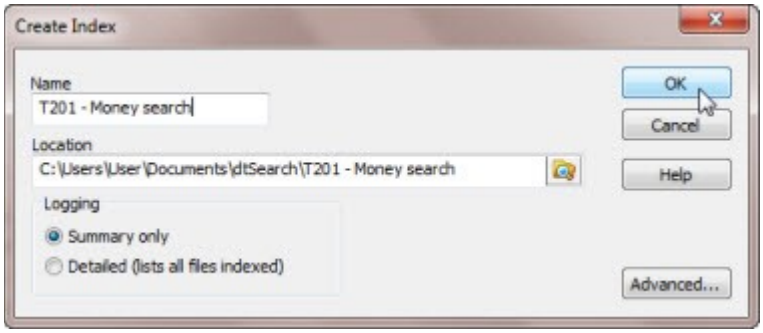
Set the **Search results** check-boxes to the basic settings as shown below:

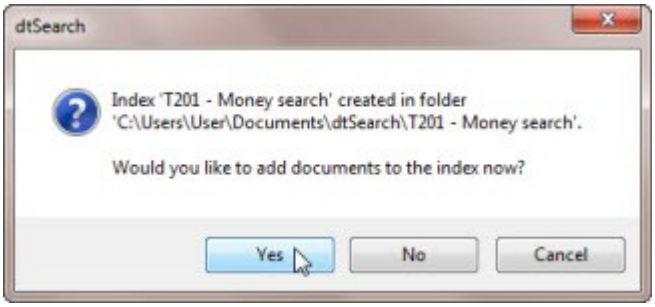


Finally click on the **OK** button to save your settings.

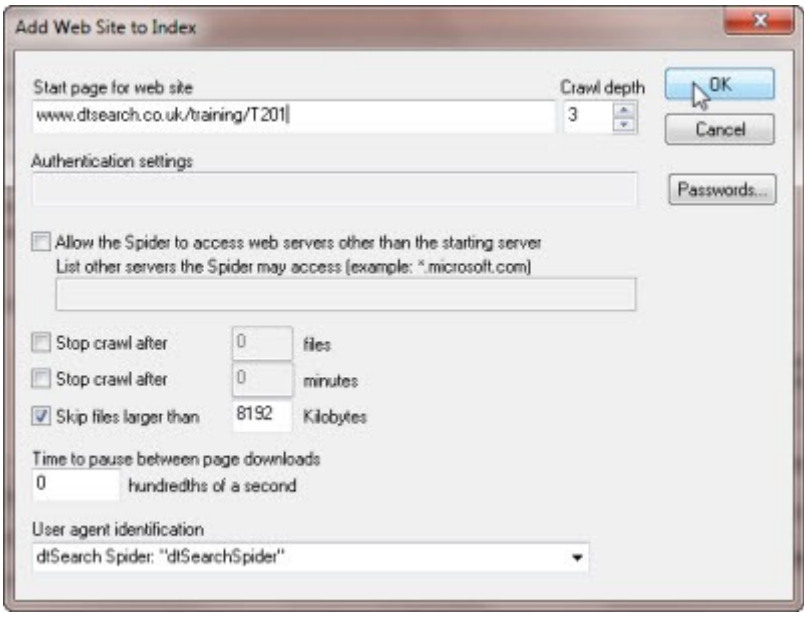
Now we are ready to create a basic index. From the **Index** menu select **Create index...**

Enter a name for the index as shown below:





In the **Update Index** dialog that appears, press the **Add Web...** button, now enter the web site address **`www.dtsearch.co.uk/training/T201/index.htm`** as shown below and press OK.



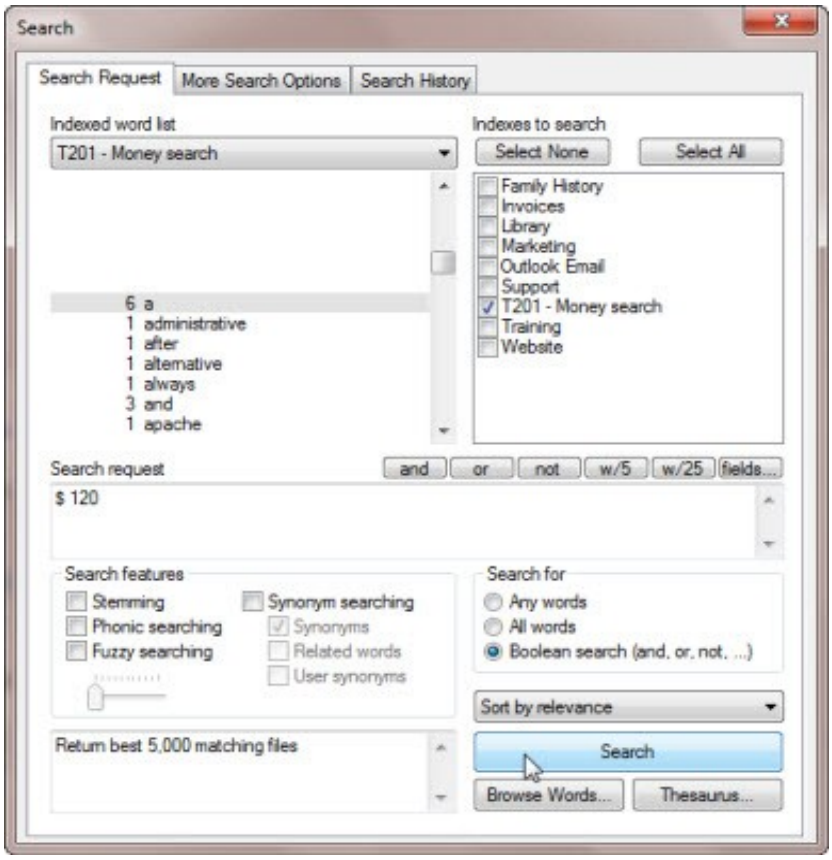
The **Update Index** dialog will re-appear, press the **Start Indexing** button, when the indexing is complete (under 30 seconds on a broadband Internet connection) click on the **Close** button.

We are now ready to start searching!

In dtSearch Deskop click-on the Search icon or press Ctrl+S to open the Search dialog.

Press the **Select None** button to unselect any previously selected indexes then select the *T201 - Money search* index. Make sure no **Search features** check-boxes are selected and that **Boolean search** is selected.

Take a moment to browse the **Indexed word list**, notice it contains no numbers or currency signs. Now enter a search for `$ 120` and press the **Search** button.



You should get a 'no files retrieved' message, click on OK.

dtSearch Desktop/Network does not index currency signs or numbers by default, this is to make indexes smaller and reduce indexing time. So the first thing we need to do is to get some numbers in the index! But before we do, select the **Synonym searching** and **Synonyms** check-boxes and then **Search** again.

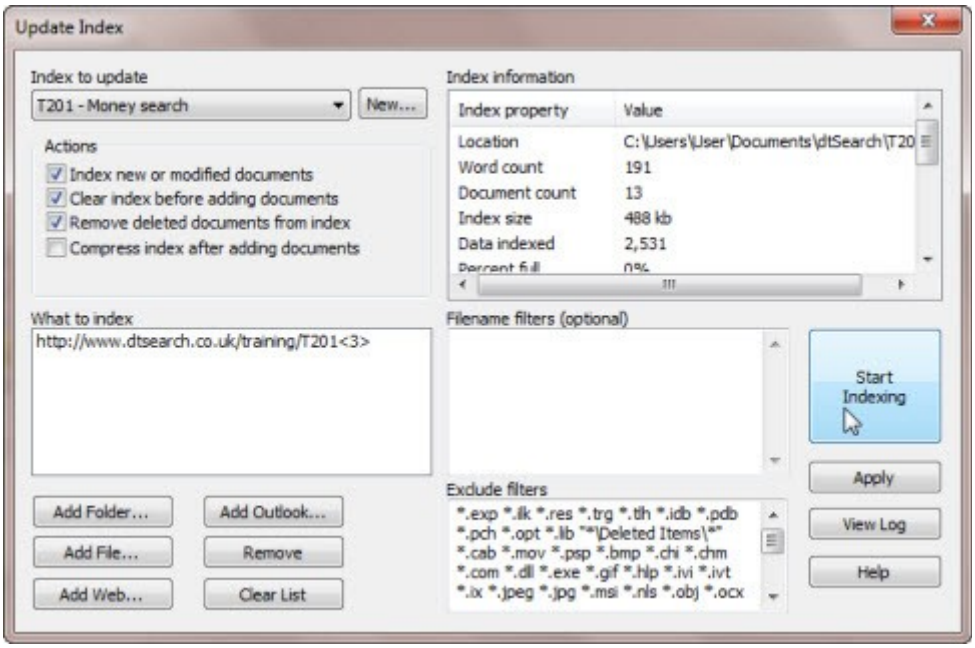
The result may surprise you given that the word list doesn't contain numbers or currency signs!

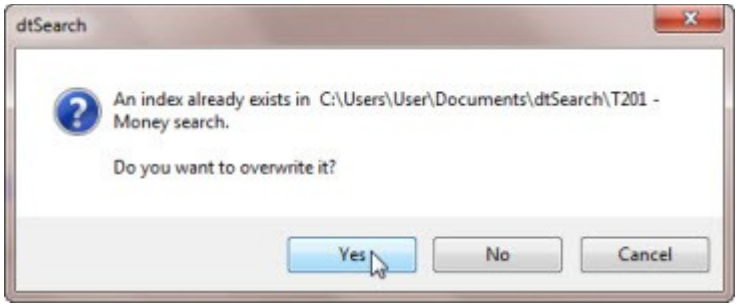
To find out why, open the **Search** dialog again and click on the **Thesaurus...** button, the **Browse Thesaurus** dialog that appears allows you to inspect the built-in Word Net Thesaurus. Enter `120` and click-on the **Lookup...** button. Click on the items that appear in the **Words** list-box to see the synonyms and related words. Close the dialog. The built-in

Word Net thesaurus is very powerful and can often help you find information by using synonyms or related words that you might not have been aware of!

Now from the **Options|Preferences** menu, select **Indexing options** and select the **Index numbers** check box, then click on the OK button.

From the **Index** menu, select **Update Index..**. and rebuild the index as shown below:





Close the **dtSearch Indexer** dialog and repeat the search.

You should now find that a search for $ 120 will find documents containing $ 120 because it will ignore the $ sign; unfortunately this means it may also find documents containing £120 or 120 € or 120 Yen or even 120 chickens! If you are searching a small document collection this may be acceptable but for many this lack of *precision* will be frustrating.

$$\text{Precision} = \frac{\text{Number of retrieved relevant documents}}{\text{Total number of documents retrieved}}$$

A search that returns very specific results is one with high precision, while a search that returns broad results is one with high recall.

$$\text{Recall} = \frac{\text{Number of retrieved relevant documents}}{\text{All relevant documents in a collection}}$$

Strictly speaking there is only one document - money_2.txt - that exactly matches the

search query, but it is generally accepted that when measuring precision and recall the judgement of one or more human experts should be used to judge relevancy, in this case three documents money_2.txt, money_1.txt ('$120') and money_3.txt ('120 dollars') are judged as highly relevant, Precision in this case is just 3/12 and Recall 3/3.

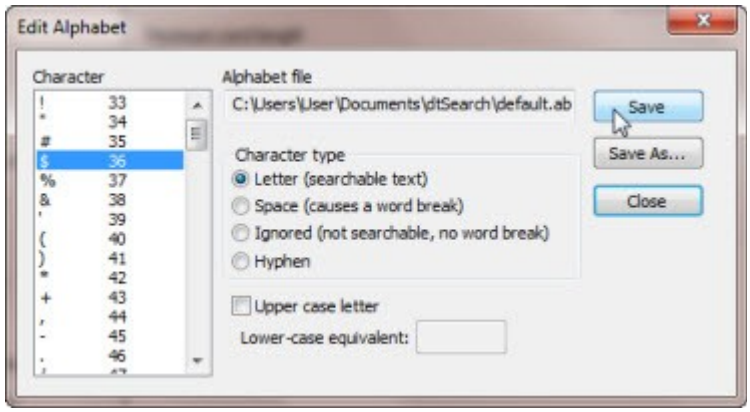
This is a Precision of : This is a Recall of:

Although you could start adding terms to get rid of unwanted results (e.g. AND NOT chickens) the only preferred method for narrowing results is to use the word proximity operator **`W/n`**, in this case the only relevant documents are those with the word dollars within one word of 120, or those with a $ sign within one word of 120.

A simple search query of **`120 w/1 dollars`** will give a single result, this is

a Precision of : a Recall of:

For a more thorough search of **`($ w/1 120) or (120 w/1 dollars)`** you need to make the **$** sign searchable. Edit the alphabet file to make **$** a **Letter** as shown.



Now update the index and repeat the search.

The **`pre/n`** directed proximity operator is more precise than the **`W/n`** operator, because it specifies the sequence of the words, now search again but modify the search query to:

```
($ pre/1 120) or (120 pre/1 dollars) or $120
```

This will return all three relevant documents and no non-relevant documents

This is a Precision of : This is a Recall of:

Clearly this is a simple example, in practice we may not always know the exact wording in a document. The **Indexed word list** in dtSearch Desktop's **Search** dialog can often be used to gain insight into the document collection, for example you may have noticed the words 'bucks' or 'USD' appear in the word list, could these be clues to relevant documents? Try searching for `120 bucks`, clearly the **Indexed word list** has its limits.

An **All Words** search ignores the word order, this type of search is similar to the way the major Web search engines work by default. Select **All Words** and repeat the search.

Select **Any Words** and repeat the search. An **Any Word** search gives the widest possible interpretation of the words in your Search request and should be considered a last resort if all else fails to find a useful result.

For currencies other than the dollar the Alphabet Editor in dtSearch Desktop is not usable. You can use the Alphabet File Editor in the User Thesaurus Plus Add-on to make other currency signs searchable so that your searches will have much higher precision.

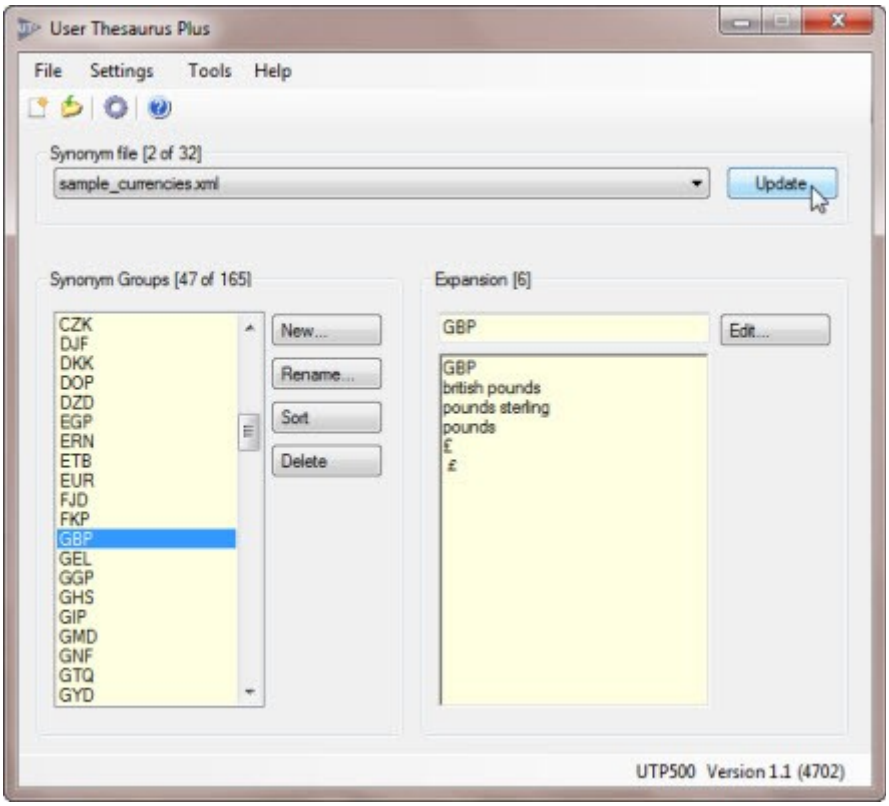
Refer to the User Thesaurus Plus Web Help: Working with Thesaurus, Macro and Alphabet files > Alphabet File Editor and make the **£** sign searchable, then update the index. This time we need to find all the documents that mention a sum of 2 million pounds,

You should be aware that you will need to adapt your search queries when you make currency signs searchable. For example if you make the **£** sign a searchable letter a search for £2000000 will only find documents containing £2000000, you will not find a document that contains £ 2000000 (i.e. with a space between the sign and the digits). To ensure that you will find documents with or without a space between the currency sign and the amount, you will need a search query of:

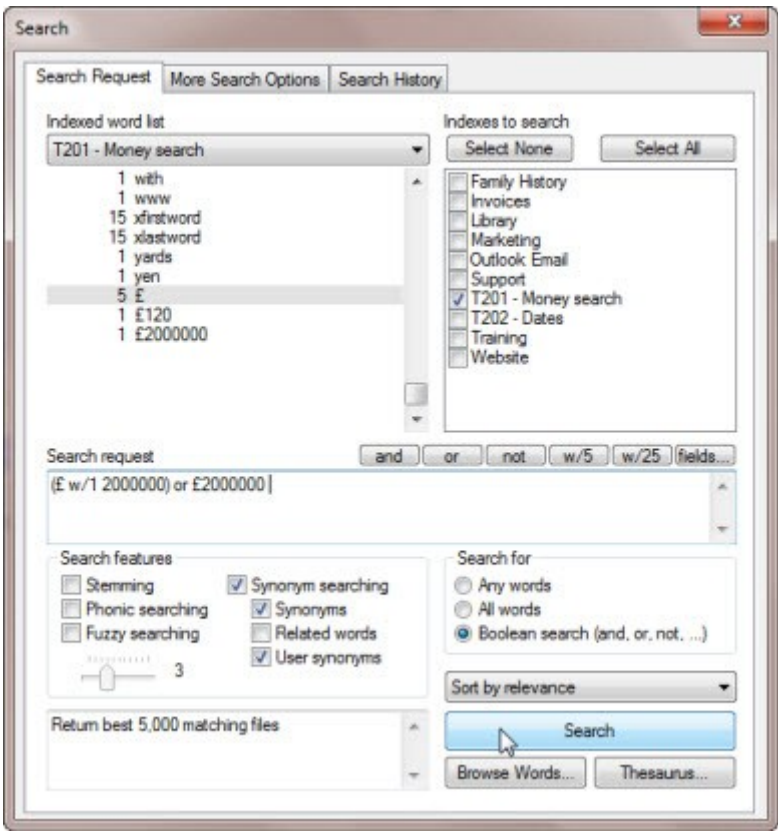
```
(£ w/1 2000000) or £2000000
```

Unselect **Synonyms** and try the search query above, you should find two documents only.

Now lets expand the search further, in User Thesaurus Plus select the **sample_currencies** file from the drop-down list and click on the **Update** button.



In the Search dialog select **Synonym Searching, Synonyms** and **User Synonyms** and repeat the **Search** again.



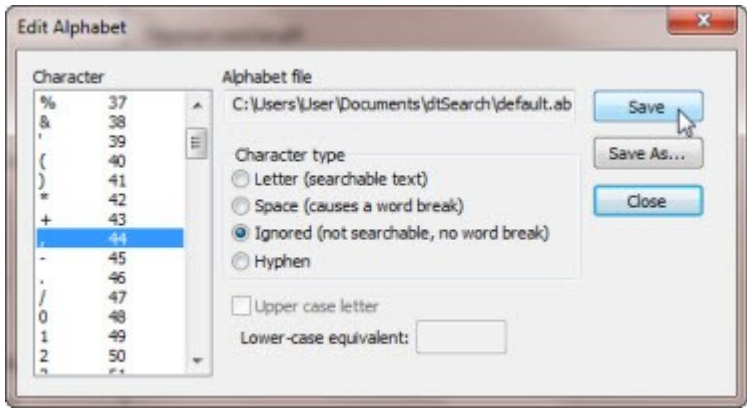
You should now find five documents:

money_3 - £2000000
money_4 - £ 2000000
money_5 - 2000000 Pounds Sterling
money_6 - 2000000 GBP
money_7- GBP 2000000

The default settings for punctuation in dtSearch Desktop/Network is to treat a comma and a full-stop as a non-searchable 'space'. To be able to find 2,000,000 or 2.000.000 you need to edit the Alphabet file from within dtSearch Desktop/Network to change the comma or full-stop from a 'space' to 'ignored', only change one of the characters depending on the format that you expect to find, for example in Germany 2 million dollars and 89 cents would appear as 2.000.000,89 whereas in the USA it would appear as 2,000,000.89.

(Changing both punctuation characters to 'ignored' would result in the index containing 200000089).

In dtSearch Desktop edit the alphabet file to make the comma (character code 44) 'ignored' and Save the settings.



Now update the T201 index and repeat the search.

You should now find two more documents:

money_8 - £ 2,000,000
money_9 - 2,000,000 pounds

To expand the search further add more OR'd terms like this and repeat the search:

```
(£ w/1 2000000) or £2000000 or (2 w/1 million)
```

Finally don't forget that people make spelling mistakes, in the **Search** dialog select a Fuzzy search feature of 3 to catch a misspelling of *'milion'* instead of *'million'* for example and select Stemming to ensure that documents containing alternative word forms are found, for example the word *pound* when searching for *pounds*.

Using these techniques you should find all 11 relevant documents.

NOTES

This page is blank for trainee notes. Trainer manuals are available with additional technical notes and training material.

