

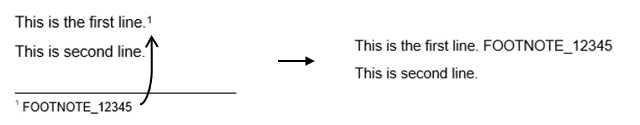
**DocIDParser.exe**

The DocIDParser utilizes Regex to extract document IDs from .txt and .docx files. While it is highly effective in extracting doc IDs, users should be mindful of how each option can impact results. Steps are outlined below:

Locating Files

Enter a path direct to your file, or to a folder containing your files. If the latter, you will be prompted to choose the file extension to target within the folder. The tool targets files in all subfolders and includes a designated column in the results to specify the subfolder path where each document is situated.

**NOTE:** When parsing from .docx files, the tool first rearranges footnotes in the *actual* order they appear in the document, like so:



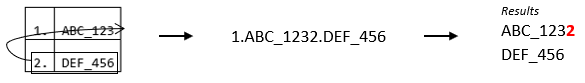
Collapsing Lines

Choosing to/not to collapse lines when reading the text of your document(s) can have varied effects on your results. For instance, if the doc ID **DEMO\_ID\_123456** was split across 2 lines...

The document **DEMO\_ID\_**

**123456** is not real.

...the tool would not recognize **‘DEMO\_ID\_123456’** as a match unless you chose to combine lines. When combining lines, the tool reads each document as if it were one long line. A common occurrence here is where the first character of the line below is undesirably included in a doc ID match, like so:



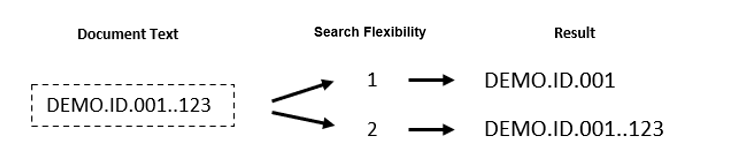
Whitespace Characters

When using the tool on OCR generated text, or any other text you suspect may have doc IDs separated by whitespace, you may choose to allow for whitespace characters in your search. If this option is chosen, the tool will allow for spaces in between each character of a doc ID.

Search Flexibility

The flexibility of your search determines the number of connectors permitted between each component of a doc ID, as well as the number of whitespaces between each character of a doc ID (only if you previously chose to allow for whitespace characters). You may wish to choose a higher search flexibility if you know your text was derived from OCR (especially poor OCR), or if you simply wish to be inclusive with results.

To illustrate search flexibility:



Suffixing Letters

This option allows you to match suffixing letters on doc IDs. Only 1 letter is matched, and it may or may not have a period beforehand. For example, both of the following IDs would be matched.

*Results*

DEMO\_ID\_123456.C

DEMO\_ID\_123456C

The tool will only ask about suffixing letters if the DocIDParser.exe file is accompanied by a text file by the name ‘Suffix.txt’, where each letter to search for is on a new line, like so:



**NOTE:** The tool will match a suffixing letter preceded by whitespace if you allowed for whitespace previously.

Character Conversion Errors

Here you can choose to allow for character conversion errors in doc IDs. This is most common in text generated from OCR. If this option is chosen, errors are considered in both a doc ID’s prefix and its number elements. Some examples you may be aware of:

**j -> i 0 -> O 5 -> S b -> 6**

The tool has a built-in dictionary of commonly misrendered characters that it works from.

Doc ID Length Adjusting

This option has been found useful for larger and messier parsing jobs. Post-parsing, the tool will shorten your doc IDs on a prefix basis, given a weighted score for each length of each prefix, taking into account both length and frequency of the length. The goal here is to remove any ‘junk’ on the end of each doc ID, which is usually the result of a more inclusive search.

**NOTES:**

* Doc IDs that have been matched with a suffixing letter remain unchanged in this process.
* Length adjusting is not document specific – it is done on all doc ID matches after searching.
* A raw version of results will also be saved if length adjusting is chosen.

Prefixes

Enter the doc ID prefixes you want to search for in your document(s). The tool first analyses your document(s) to generate a list of suggest prefixes (you will be asked to specify a minimum length cutoff for any suggested prefixes in order to remove clutter – 3 is usually a safe option). For ease of working with this generated list, use these keyboard shortcuts:

**s** -> Select first item

**r**  -> Remove first item

**sa** -> Select all items

You can also just type in the prefixes as they appear in the list. Note that entering a prefix once you have already selected it will remove it from your selected. You don’t have to enter any prefixes from suggested.

De-duplication

Here you can choose to de-duplicate your references. This is done on a document level. **Only the first occurrence of a doc ID remains in the results.**

Results

In addition to the list of doc IDs found, you’ll be provided with a hit report. This report will tell you how many hits the tool got for each prefix, allowing you to compare the number of hits you get within the actual document(s) (Ctrl + F) to what the parsing tool says are the hit counts. You’ll be provided with the number of hits from both before and after de-duplication (if you chose to de-duplicate). Note that if you chose to combine lines, your Ctrl + F hits may not match your prefix hits (unless you were to combine lines in your document before using Ctrl + F).