Table of Contents

Regular Expressions

Regular expressions (aka regex) is useful for replacing patterns of text, such as headers/footers with page breaks or simply removing them, or replacing line breaks as is common when text is converted from a PDF (to remove middle of word or middle of sentence breaks).

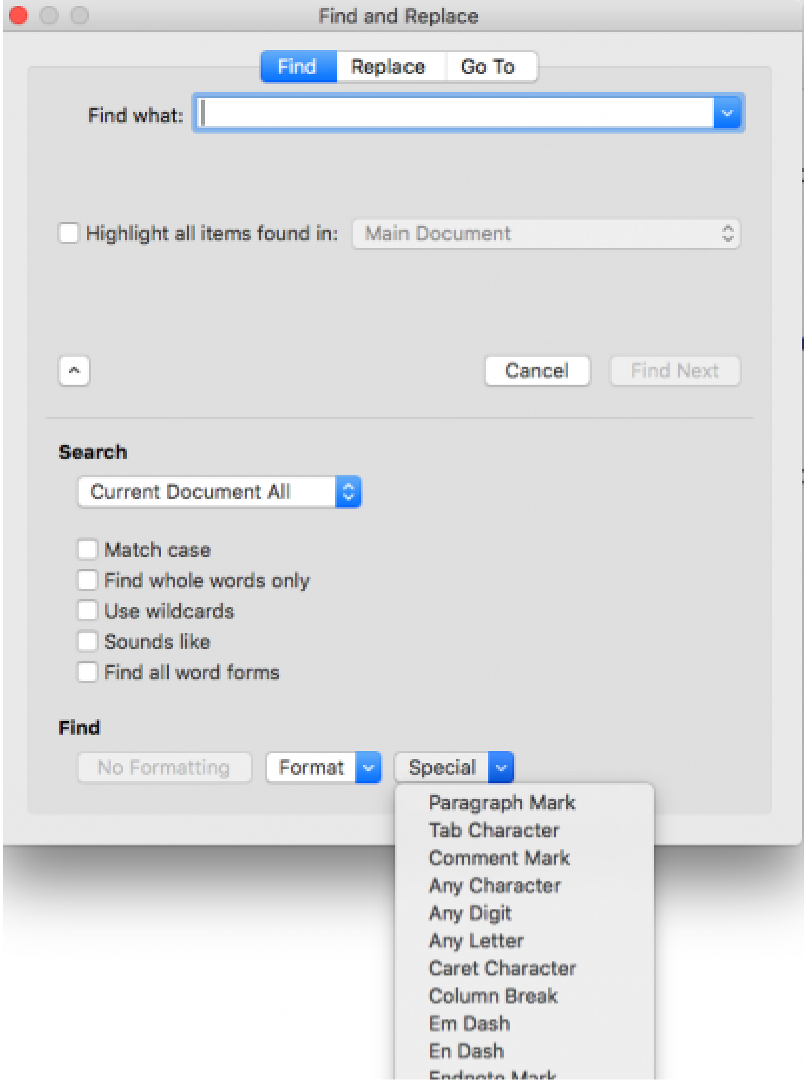
With regex, you can define patterns of text in a number of different ways, but the most commonly used ones for our purposes are **Ranges** and **Groups**. For more information about others, you can take a look at this helpful webpage:

- Ranges
  - Square brackets are always used in pairs and are used to identify *specific characters* or *ranges of characters*. You can use any character or series of characters in a range [ ], including the space character. For example:
    - [A-Z] will find any upper case letter;
    - [a-z] will find any lower case letter;
    - [A-z] will find any letter (upper or lower case);
    - [0-9] will find any number
    - [abc] will find any of the letters a, b, or c.
    - [F] will find upper case “F”
    - [Fred] will find "Fred"
- Groups
  - Round brackets are used in pairs to enclose *groups*. For example:
    - `([A-Z][A-Z])-([0-9])` Will find any two capital letters followed by a hyphen and a number, like `BB-8` or `LY-5`
  - They must be used in pairs and are addressed by number in the replacement. In the replace field, \1 represents the first group, \2 represents the second group, and so on. For example:
    - If you wanted to remove the hyphen from "BB-8" you would enter `\1\2` (i.e., the two groups with nothing between them) into the Replace field. Or, if you wanted to change the hyphen to a space, you would enter `\1 \2` (i.e., the two groups with a space between them) into the Replace field.
    - Another example: `(John) (Smith)` replaced by `\2 \1` (note the spaces in the search and replace strings) – will produce `Smith John`

Tips

Using wildcards in Microsoft Word (this is similar to regular expressions, but Word has a lot of its own syntax)

- Word has a lot of options to find letters (`^$`) and numbers (`^#`) when using the non-regex Find & Replace, but these only work with the wildcard option *off* (which it is by default). Only turn the wildcard option on if you're using regex options. Read the info page carefully on when things apply with the wildcard option on/off.
- A lot of the codes for special characters (e.g. page break) are under the "Special..." button.



In LibreOffice & OpenOffice

Make sure that the `Regular expressions` box is checked on the Alternative Find & Replace dialog for all of the search and replace actions below.

Regular expressions in LibreOffice Regular Expressions in OpenOffice

Conversion Fixes

The following fixes assume you are using Word, unless otherwise stated.

Contribute your problems and regex solutions below. Attach your screenshots of both the problem and solution.

PROBLEM: Each line ends with a paragraph break.

SOLUTION: There is no single solution to this, but the typical pattern is to search for the pattern: `not a period`, followed by `paragraph break`, followed by `letter` and replace with the same thing minus the paragraph break.

In Word, this will only work with wildcards turned on.

Find: `([A-z] )^13([A-z])`

Replace with: `\1\2`

This looks for the pattern: `any-letter` `space` `paragraph-break` `any-letter`

The parentheses are used to group what it finds, so \1 refers to the first "any-letter" group and \2 refers to the second "any-letter" group.

In this way, you are putting back exactly what it found minus the paragraph break.

PROBLEM: Hyphenated words that break over two lines.

SOLUTION: Replace with the same text minus the hyphen.

Find: `([a-z])-^13([a-z])`

Replace with: `\1\2`

Using a-z restricts what it finds to lowercase.

You will likely have to do it again for lines that end with a comma, and possibly en and em dash. Look through your document for patterns of anything else it might have missed.

PROBLEM: OCR converted some "1" digits to "i/l" letters, resulting in dates like "i984" or numbers like "3l".

SOLUTION: Replace "i/l"s that come immediately before of after a number with "1"s. This will be done in two steps

1. Find: `([iI])([0-9])` This will find both lower and upper case "i"s, immediately followed by a digit.
2. Replace: `1\2` This replaces the first group `([iI])` with the number **1**, and leaves the second group `([0-9])` as is.

1. Find: `([0-9])([iI])` This will find the digit immediately followed by the letter i (e.g., 3i).
2. Replace: `\11` This leaves the first group `([0-9])` as is, and replaces the second group `([iI])` with the number **1**.

PROBLEM: OCR did not recognize spaces around quotation marks.

- Example A: As one of Montgomery's British staff officers later put `it,"I` feel Monty was astonishing in his relationship with all the Dominion troops.
- Example B: The "nasty little `troublemaker,"as` Montgomery was widely known in the British army…

This problem has an added complexity; the pattern has two different solutions:

- Example A will need to say: … later put `it, "I` feel Monty… (or, comma-space-quotation mark)
- Example B will need to say: The "nasty little troublemaker`," as` Montgomery… (or, comma-quotation mark-space

SOLUTIONS: Example A:

Find: `([,])(["])([A-z])`
Replace: `\1 \2\3`

Example B:

Find: `([,])(["])([A-z])`
Replace: `\1\2 \3`

Notes:

- You will **not** be able to use "replace all" in this situation. You will need to keep hitting `Find Next` and replacing the pattern with the appropriate solution.
- You will also need to re-do this, searching for periods instead of commas.

PROBLEM: There are extra paragraph breaks. We want to keep the real paragraph breaks and remove the fake extra paragraph breaks.

SOLUTION: See: Find & Replace

PROBLEM: There are newlines/line breaks (↵) instead of paragraph marks (¶).

SOLUTION: See: Find & Replace

> **PROBLEM**: Running headers. Example, where the first three numbers and the three numbers after the filename is the page number: `231(paragraph break)MacG_9781770494220_5p_all_r1.indd 231(paragraph break)10/27/14 11:56 AM(paragraph break)`
>
> **SOLUTION**: See: Find & Replace

In LibreOffice:

- Verso (left hand)
- `\p[0-9OoIil]{1,3}\s+.+\p`
  - taken piece-by-piece, this means:
  - `\p` : a paragraph marker
  - `[0-9OoIil]{1,3}` : between one and three numbers or "number like" symbols. (OCR programs often mistake o or O for `0` and `I`, `i`, or `l` for 1.)
  - `\s+` : one or more whitespace character (spaces, tabs, etc.)
  - `.+` : one or more of any character
  - `\p` : a final paragraph marker
- Recto (right hand)
- `\p.+\s+[0-9OoIil]{1,3}\p` ### Detect bad line breaks ###
- `[^\."?!]$`