Filename Matcher

Table of contents

Filename matcher script. What is it?

The filename matcher script traverses all subdirectories on a computer from a directory starting point and lists all files whose filenames match the processed filename by selected file type (e.g., pdf or docx or even * for any file type).

Why bother with a file matcher script?

Example 1: Journal articles. Suppose that, like me, you download different journal articles. You keep them all in one “Journal articles” directory. But then, you start copying many of them around in various other directories, may be because you use them as references in some papers or is some courses you are taking or teaching. You may want to know where all these files are perhaps because you have highlighted the content of one of them.

Example 2: Newspaper articles with metadata in filename. You have downloaded thousands of newspaper articles in pdf format. But as we know, with NLP tools we can only use txt type files. If like me, you do historical research these newspaper articles may simply be of too poor quality for automatic conversion to txt. So, you do it by hand. Or you have different people helping you to transcribe the pdf files. In either case, you may want to know which files have been transcribed already not to duplicate very labor-intensive work. The file matcher would list all the files that have the same filename and different extensions (pdf and docx).

Exact filename match and partial match

The filename matcher algorithm allows for an exact match of the entire filename or a partial match based on a subset of items embedded in the filename (e.g., you search only for files matching the newspaper name in such filenames as The New York Times_8-9-1827_1_3, i.e., the New York Times article of 8-9-1827 on page 1, column 3, ignoring any information after the newspaper name, i.e., ignoring the three _ separated items 8-9-1827_1_3 that come after the newspaper names; in practice, you would be searching for any file whose name contains The New York Times.

Using a csv list of files for finding matches

You can also use a csv file that contains a list of filenames to be used for finding matches.

This option is particularly useful if you want to process partial matches. Currently, in fact, the partial match option is not available.

Creating the csv list

You can create the csv list running file_manager_main.py with the list option and any filter you wish to apply. **To generate a csv file for partial matches, use the filter "By number of embedded items."**

Source and target extensions

The script identifies files having the same filename and different extensions (e.g., The Atlanta Journal_3-12-1956_4_2.pdf and The Atlanta Journal_3-12-1956_8_2.txt). All subdirectories of a selected directory will be searched for a selected pair of **source** and **target** extensions (e.g., pdf and docx).

Using * * for both source and target file types will identify any file with the same exact filename and different extensions of any type.

Input

In INPUT the script takes a directory to start traversal.

Output

In OUTPUT the script produces three csv files:

duplicates_pdf_star.csv containing a list of duplicate files with source extension pdf and target extension *; these values would change depending upon what you select as source and target extensions;

matched_pdf_star.csv containing a list of files matched by extension types;

unmatched_pdf_star.csv containing a list of files for which no matches were found by extension types.