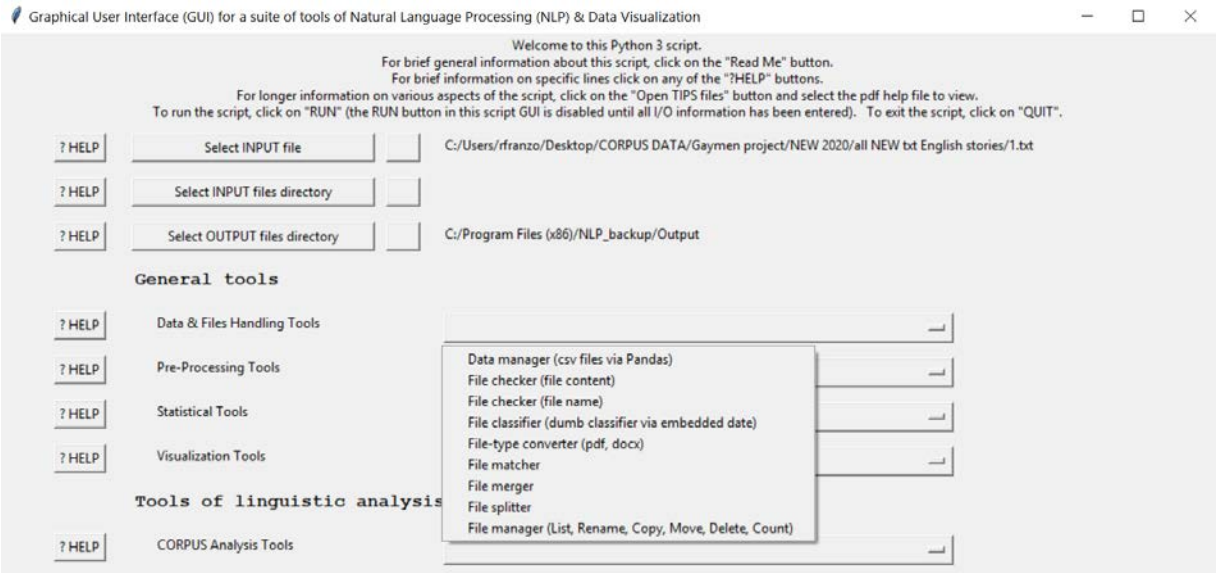
File Splitter

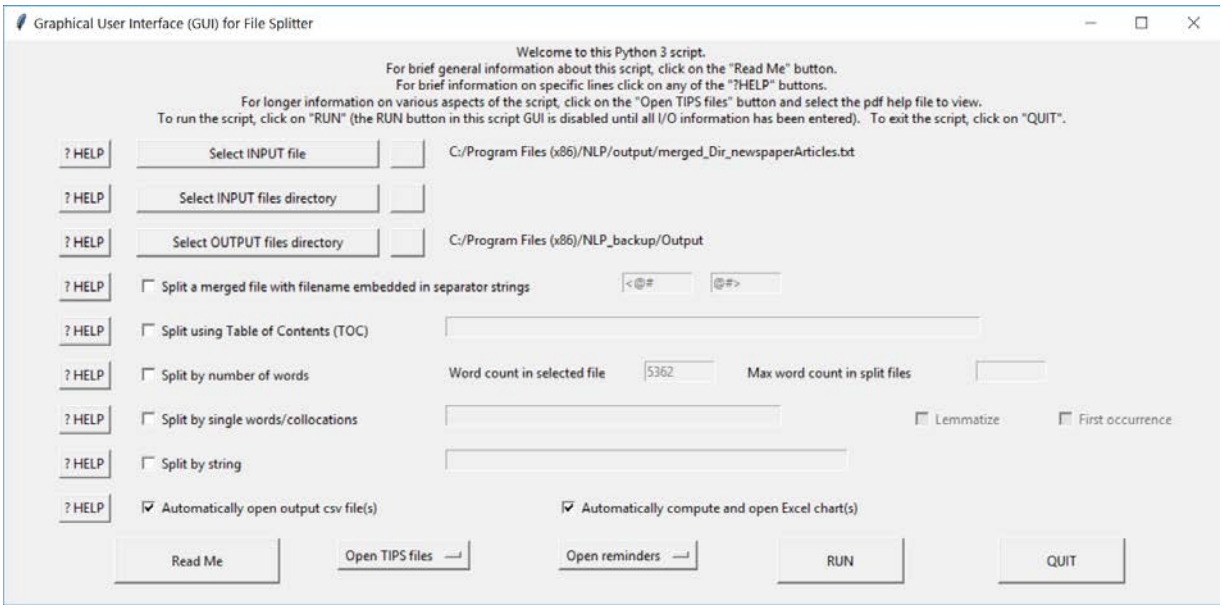
Table of contents

A file splitter GUI

When you run in command line python NLP_main.py, under General tools, Data & File Handling Tools, select the option File splitter to open the File splitter GUI.



You can also run directly in command line python file_splitter_main.py to open that same GUI. Once active, the file splitter GUI provides several options for splitting files.



What does a files splitter do?

It takes a text file and splits it into a set of smaller subfiles using a variety of criteria.

Split options

The GUI provides several options for splitting.

Split a previously merged file

You can split a previously merged file with filenames embedded in start/end strings (e.g., <@#The New York Times_11-02-1992_4_1#@>). **You can only split a merged file if the merged file contains special start/end strings. The algorithm uses these strings to split the file.**

Split using Table of Contents (TOC)

You can split a text file into separate files using a Table of Contents as the criterion for splitting. **The tool will work well for books or any document that has a Table of Contents.**

In INPUT the splitter script expects two types of txt-type files

1. a main txt file (e.g., The Philosopher's Stone.txt) with the body of a text and section headings (e.g., chapter titles of the Harry Potter book);
2. a TOC (Table of Content) file that contains all the section headings of the main document.

In OUTPUT, the script will split the main file into sub-documents, one document for each of the headings listed in the TOC file.

Split using a maximum number of words

You can split a txt file into separate files using a maximum number of words as the criterion

for splitting.

The number of words in the selected file is displayed in the second widget of the GUI option, "Word count in selected file." You will need to enter to desired maximum number of words in each split file in the third widget, Max word count in split files.

Split using single words or collocations

You can split a text file into separate files using single words or collocations, i.e., combinations of words such as "coming out," "standing in line," as the criterion for splitting.

You have the option to **lemmatize** the expression you entered (thus, the expression "coming out", when the "Lemmatize" checkbox is ticked, would be checked for "coming out", "come out", "came out", "comes out").

You also have the option to split a file by the **First occurrence** of the expression entered (which would always result in two txt output files) or of splitting the file at every occurrence of the expression entered (thus leading to multiple output txt files, one for each occurrence of the expression).

THE SCRIPT CAN EITHER SEARCH IN A CONLL TABLE OR IN TEXT FILE.

Split using special string values

You can split a text file into separate files using special string values as the criterion for splitting. **The option is currently under generalizing and testing.**

Why would you split files?

There are scripts that can only deal with a maximum number of characters (more than words). Thus, Stanford CoreNLP can process up to 100,000 characters. The DBpedia script also seems to break with longer files. The NLP Suite script for CoreNLP and DBpedia automatically split the text below the maximum allowed.

But there may be also analytical reasons for splitting a text. Thus, if you are analyzing novels, you may wish to investigate what changes in the different chapters: scenes, characters, linguistic features. Splitting by TOC would help you approach these differences. If you are investigating gay men autobiographical short stories, you may wish to see what changes before and after the "coming out": does sentiment analysis highlight more or less happiness in the two chunks of text? Do social relations (family, friends) change? Splitting the stories at the first occurrence of "coming out" would allow you answer these questions.

Input

In INPUT the different split scripts expect either a single file or a set of text files in an input directory.

The option "Split using Table of Contents (TOC)" has different arguments (see above).

Output

In OUTPUT, the script will generate the split files in a subdirectory, named split_files, of the directory of the input file or directory.

The option “Split using Table of Contents (TOC)” has different arguments (see above).

References

TIPS_NLP_File merger.pdf