Setup INPUT/OUTPUT Options

Table of contents
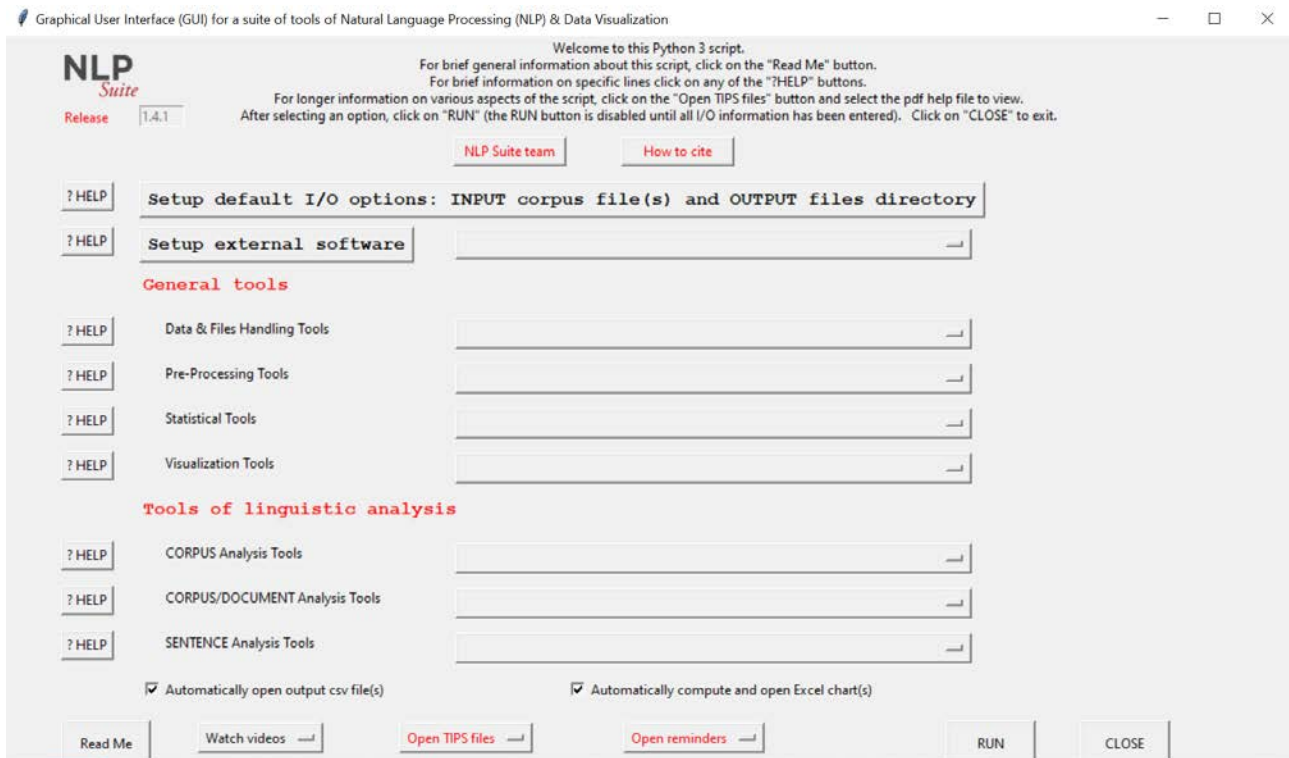
Setting up I/O configuration options

The computational linguistic tools in the NLP-Suite by and large require txt files in **input** for analysis, whether a single file or a set of files in a directory (the *corpus*). Occasionally, a specific tool may expect a df file in input (e.g., the pdf converter tools) or csv files (e.g., the Excel charts tool or the GIS – Geographic Information System – tools. But unless otherwise specified, the typical file type in input is txt. In **output**, the typical file type produced by the NLP-Suite tools is csv, but could also be txt (e.g., pdf file converter or co-reference resolution tools), HTML (e.g., the various annotators), KML (e.g., the GIS tools that produce files to be displayed in Google Earth Pro or Google Maps).
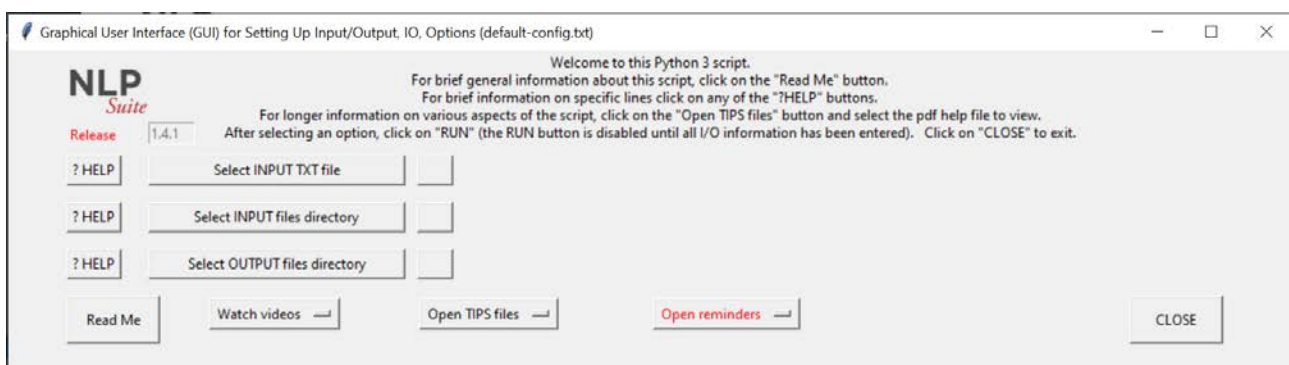
The default I/O configuration

Typically, a researcher will analyze the same corpus through various NLP tools. It is convenient to have a **default set of Input/Output, I/O, options**. These I/O configuration options will be used by every GUI (Graphical User Interface) NLP tool. You can setup the default I/O configuration in the opening GUI, the NLP_menu_main.
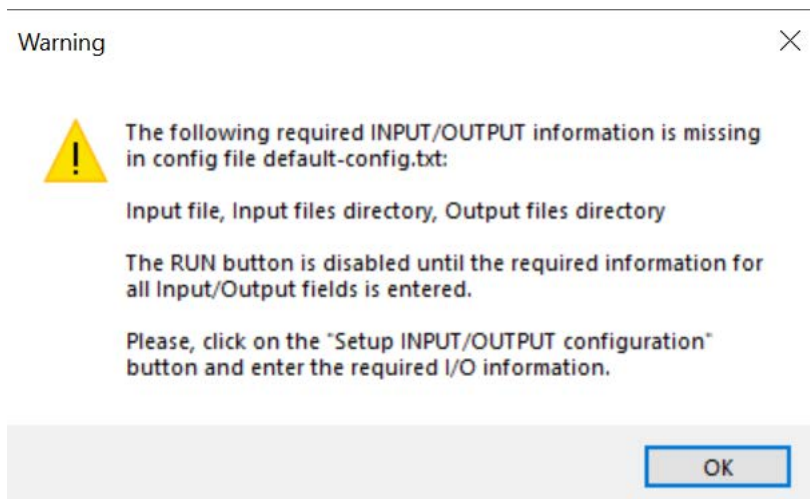
To setup the default I/O configuration click on the top button **Setup default I/O options: INPUT corpus files(s) and OUTPUT files directory**.

After clicking on the **Setup default I/O options** button, the following GUI will open where you can enter either the filename (in txt format) or the directory containing a set of txt files that will be used in input as the default corpus. The title line of the GUI tells you that the selected options will be saved to saved in the *default-config.txt* file in the config subdirectory.

If any I/O values are missing (as in the GUI displayed here), a warning message will pop up.

Warning ✕

⚠️ The following required INPUT/OUTPUT information is missing
in config file default-config.txt:

Input file, Input files directory, Output files directory

The RUN button is disabled until the required information for
all Input/Output fields is entered.

Please, click on the "Setup INPUT/OUTPUT configuration"
button and enter the required I/O information.

OK

Select INPUT TXT file

Click on the *Select INPUT TXT file* button if you wish to process a single txt file.

Select INPUT files directory

Click on the *Select INPUT files directory* button if you wish to process all the txt files contained in a directory.

The INPUT/OUTPUT options are mutually exclusive

The two I/O options *Select INPUT TXT file* and *Select INPUT files directory* are mutually exclusive, i.e., you can only enter one or the other. If you have selected a txt file (as displayed on the I/O widgets) and would rather select a txt file directory, simply click the *Select INPUT files directory* button and select a directory. The option associated with the *Select INPUT TXT file* button will be cleared and the next directory option will be displayed.

Select OUTPUT files directory

The output is a text file containing the candidate documents for each set of duplicates and the list of its duplicates. The features extracted by Lucene and used to compare the documents are also listed as well as the similarity percentage.

Values saved in the *default-config.txt* file will be used for all GUIs in the NLP-Suite, unless otherwise specified for a specific GUI.

The RUN button will be disabled until a default I/O configuration will have been saved and you will not be able to access any of the NLP-Suite options.
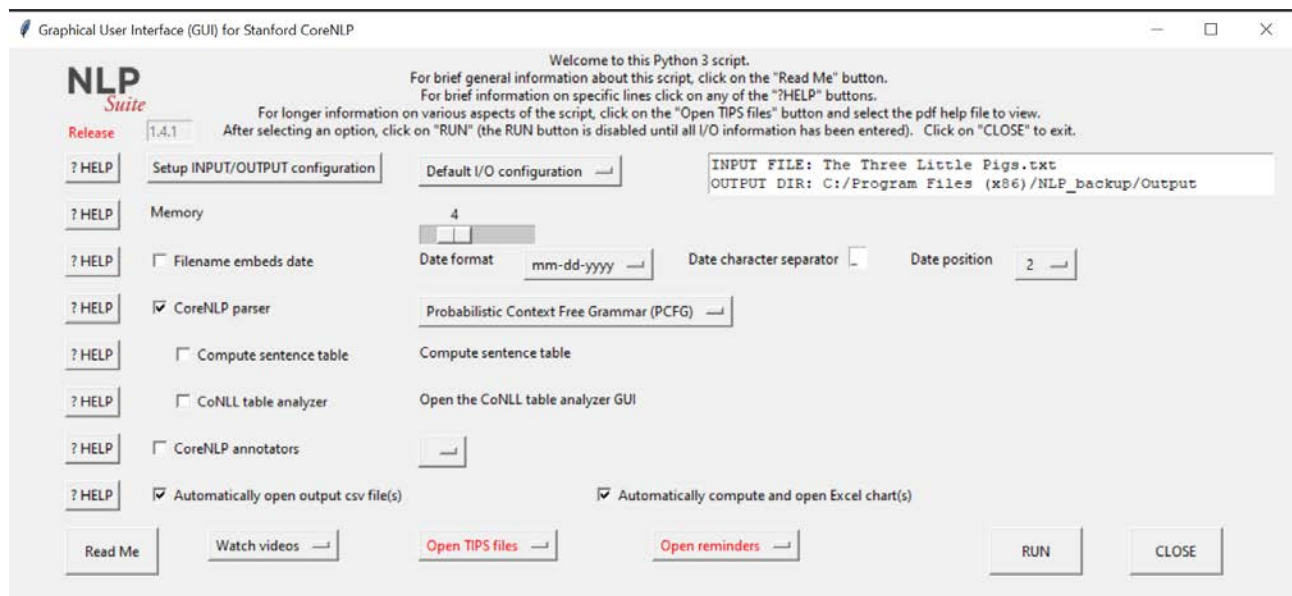
The alternative, GUI-specific I/O configuration

As a data analyst, it is very likely that you will analyze a single corpus with a variety of NLP tools. In this case, the default I/O configuration will be ideal. You will not need to enter a new configuration for the different tools you use. However, for a specific tool accessed via a specific GUI (e.g., GIS_main) you may need different I/O options. For instance, the tool may require a csv

or an html file in input, or you may wish to run a test on a sample txt file rather the default options. Whatever the reasons…
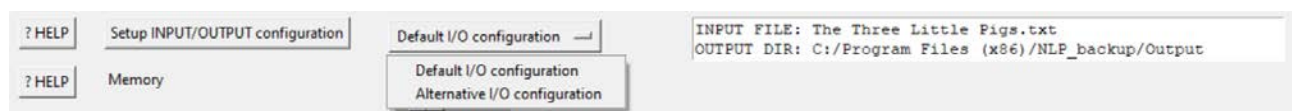
How do you setup an alternative I/O configuration?

When you run any NLP-Suite tool, e.g., Stanford CoreNLP, the GUI will display at the top the standard, default configuration (the dropdown menu is set to Default I/O configuration).

The selected I/O values will be displayed in the setup window.

But, by clicking on the dropdown menu, you can also choose the option *Alternative I/O configuration*.
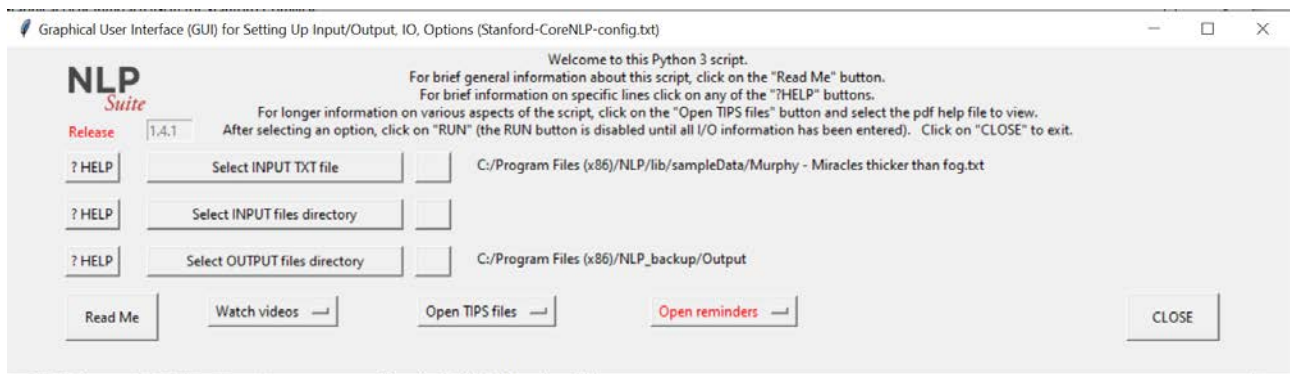

Now, select the option *Alternative I/O configuration*.

The setup window at the top of the GUI will now display a different set of available options. How do enter these options if none are available or change them?

Click on the *Setup INPUT/OUTPUT configuration* button. This will open the I/O setup GUI. The top title line, however, will tell you that whatever options you choose will be saved in the Stanford-CoreNLP-config.txt; this file name will change depending on the NLP tool you are using.

Saving I/O configuration options

When hitting the CLOSE button in the Input/Output configuration setting GUI, if the selected options are different from the currently available options, you will be asked if you want to save the new I/O options in either the default or GUI-specific config file.