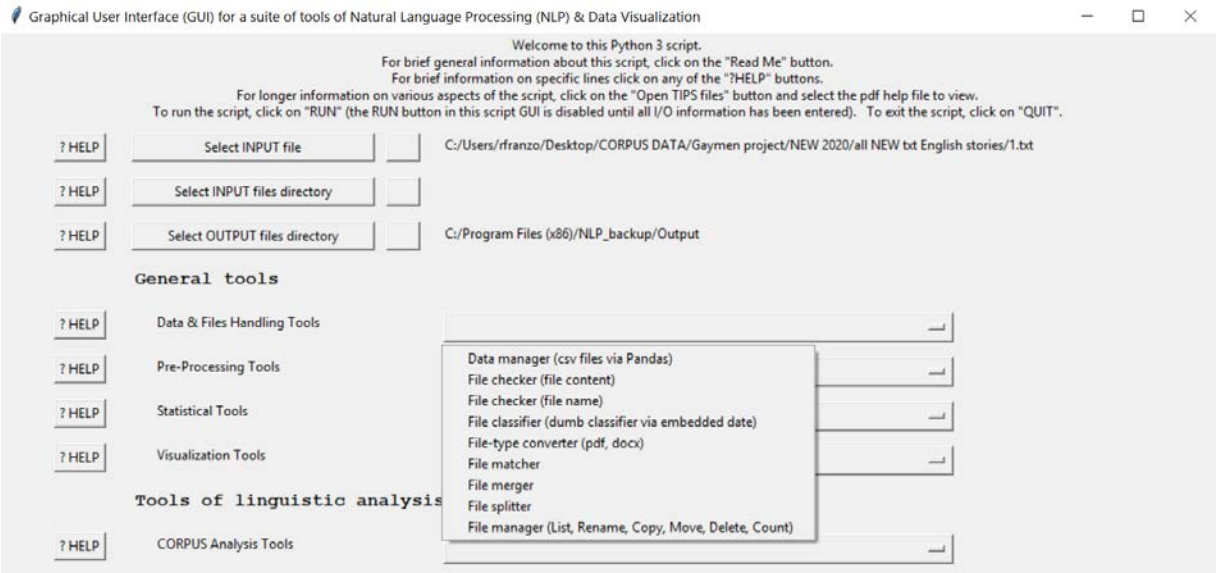
File Checker & Converter

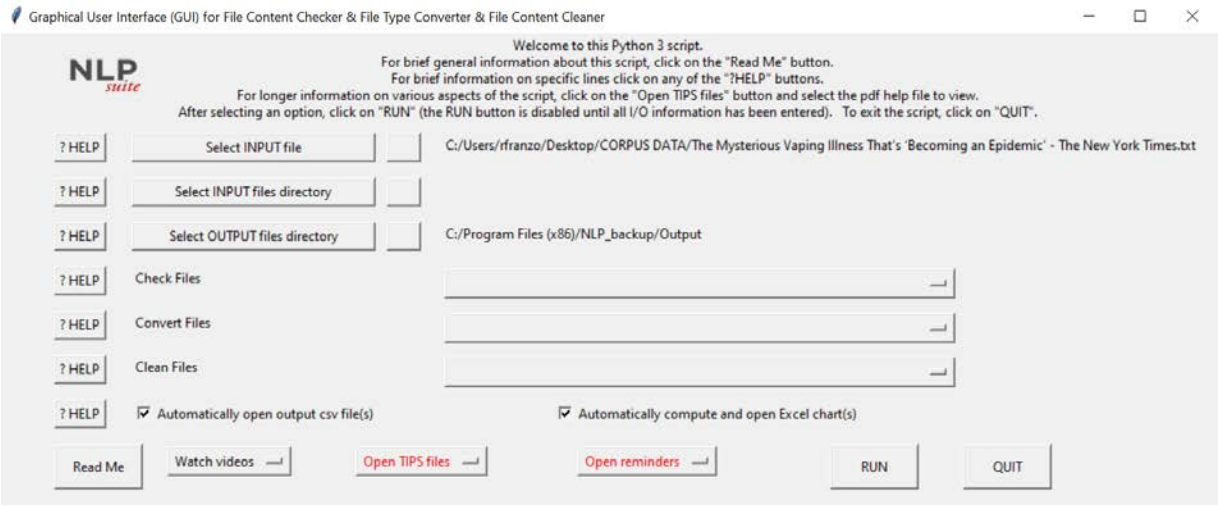
Table of contents

A file checker & converter & cleaner GUI

When you run in command line python NLP_main.py, under General tools, Data & File Handling Tools, you will find two options – File checker (file content), File-type converter (pdf, docx) – that will allow you to open the GUI with several options to check the content of files and convert the type of files.



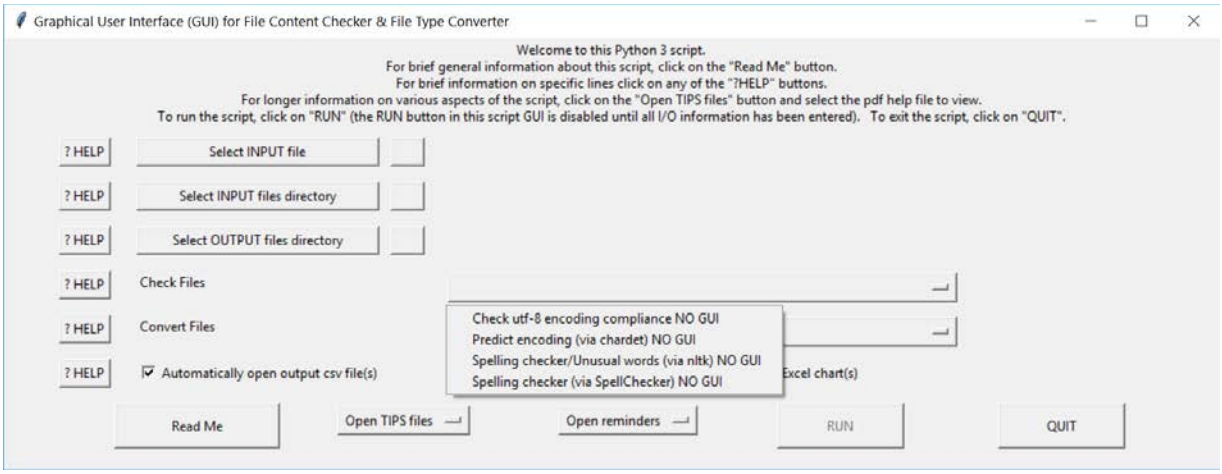
You can also run directly in command line python file_checker_converter_main.py to open that same GUI. Once active, the file checker & converter GUI provides several options for each task.



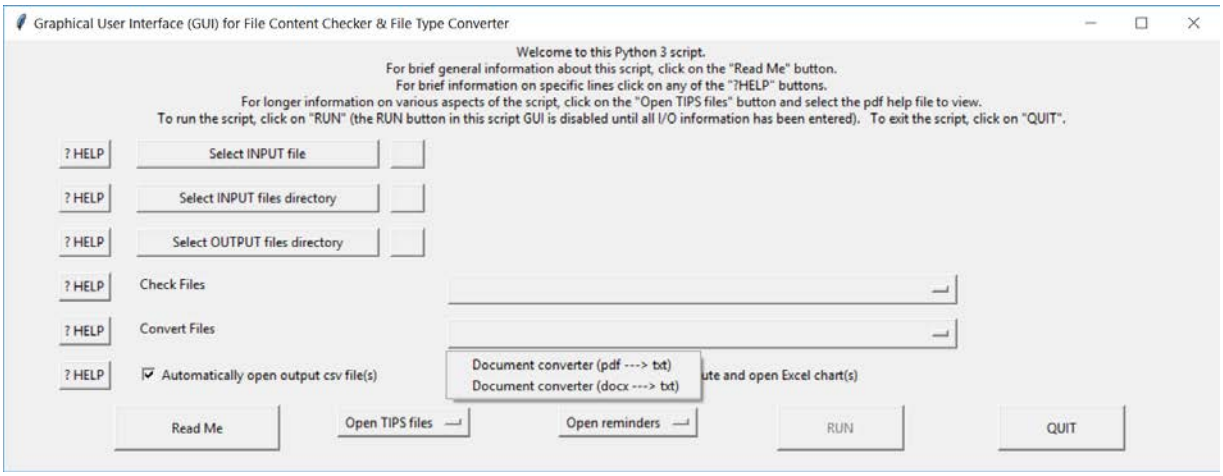
Check Files option

The Check Files dropdown menu provides several options. Please, refer to the following TIPS for more information on these options.

TIPS_NLP_Text encoding (utf-8).pdf
TIPS_NLP_Spelling checker.pdf



Convert Files option



Please, using the dropdown menu, select one of the options available for converting the file type (from pdf to txt, or docx (NOT doc) to txt.

pdf to txt converter. The NLP Suite relies on the Python package pdfminer to convert pdf files into txt files (https://pypi.org/project/pdfminer/). Pdfminer can convert 2-column texts, such, for instance, as journal articles.

In INPUT, when a directory is selected, all files in a directory and its subdirectories can be converted. The script will ask users whether they want to convert files in subdirectories.

docx to txt converter. The NLP Suite relies on the Python package Python-docx (https://python-docx.readthedocs.io/en/latest/) to convert docx (not doc!) files into txt files (https://www.geeksforgeeks.org/python-working-with-docx-module/).

rtf to txt converter (Mac OS only). In a **Mac Operating System**, there is a simple way to batch convert a set of rtf files to txt. Open the command prompt and change directory to where the rtf files are stored, then type:
nfind . -name *.rtf -print0 | xargs -0 textutil -convert txt
Hit return. All txt converted files will be found in the same input directory as the original rtf files.
For more information, see the post by Alexander Refsum Jensenius at:\nhttps://www.arj.no/2013/01/08/batch-convert-rtf-files-to-txt/.

Clean Files option

The script provides several options for cleaning the content of a file. Two options are particularly helpful.

Change to utf-8 non-utf-8 apostrophes & quotes

The option provides a solution to a nagging display format in csv files on Windows machines.

TIPS_NLP_Text encoding (utf-8).pdf

If when you open a csv file created by an NLP Suite Python script (e.g., n-grams) you get

some weird characters, such as â€œ on a Windows laptop (not on a Mac). Such characters are not in the text file processed by the script in input.

	A	B	C	D	E	F	G	H
1	4-grams	Frequency	Document	FileName				
2	â€œ What we need	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
3	What we need to	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
4	we need to know	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
5	need to know is	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
6	to know is how	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
7	know is how to	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
8	is how to live	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
9	how to live a	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
10	to live a life	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
11	live a life to	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
12	a life to make	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
13	life to make it	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
14	to make it the	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
15	make it the best	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
16	it the best possible.	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
17	the best possible. â€	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				
18	best possible. â€ â€“	1	1	/Users/Myself/Desktop/NLP/test_input/1.txt				

	A
1	4-grams
2	“ What we need
3	What we need to
4	we need to know
5	need to know is
6	to know is how
7	know is how to
8	is how to live
9	how to live a
10	to live a life
11	live a life to
12	a life to make
13	life to make it
14	to make it the
15	make it the best
16	it the best possible.
17	the best possible. ”
18	best possible. ” –

And the weird character is not in the input text file either.

```
"What we need to know is how to live a life to make it the best possible."

-   Socrate
– And this applies to everyone, in every culture and every country. I am very fortunate to live in a city where people are open-minded and aware of the rights and
freedoms of everyone. In Montreal, being gay is pretty well accepted, even very well accepted. There is very little discrimination and the gay community is very
present. I would even say that here, homophobic are judged more harshly than homosexuals!
Yet even here, to “come out” is not always simple. I came out at the age of 20. Today, in retrospect, I wonder why I waited so long.
```

The weird character â€œ seems to be a csv display of the character “.

> while quotation marks and apostrophes seem innocuous, they can cause problems when the CSV is opened, as opposed to imported into Excel, partly because quotation marks or other symbols such as apostrophes come in different forms. Some are curly “, some are straight ". It can depend on what application generated the characters to begin with. It can be hard to explain exactly why Excel has an issue with some characters but read on for some solutions.
> A **solution** is to replace all curly quote marks to straight ones before saving a csv file. (https://help.shotfarm.com/hc/en-us/articles/115004652968-Why-are-there-odd-characters-when-I-open-a-CSV-in-Excel-)

Find & Replace string

The option will allow you to automatically replace any string in a file or set of files.

Input

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

Output

In OUTPUT, the scripts will generate the converted files.

References

TIPS_NLP_File handling.pdf
TIPS_NLP_Text encoding (utf-8).pdf