

File Merger

Table of contents

| | |
|--------------------------------------------------------|---|
| Merging files in command prompt/terminal | 1 |
| Mac | 2 |
| Windows | 2 |
| Merging files in the NLP Suite: A file merger GUI..... | 2 |
| What does a files merger do? | 3 |
| Merge options | 3 |
| What does a merged file look like? | 4 |
| Why would you merge files? | 4 |
| Input | 4 |
| Output..... | 4 |
| References | 5 |

Merging files in command prompt/terminal

Open command prompt/terminal

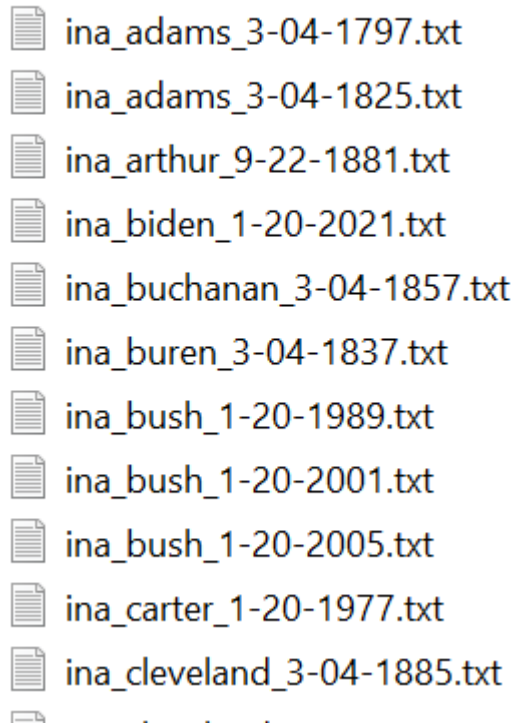


Type CD and path name to select the directory where the txt files you wish to merge are located, for example

CD C:\Users\rfranzo\Desktop\CORPUS DATA\CORPUS POTUS speeches\POTUS speeches\ina

A directory that contains the set of United States' presidents' inaugural speeches.

Name



ina_adams_3-04-1797.txt
ina_adams_3-04-1825.txt
ina_arthur_9-22-1881.txt
ina_biden_1-20-2021.txt
ina_buchanan_3-04-1857.txt
ina_buren_3-04-1837.txt
ina_bush_1-20-1989.txt
ina_bush_1-20-2001.txt
ina_bush_1-20-2005.txt
ina_carter_1-20-1977.txt
ina_cleveland_3-04-1885.txt

Mac

Type:

```
cat *.txt >> merged_files.txt
```

Windows

Type:

```
type *.txt >> merged_files.txt
```

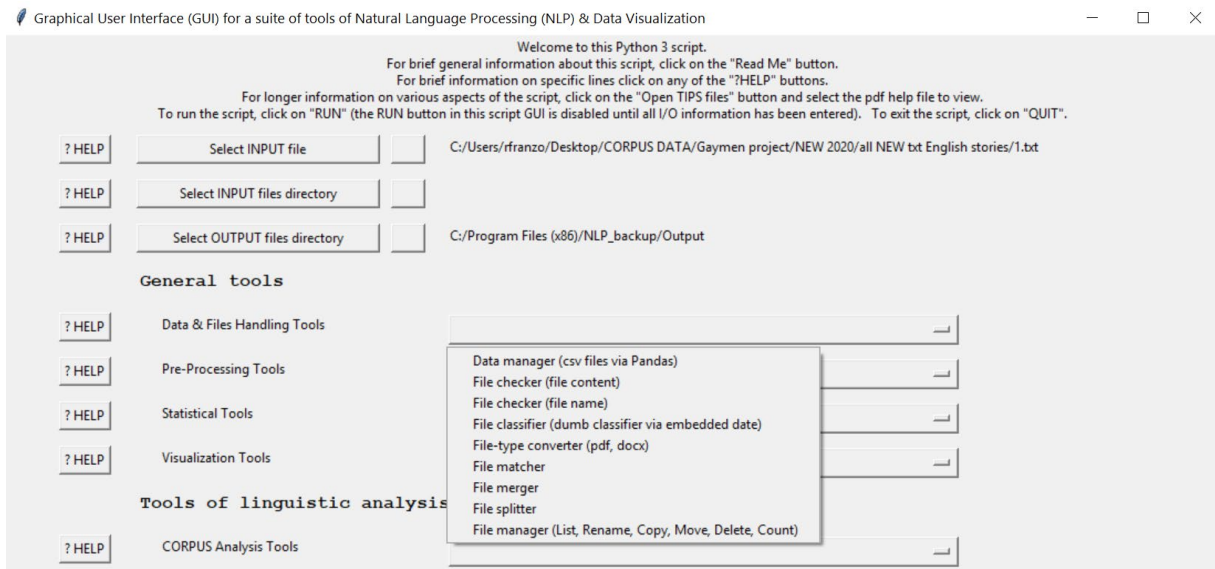
```
Microsoft Windows [Version 10.0.19044.3448]
(c) Microsoft Corporation. All rights reserved.

C:\Users\rfranzo>CD C:\Users\rfranzo\Desktop\CORPUS DATA\CORPUS POTUS speeches\POTUS speeches\ina
C:\Users\rfranzo\Desktop\CORPUS DATA\CORPUS POTUS speeches\POTUS speeches\ina>type *.txt >> merged_files.txt
```

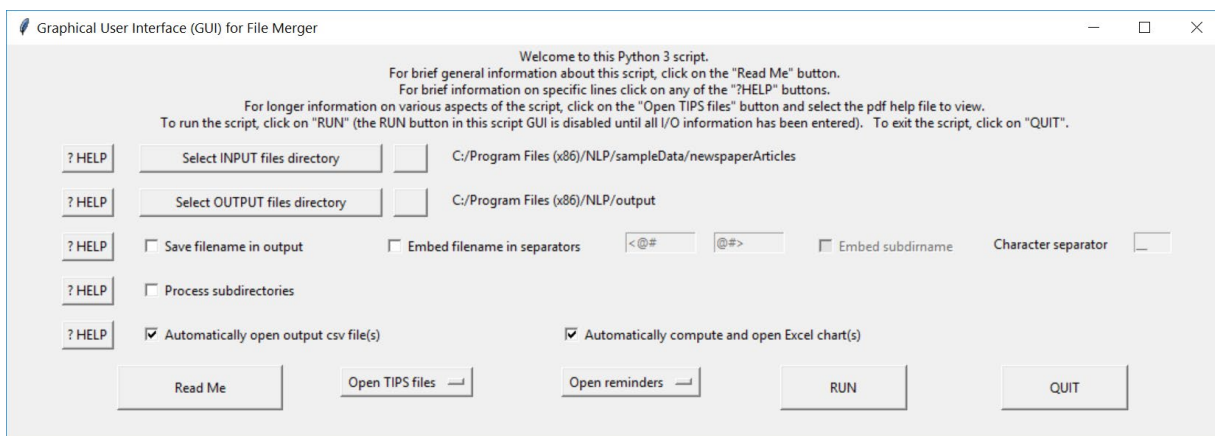
The command will create a file merged_files.txt in the same directory that contains all the txt files in the input directory.

Merging files in the NLP Suite: A file merger GUI

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



You can also run directly in command line `python file_merger_main.py` to open that same GUI. Once active, the file merger GUI provides several options for merging files.



What does a files merger do?

It takes a set of text files in input, as found in the INPUT files directory and merges them together in a single output file.

Merge options

The GUI provides several options for merging.

1. You can process the files in the input directory and all its subdirectories.
2. In the output file, along with the text of each file, you can also include the filename.
 - a. When saving the filename, you can embed the filename in special symbols (default `<@# #@>`). This will greatly facilitate file searches focusing on the symbols `<@#`
 - b. You can save in output, along with the filename, the name of the subdirectory of the file; filename and subdirectory are separated by a special symbol (default `__`).

- c. The filename is saved with the full path
(<@#C:\Users\Myself\Desktop\CORPUS DATA\Sample text\Atlanta Constitution_02-09-1888_2.txt@#>).
- d. But when the subdirectory name is saved, the filename is saved without its full path (<@#Atlanta Constitution_02-09-1888_2__1.txt@#>).

What does a merged file look like?

When saving files with their filenames embed in the default strings, this is what the output will look like using the three sample newspaper articles.

```
Current directory C:/Users/Myself/Desktop/CORPUS DATA/Sample text.
<@#C:/Users/Myself/Desktop/Atlanta Constitution_02-09-1888_2.txt@#>|.
SHOT TO DEATH.
An Incendiary Put Out of the Way.
ARMED MEN VISIT HINESVILLE JAIL,
Overpower the Jailer, Seize a Negro Prisoner and Riddle Him With Bullets-Great Excitement Prevails.
Savannah, February 8.-[Special.] -A few weeks ago a house and a warehouse were destroyed by fire in Hinesville, and all the circumstances pointed to its being the work of an incendiary. The people have been greatly wrought up in consequence. Intelligence received here tonight states that a negro was arrested there yesterday on the charge of burning the houses aforesaid. He is said to have confessed the deed, and implicated several in the crime. After a preliminary investigation, he was committed to jail in Hinesville. Last night a band of armed men overpowered the deputy sheriff, who had the prisoner in charge, and carrying him off the woods shot him to death. Great excitement prevails in that section.
<@#C:/Users/Myself/Desktop/Atlanta Constitution_02-10-1888_2.txt@#>.
THE LYNCHING IN LIBERTY.
Fuller Particulars about the Killing of an Incendiary.
Savannah, Ga., February 9-[Special]-Very little information can be obtained from Liberty County about the lynching of the negro incendiary Tuesday night. About a fortnight ago The Constitution published an account of a fire at Johnson's station, on the Savannah, Florida and Western railway. Mr. Chapman lost a store and the railway company's warehouse was burned, along with several other buildings. It was suspected that the fire was started by an incendiary, and on Tuesday a negro was arrested on suspicion. He was given a preliminary hearing, and confessed that he was one of a party of five who broke into Chapman's store. After stealing all they could carry off, the burglars sprinkled kerosene about the building and set fire to it. The magistrate committed the negro to jail. While the deputy sheriff was on his way to the Hinesville jail, he was surprised by a crowd of fifteen men, who took the prisoner away from him. The negro was carried in to woods, and it is supposed he was hung or burned. The officer never saw him more. It is expected that the other incendiaries will share the same fate.
<@#C:/Users/Myself/Desktop/Atlanta Constitution_02-10-1888_3.txt@#>.
THE LYNCHING IN LIBERTY.
Fuller Particulars about the Killing of an Incendiary.
Savannah, Ga., February 9-[Special]-Very little information can be obtained from Liberty County about the lynching of the negro incendiary Tuesday night. About a fortnight ago The Constitution published an account of a fire at Johnson's station, on the Savannah, Florida and Western railway. Mr. Chapman lost a store and the railway company's warehouse was burned and torn down, along with several other buildings. It was suspected that the fire had been previously started by an
```

Why would you merge files?

The answer to that question depends upon the NLP tools you are planning to use to analyze your corpus.

1. Some NLP algorithms require in input a set of files, rather than an individual file. Such are Gensim or Mallet topic modelling tools, the NGrams_CoOccurrences tool, or the Shape of Stories tool. **Merging a set of files into a single file will not do you any good if you are planning to use one of these algorithms.** But most tools in the NLP Suite can process indifferently a single file or a set of files.
2. Some NLP tools when processing a set of files in an input directory produce a single output file with clearly marked document ID and document names for each document. Such are, for instance, the Stanford CoreNLP parser or the Python wordcloud tool that will produce images for individua files and the merged file. **Again, no point merging files in these cases.**
3. When processing a set of files in an input directory most algorithms produce results separate for each individual input file. Yet, it may be beneficial to know the results for the corpus as a whole. Such are, for instance, WordNet, the DBpedia/dictionary Annotator, or any other tool where synthetic results are meaningful. **In these case, merge, by all means!**

Input

In INPUT the routine expects a set of text files in an input directory.

Output

In output, the script produces a text file with all the text files processed in the input directory.

References

TIPS_NLP_File splitter.pdf