

NLP Suite

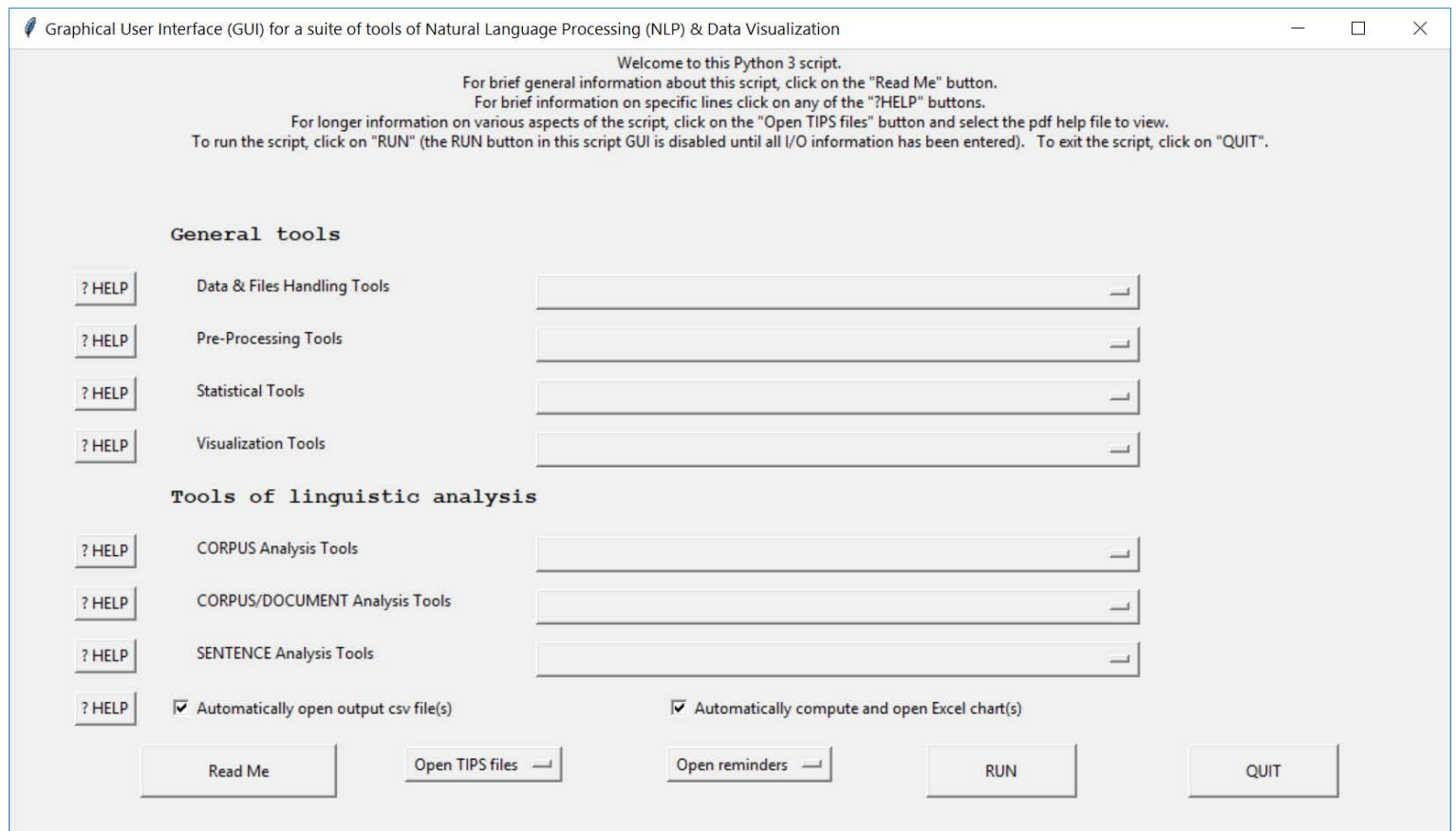
General Tools

Table of contents

The NLP Suite main GUI (Graphical User Interface)	1
General tools	1
Data & Files Handling Tools	2
Pre-Processing Tools	2
Statistical Tools.....	2
Visualization Tools	3

The NLP Suite main GUI (Graphical User Interface)

When you fire up NLP_main.py in command line, you will display the main NLP Suite GUI.



As the GUI makes clear, the NLP Suite provides two sets of tools: General tools and tools for linguistic analysis of texts via Natural Language Processing (NLP). This TIPS file provides a brief introduction to General tools. Read the TIPS_NLP_NLP Suite Tools of linguistic analysis.pdf.

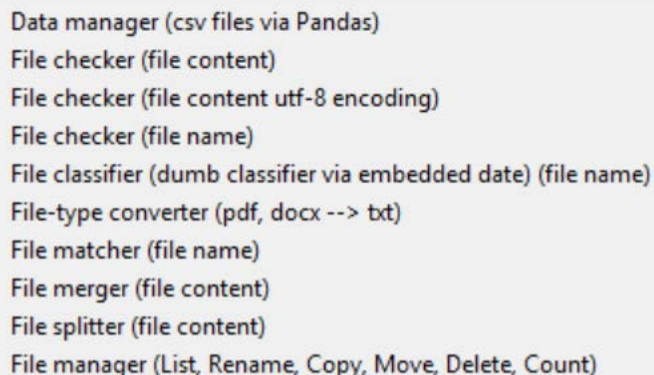
Select any of the tools using the dropdown menu and then click on RUN to open more specialized GUIs. The NLP GUI is only a front-end display of all the things you can do with the NLP Suite.

General tools

The label groups together a number of tools for dealing with data and files, very basic statistical analyses, and data visualization.

Data & Files Handling Tools

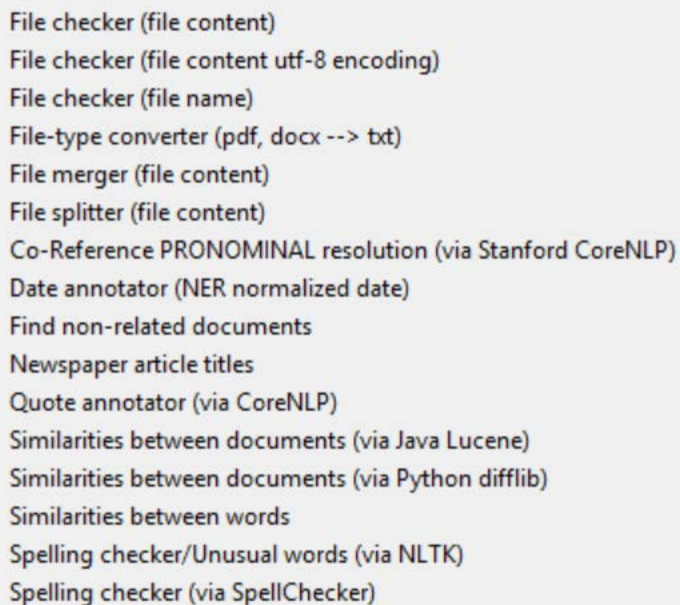
The scripts provide multiple ways of dealing with data and files. The **Data manager**, for instance, allows you to merge, concatenate, append, and extract fields from csv files. The **File-type converter** provides good options to convert to txt format any pdf file you may have downloaded from the web. **txt format is the only file-type format NLP tools accept**. There are several tools that deal with **file content** (merging or splitting files, for instance) and **file names** (checking for the well-formedness of filenames that embed metadata information). Among the file checker tools, the **File checker (file content utf-8 encoding)** is of special importance since most NLP algorithms expect in input txt files that are utf-8 encoded.



Data manager (csv files via Pandas)
File checker (file content)
File checker (file content utf-8 encoding)
File checker (file name)
File classifier (dumb classifier via embedded date) (file name)
File-type converter (pdf, docx --> txt)
File matcher (file name)
File merger (file content)
File splitter (file content)
File manager (List, Rename, Copy, Move, Delete, Count)

Pre-Processing Tools

Once again, among the file checker tools, the **File checker (file content utf-8 encoding)** is of special importance since most NLP algorithms expect in input txt files that are utf-8 encoded. But several tools offer specialized options, like checking the **Similarities between words** in a document or **Similarities between documents (via Java Lucene)** in all the documents in your corpus, or spelling checkers.



File checker (file content)
File checker (file content utf-8 encoding)
File checker (file name)
File-type converter (pdf, docx --> txt)
File merger (file content)
File splitter (file content)
Co-Reference PRONOMINAL resolution (via Stanford CoreNLP)
Date annotator (NER normalized date)
Find non-related documents
Newspaper article titles
Quote annotator (via CoreNLP)
Similarities between documents (via Java Lucene)
Similarities between documents (via Python difflib)
Similarities between words
Spelling checker/Unusual words (via NLTK)
Spelling checker (via SpellChecker)

Statistical Tools

The NLP Suite is not designed for statistical analyses. It provides very basic statistical tools either for numeric data in a csv file or basic statistics for your corpus (e.g., number of documents, sentences, words, n-grams computation). For more sophisticated analyses, please use R, or any other statistical package you are familiar with (e.g., SPSS or STATA) on any of the csv files generated as output by the various NLP scripts.

Statistics (csv & txt files)

Visualization Tools

The NLP Suite provides a variety of data visualization tools, from Excel charts, to maps in Google Earth Pro, network graphs in Gephi, word clouds in a variety of software options. Of special interest are the annotators that produce annotated html files from input txt files either via user dictionaries, knowledge-base systems such as DBpedia or YAGO, or the special gender annotator that annotates a document for male/female proper names and pronouns. **It should be noted that nearly all scripts produce in output not just csv files of results but also Excel charts based on those results.**

Annotators (dictionary, gender, DBpedia, YAGO)
Excel charts
GIS geocoder & Google Earth maps
Network graphs (Gephi)
Sentence visualization: Dependency tree viewer (png graphs)
Word clouds