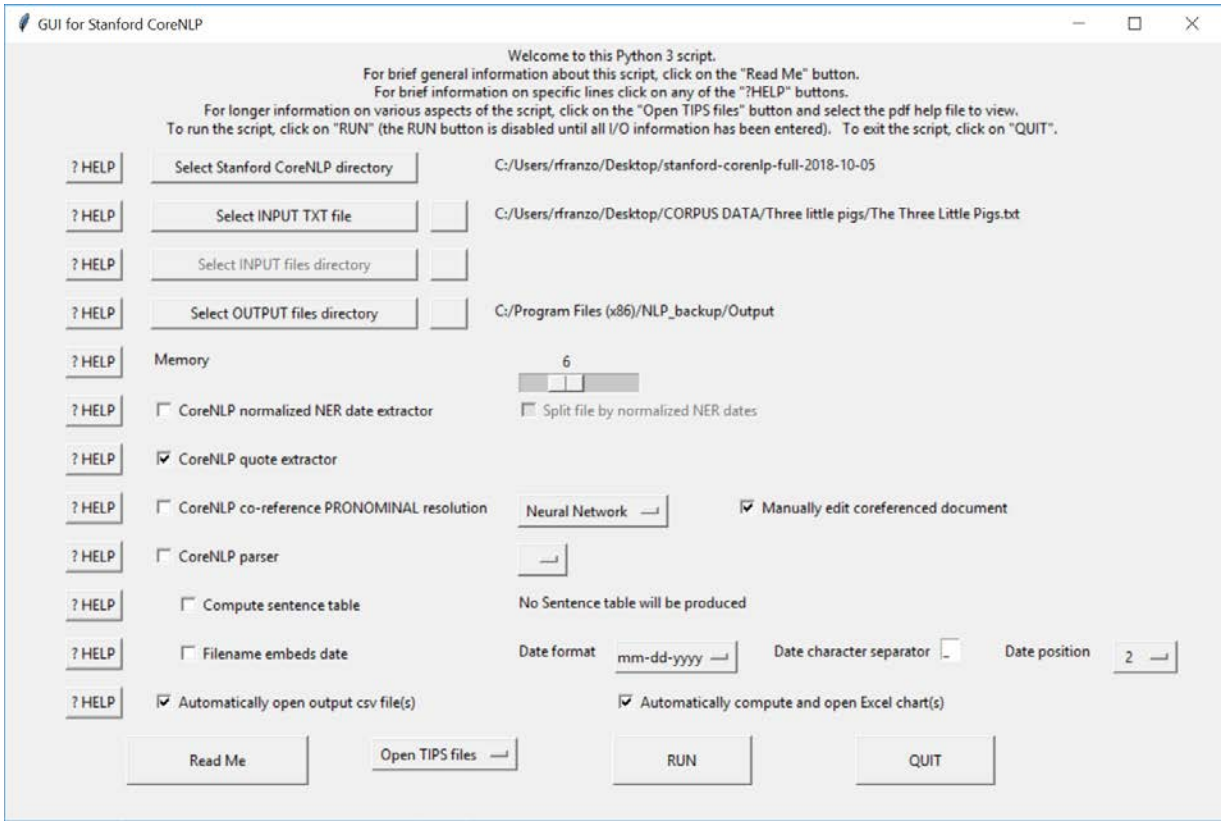
Stanford CoreNLP Date Extractor

The Stanford CoreNLP normalized date extractor



Stanford CoreNLP provides a way to extract a normalized date from a text (see Chang and

Manning, 2012).

What does a normalized date look like?

Let's take a look at what the temporal expressions used in Murphy's *Miracles Thicker Than Fog* look like.

words	Normalized date
29-Dec-44	12/29/1944
45 years ago	OFFSET P-45Y
a minute later	OFFSET PT1M
past	PAST_REF
now	PRESENT_REF
now	PRESENT_REF
Now	PRESENT_REF
now	PRESENT_REF
this day	THIS P1D
the day	THIS P1D
45 years since the day	THIS P1D OFFSET P45Y
Christmas	XXXX-12-25
Christmas	XXXX-12-25
Christmas Day	XXXX-12-25
the day after Christmas	XXXX-12-26
29-Dec	XXXX-12-29
winter	XXXX-WI
Tuesday	XXXX-WXX-2TMO

Thus, expressions such as Christmas or Christmas days are converted to XXXX-12-25 where the XXXX indicates that the year is unknown. Similarly, the day after Christmas is converted to XXXX-12-26. Any unknown year is converted to XXXX. And so is any unknown month, similarly converted to XX.

What can you use a normalized date for?

There are several things you can do with normalized dates. For one thing, you can split the file into subfiles by normalized date.

Splitting files by date

The Stanford CoreNLP GUI gives you the option to split a file by normalized dates. The script takes the text between dates and saves each chunk of text between normalized dates as separate files. **Not all normalized dates are used to split the text. Only complete dates or dates with an XXXX year.** For the Murphy dates, these would be the split files.

Output

Split files are saved in a subfolder of the INPUT directory with the name *split_files_byDate_Murphy* where the suffix Murphy is the input filename (without the txt file type).

Analyzing the subfiles separately with NLP tools

With text split into subfiles by date, you could analyze separately the different subfiles with any of the NLP tools to investigate, for instance, the appearance of different actors, actions, locations.

Studying narrative strategies: story and plot

When properly parsed, these normalized dates can be used to study a story in terms of the narrative difference between story and plot. And what are these?

In *From Words to Numbers* Franzosi (2004: 56) thus introduces the distinction between story and plot:

> It is the Russian formalists who introduced the distinction between story versus plot in narrative (*fabula* versus *sjužet*). Building upon Aristotle's idea of plot-structure – mythos, in the master's own words: "[B]y this term 'plot-structure' I mean the organisation of the events" – Tomashevski (1965, p. 67) wrote: "Plot is distinct from story. Both include the same events, but in the plot the events are arranged and connected according to the orderly sequence in which they were presented in the work. [Continued in a note; the story is] "the action itself, … [the plot] how the reader learns of the action." A story, in other words, refers to a skeletal description of the fundamental events in their natural logical and chronological order (perhaps, with an equally skeletal listing of the roles of the characters in the story).

A *story* then moves the events chronologically; a *plot* scrambles them for rhetorical effect, to keep the audience interested, like Hitchcock starting his movies from the murder scene and then moving back to the antecedents.

Normalized dates may give us automatic clues on how an author organizes a narrative.

Input

In INPUT the routine expects a text file or a set of text files in a directory.

Output

csv file

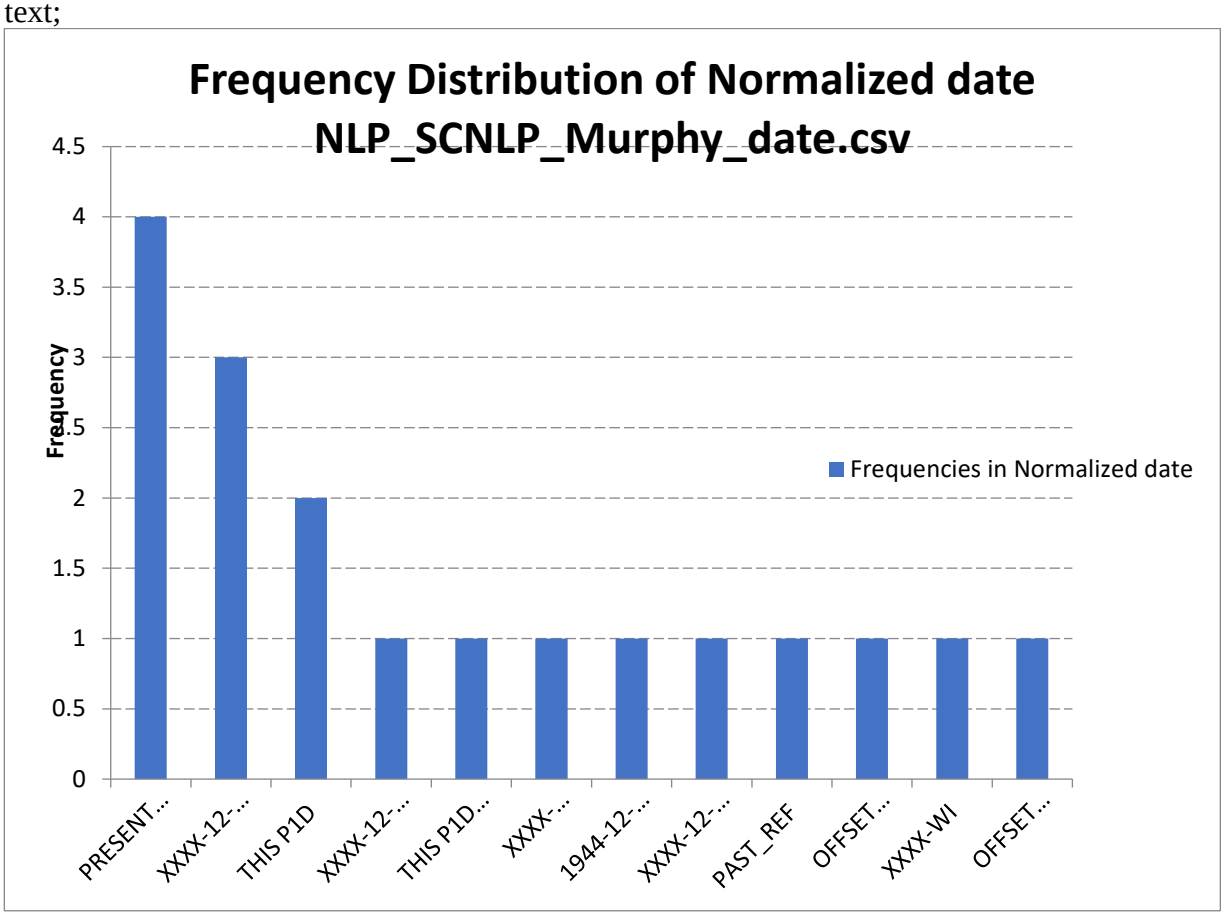
In output the script produces a csv file containing all normalized temporal expressions.

	A	B	C	D	E	F	G	H	I	J	K
1	words	Normalized date	SENTENCE	Sentence	DOCUMEN	DOCUMENT_NAME					
2	29-Dec-44	12/29/1944	33	It was Dec	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
3	45 years ago	OFFSET P-45Y	225	I was by r	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
4	a minute later	OFFSET PT1M	82	Half a mir	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
5	past	PAST_REF	71	The first a	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
6	now	PRESENT_REF	102	But now t	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
7	now	PRESENT_REF	131	We saw tl	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
8	Now	PRESENT_REF	142	Now we c	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
9	now	PRESENT_REF	157	It had bee	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
10	this day	THIS P1D	3	As I was v	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
11	the day	THIS P1D	226	Truly, this	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
12	45 years since the day	THIS P1D OFFSET P45Y	3	As I was v	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
13	Christmas	XXXX-12-25	4	The mirac	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
14	Christmas	XXXX-12-25	15	I recalled	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
15	Christmas Day	XXXX-12-25	19	I rememb	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
16	the day after Christmas	XXXX-12-26	2	It was the	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
17	29-Dec	XXXX-12-29	47	An extren	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
18	winter	XXXX-WI	147	We climb	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					
19	Tuesday	XXXX-WXX-2TMO	12	I rememb	1	C:/Users/Myself/Desktop/CORPUS DATA/MURPHY/Murphy.txt					

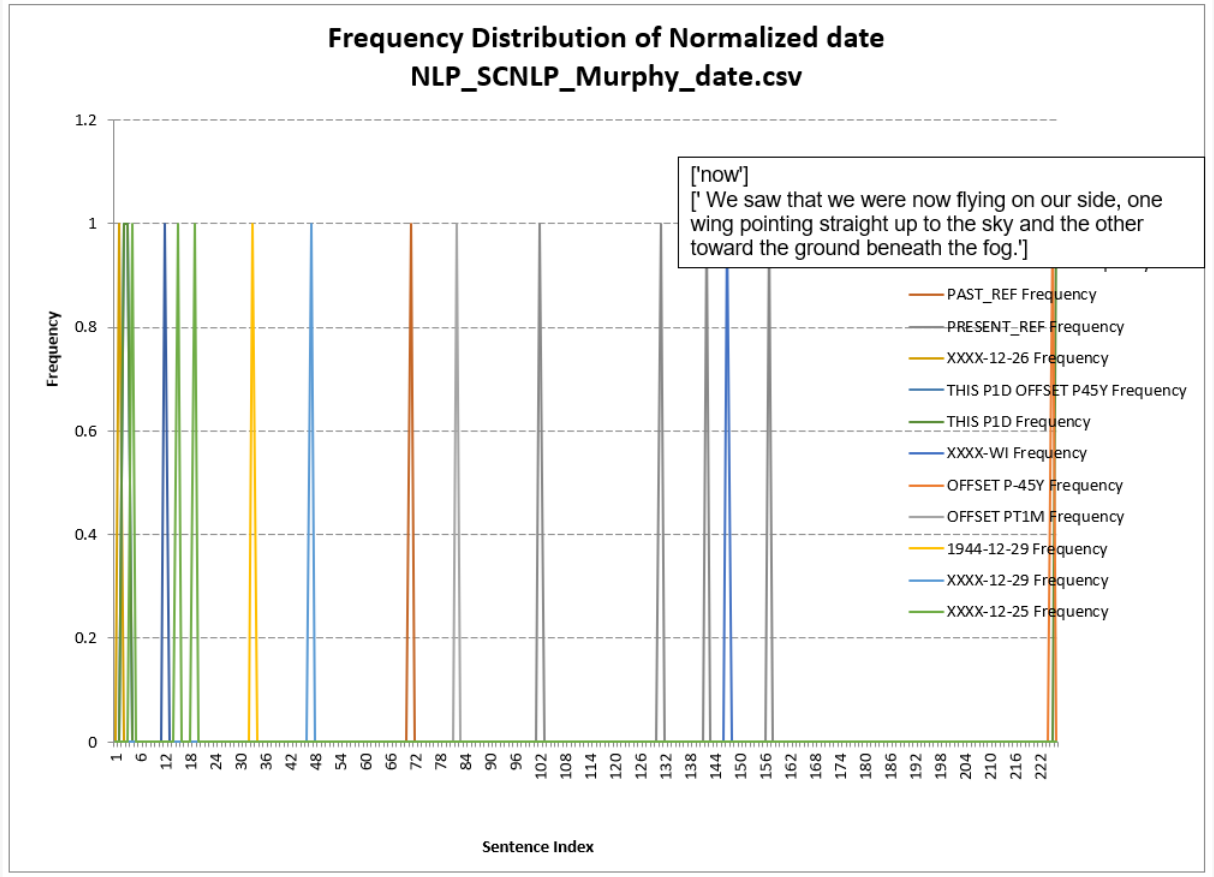
Excel charts

If the option of computing and opening Excel charts is ticked, the script will also produce in output two different types of charts:

1. a bar chart showing a frequency distribution of the temporal expressioons used in a text;



2. a line chart where temporal expressions are plotted as a function os sentence index; in other words, at what points in a text do certain temporal expessions appear?



References

Chang, Angel and Christopher Manning. 2012. "SUTIME: A Library for Recognizing and Normalizing Time Expressions." *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* Istanbul, Turkey.

Franzosi, Roberto. 2004. *From Words to Numbers: Narrative, Data, and Social Science*. Cambridge, UK: Cambridge University

TIPS_NLP_Stanford CoreNLP parser.pdf