Table of contents

Dates and the NLP Suite

The NLP Suite has several ways of extracting a date from files. BERT, spaCy, Stanford CoreNLP, and Stanza NER (Named Entity Recognition) annotators can extract a date from a text, the Stanford CoreNLP NER normalized date annotator can extract dates in ways that become comparable.

How to embed a date in the filename

You can also embed a date in the filename, such as the date of publication of the document (e.g., New York Time_01-15-1999_4_3, see Fig. 1 below). All you need to do is to separate the filename (e.g., New York Times) from its publication date (e.g., 01-15-1999) by a character separator of your choice (e.g., _ but could be $ or __ or @ as you prefer). The date itself can be in different formats of your choosing, such as dd/mm/yyyy or mm/dd/yyyy or mm/dd/yy…). Besides the date, you can embed other useful information in the filename, each item of information separated by the other items by the character separator (e.g., page number or column number).

Name	Date modified	Type	Size
Charlotte Daily Observer_09-23-1906_1_1.txt	2/8/2020 2:42 PM	TXT File	6 KB
Chicago Daily Tribune_09-23-1906_1_5.txt	2/8/2020 2:42 PM	TXT File	7 KB
Chicago Daily Tribune_09-24-1906_8_4.txt	2/8/2020 2:42 PM	TXT File	3 KB
Daily Press_09-23-1906_1_1.txt	2/8/2020 2:42 PM	TXT File	5 KB
Dallas Morning News_09-23-1906_2_1.txt	2/8/2020 2:42 PM	TXT File	1 KB
Dallas Morning News_09-23-1906_2_1_2.txt	2/8/2020 2:42 PM	TXT File	1 KB
Dallas Morning News_09-23-1906_2_1_3.txt	2/8/2020 2:42 PM	TXT File	1 KB
Dallas Morning News_09-23-1906_2_1_4.txt	2/8/2020 2:42 PM	TXT File	1 KB
Dallas Morning News_09-23-1906_2_1_5.txt	2/8/2020 2:42 PM	TXT File	1 KB
Dallas Morning News_09-23-1906_2_1_6.txt	2/8/2020 2:42 PM	TXT File	1 KB
Los Angeles Time_09-23-1906_2_4.txt	2/8/2020 2:42 PM	TXT File	4 KB
Los Angeles Times_09-24-1906_1_2.txt	2/8/2020 2:42 PM	TXT File	8 KB
Los Angeles Times_09-24-1906_6_1.txt	2/8/2020 2:42 PM	TXT File	1 KB
New York Times_09-23-1906_1_1.txt	2/8/2020 2:42 PM	TXT File	9 KB
New York Times_9-24-1906_2_5.txt	2/8/2020 2:42 PM	TXT File	11 KB

Figure 1 – Filenames with dates embedded in the filename

How the NLP Suite uses dates embedded in the filename

When the filename embeds a date several NLP Suite algorithms can use such metadata. Network graphs in Gephi (Fig. 3) and geographic maps in Google Earth Pro (Fig 4) can use the date to build dynamic networks and maps, i.e., networks and maps that change overtime as measured by the embedded date.
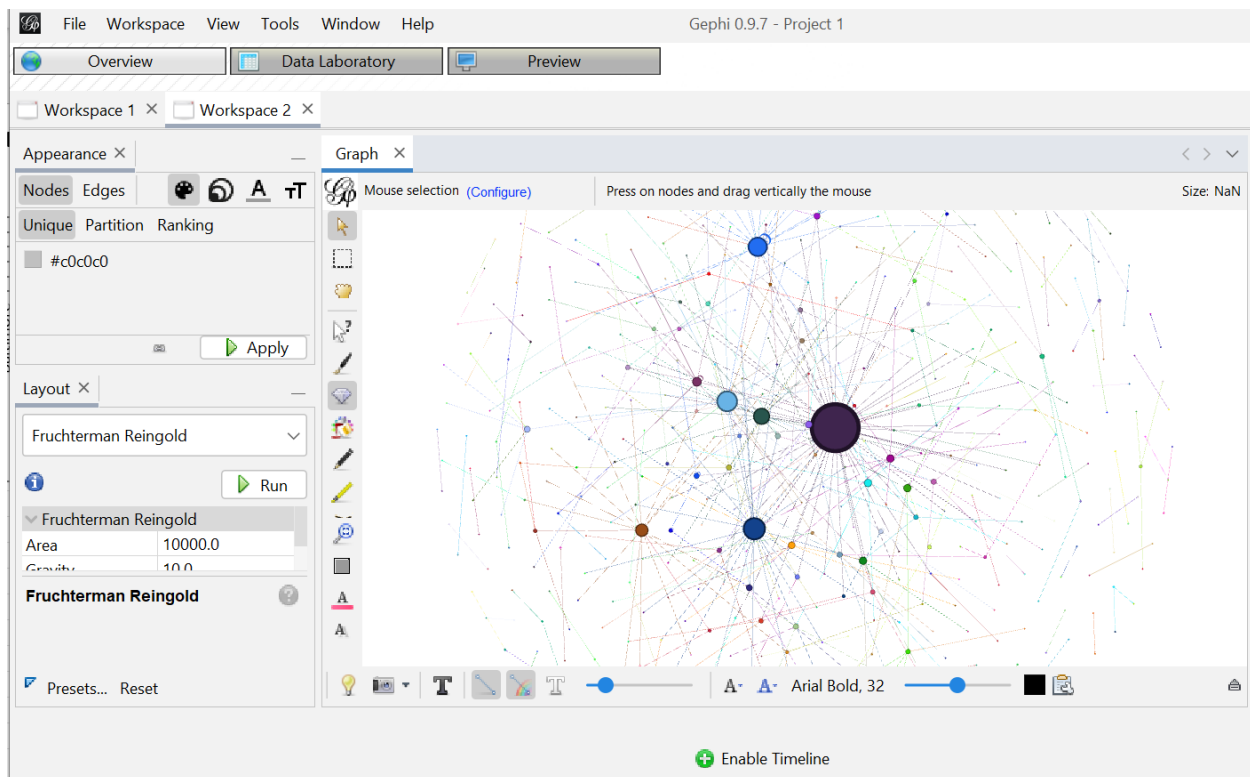

Figure 3 – Gephi dynamic network graph (as indicated by the Enable timeline widget at the bottom of the graph)
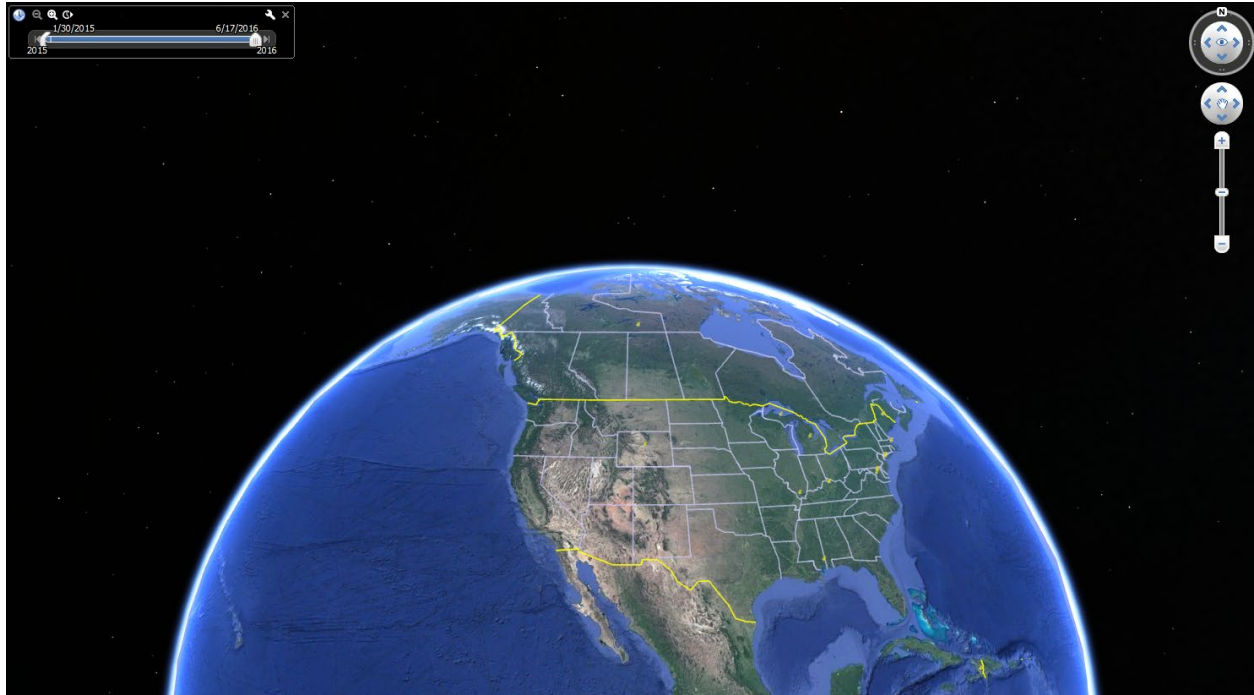
Figure 4 – Google Earth Pro dynamic map (as indicated by the slider widget on the top left-hand side of the map)

The NLP Suite Ngram VIEWER can build time plots for user-selected words similar to the Google Ngram Viewer (https://books.google.com/ngrams/) but applied to your specific corpus rather than Google books (see Figure 2 below). The Timechart visualization algorithm can use the embedded date to build dynamic time charts.
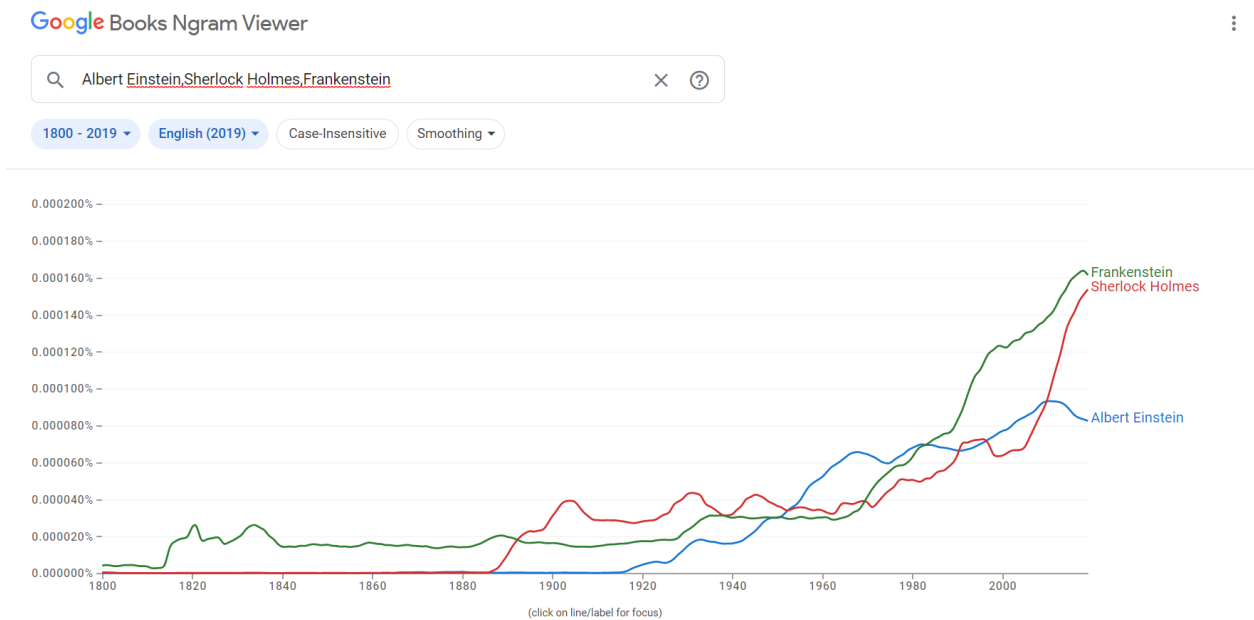

Figure 2 – Google Ngram Viewer