**Specialized visualization tools**

Contents

[**Introduction 1**](#_Toc131381850)

[**The Sunburst pie chart 1**](#_Toc131381851)

[**The Treemapper 2**](#_Toc131381852)

[**The Sankey flowchart 3**](#_Toc131381853)

[**The animated Timemapper 4**](#_Toc131381854)

# **Introduction**

This GUI contains specialized visualization tools, they are uncommon visualization tools with many customizable options to allow for a variety of analysis. You will find four PLOTLY visualization options: the Sunburst pie chart, the Treemapper, the Sankey flowchart, and the time mapper, an animated barchart where each frame corresponds to a jump in time based on the input file.

# **The Sunburst pie chart**

The Sunburst pie chart is used to separate data based on different categories. To select it, tick the “Visualize data in interactive Sunburster graph” box, the sunburst options will then be available to you.

In the Filename label/part widget, fill in the categories along which you want to split the data. For example, assume you are using the harry potter corpus and you would like to separate the data for each book, you will write the following: Book1, Book2, Book3, Book4, Book5, Book6, Book7. In general, you simply need to input a term that is included in the filename. For multiple terms, have them as comma-separated values.

Choose a variable from the CSV input file you would like to analyze in the CSV field widget. For useful data, we recommend choosing categorical data with a small number of different values, such as the sentiment label variable in sentiment analysis data.

You can choose the number of sentences from the start and from the end to conduct the sunburst graph on, or you can split each document in your corpus perfectly equally by ticking that box. You can also forgo this option by NOT splitting documents by ticking the final box.

This is an example of what the sunburst graph looks like (on the folktales corpus):

Chart, sunburst chart

Description automatically generated

The input for this graph in the filename label widget was the following: Arabian, Chinese, German, Russian, Indian, English

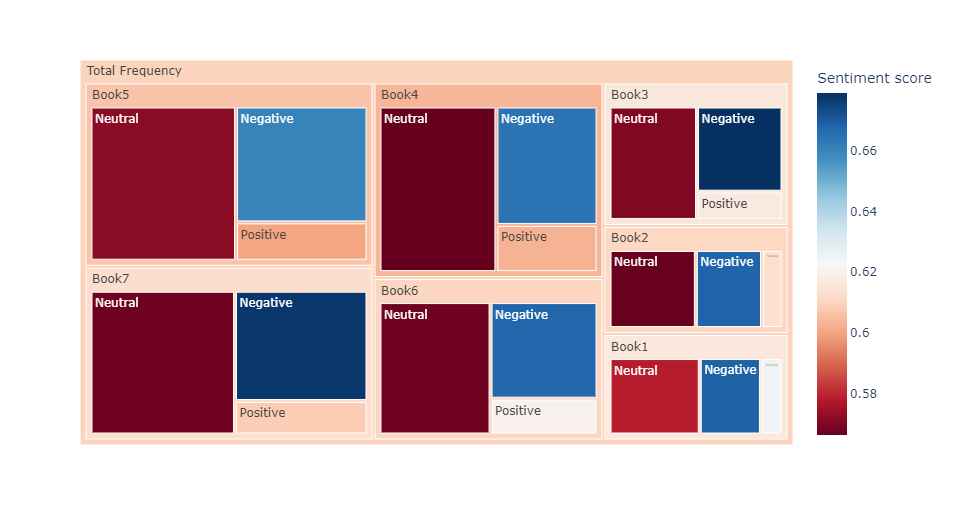
The inner layer of the graph the documents based on that input. The middle layer is dictated by the K-sentence options discussed above. The input is 1 sentence from the beginning and 1 from the end. As such, for each document, it only takes the first sentence and the last sentence. Finally, the outer layer is the CSV field choice. In this case, we chose sentiment label from sentiment analysis. Overall, this graph allows us to see the frequencies of the present emotions for the first and last sentence of all texts from each culture.

It is worth noting the graph, once outputted is interactive. You can run your mouse over the sections for more information and click on the sections to focus on a specific one. For example, clicking on the “Arabian” section would produce the following:

Chart, sunburst chart

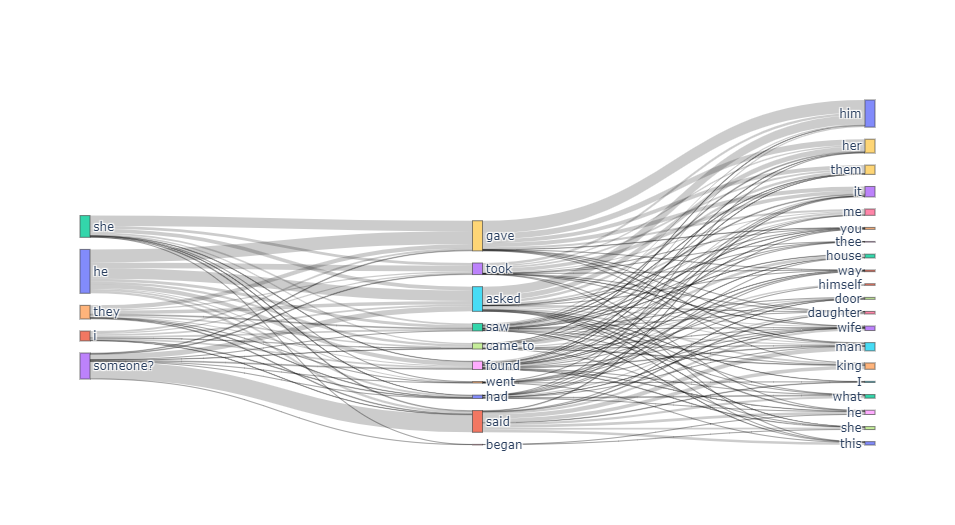
Description automatically generated

# **The Treemapper**

The Treemapper is very similar to the Sunburst pie chart, with the difference that it can also convey numerical data. The option to split the documents is not available for the treemapper. You still input filename labels to split the data, like the cultures in the sunburst example, and choose a CSV field, like sentiment label. Now, you have the option to choose to convey numerical data, and you can choose a second CSV field to do so. This will show you the average of that numerical data for every category in which we split the corpus. The following Treemapper example is one where we split the Harry Potter corpus by book, representing sentiment label, and where the numerical variable is sentiment score.

The color scale legend on the right tells you what color corresponds to what sentiment score. Much like the Sunburst pie chart, you have the option to click on the boxes to zoom into them.

# **The Sankey flowchart**

The Sankey flowchart allows you to visualize the logical flow between either two or three variables of your choice. The three variables should be categorical. However, you can choose them without worrying about the number of values the variables take, unlike the previous two options. You can choose a first CSV file field, click the + widget to add another one, and a third if you’d like to if the option to do so is selected. You also have a few options regarding how many of the most common values for each CSV file field you’d like to see: 5 or 10 for the first variable, 5, 10 or 20 for the second variable and 5, 10, 20 or 30 for the third variable. The options are limited to account for potential performance issues. Here is an example of a three variable Sankey flowchart, showing the top 5 subjects, the top 10 verbs and the top 20 objects and how they are linked to each other:

This graph is interactive. You can run your mouse over the boxes or the lines for more information. You can also move them by dragging them to simplify reading them.

# **The animated Timemapper**

The Timemapper is an animated bar chart, showing the evolution of the frequency of a chosen variable across time. For this to work, the inputted corpus must have documents in which dates are imbedded into the filenames. The filename-imbedded date can be in any format, you will simply have to choose the corresponding format in the Date Format widget. The Timeline widget will allow you to choose if you’d like to obtain the daily evolution of the data, the monthly evolution, or the yearly evolution. You can tick the cumulative box to show the cumulative evolution or not. Cumulative data means you see the frequency of a certain variable all the way up to that date. Non-cumulative data represents the frequency on the EXACT given date. Finally, you can choose the CSV file field that you would like to analyze through the timemapper. Once you obtain your output, the triangular button will automatically run through all the dates, the square button will stop the animation, and the cursor underneath the graph will allow you to manually go through the dates at your own liking. You can also run your mouse over the bars to see the current values of the frequencies. The following example is yearly cumulative data of the NER variable from the ConLL table:

