

Specialized visualization tools in Plotly

Contents

Introduction	1
The Sunburst pie chart	1
The Treemap chart	2
The Sankey flowchart	4

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 Treemap, the Sankey flowchart.

Check the other Visualization GUI for more PLOTLY visualization options: boxplots, multiple bar charts, and, and the time mapper, an animated bar chart 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 Sunburst 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.

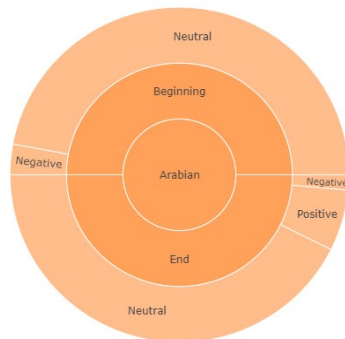
This is an example of what the sunburst graph looks like (on the folktale corpus):



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.

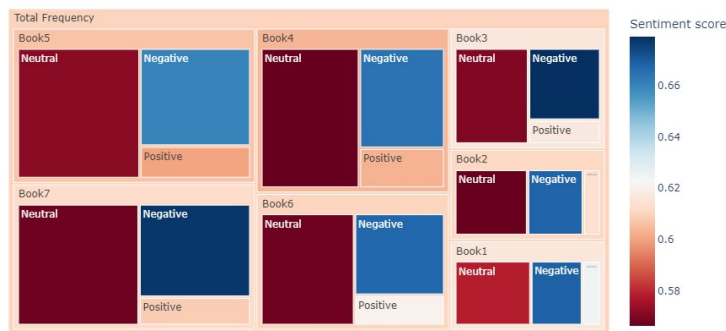
It is worth noting the graph, once outputted is interactive. You can run your mouse over the sections for more information and click on the sections to focus on a specific one. For example, clicking on the “Arabian” section would produce the following:



The Treemap chart

The Treemap chart 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 treemap. 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 Treemap 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.

There are many novel uses of the Treemap Tool. For the following example, the tool was used to analyze co-occurring adjectives for the word “government” in a POTUS State of the Union Corpus. The word “government” and its co-occurring adjectives was searched for using the CoNLL search function. With the Treemap Tool, the co-occurring adjectives were split by President. At the top of each same-colored group of boxes you can observe the name of a different president (Roosevelt, Taft, Tyler, Wilson, Taylor, Truman, Washington, and Trump). The individual boxes are the adjectives that co-occurred with that specific president’s usage of the word “government”.



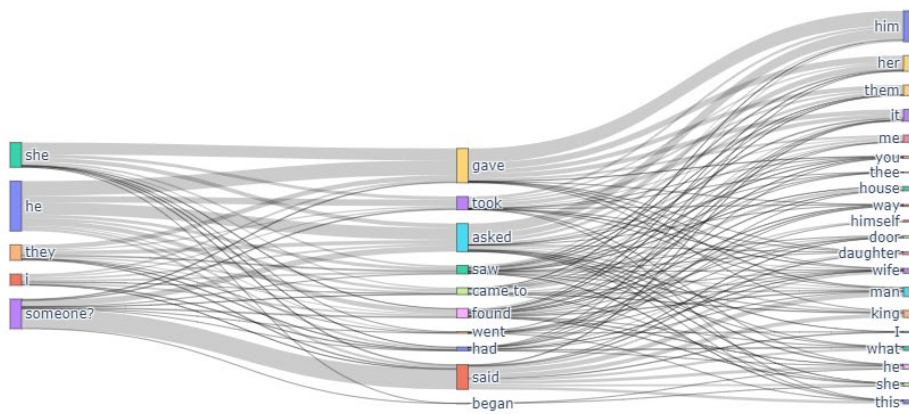
If your files encode date in their file name, you can also use the Treemap to separate by date. In this example, co-occurrences for “HIV” are examined in the HIV in Africa Corpus, which was collected in 2005, 2008, and 2014. The co-occurring verbs for the word “HIV” are graphed across the three years.



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.