Specialized visualization tools in Plotly

Contents

Introduction

This GUI contains specialized visualization tools, they are uncommon visualization tools with many customizable options to allow for a variety of analysis.

Check the other Visualization GUI for more PLOTLY visualization options: the Sunburst pie chart, the Treemap chart, and the Sankey flowchart.

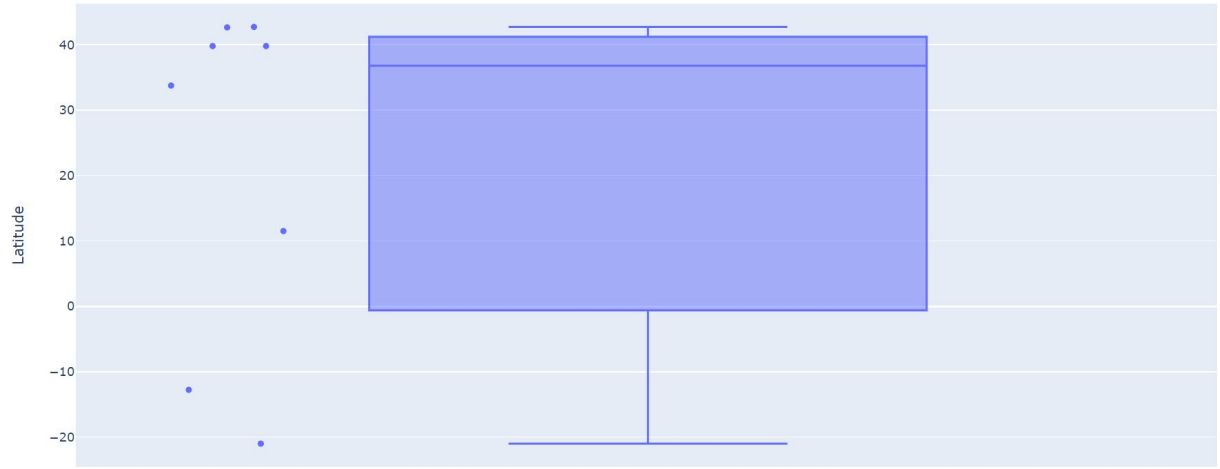
Boxplots

Boxplots are used to show distributions of **numeric** data values. They provide descriptive statistical information at a glance about a group of data's symmetry, skewness, variance, and outliers. Boxplots can display the **five-number summary** of a set of data: the minimum, first quartile, median, third quartile, and maximum. In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median.

The NLP Suite provides several options for Boxplots.

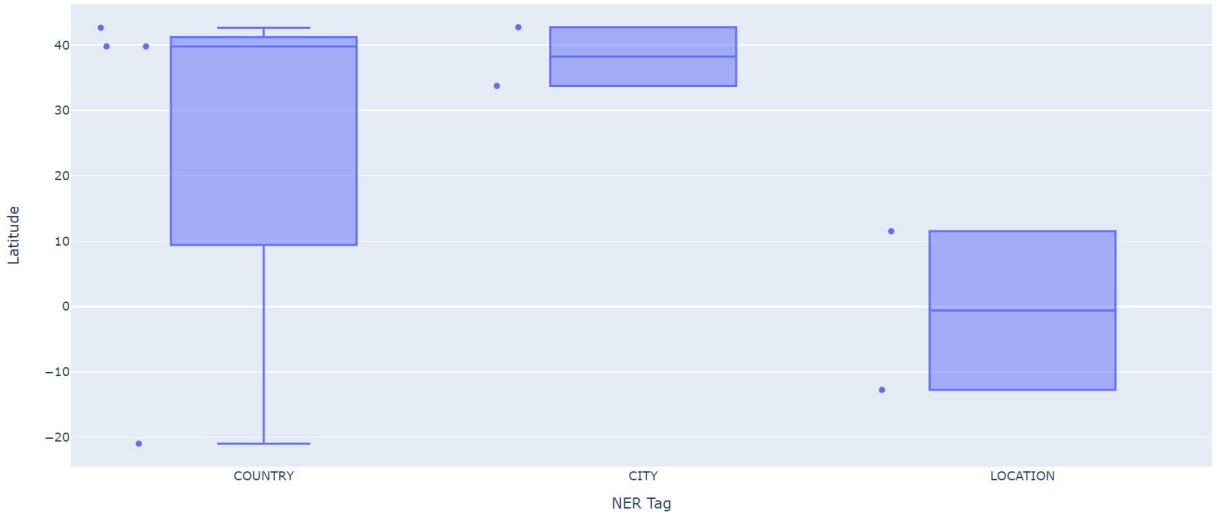
The Boxplot tool typically graphs each data point alongside a boxplot produced from the data. The result is a side-by-side frame with a scatter plot and a boxplot. However, this scatterplot can be adjusted under the Data Points widgets in the Boxplot parameters section. You will select "All points" most often, but you can select the "Outliers only" to examine the distribution and scope of data points that might be skewing your data. However, if you wish to only generate a boxplot and exclude the scatterplot with all the points, simply select the option for "No points" in the Data points widget.

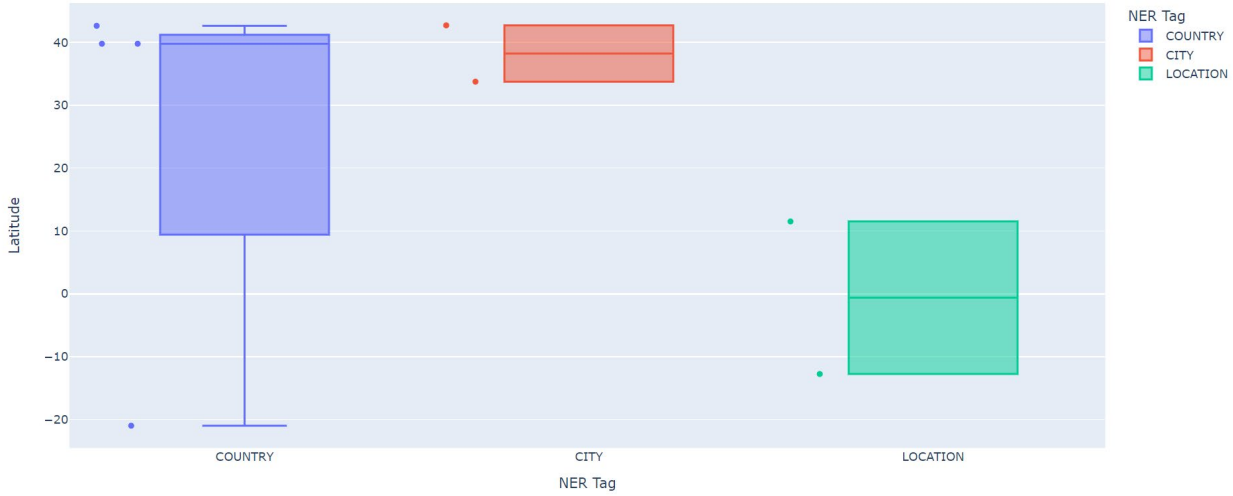
Suppose that you want a boxplot for the Latitude values you have obtained from running the GIS scripts.



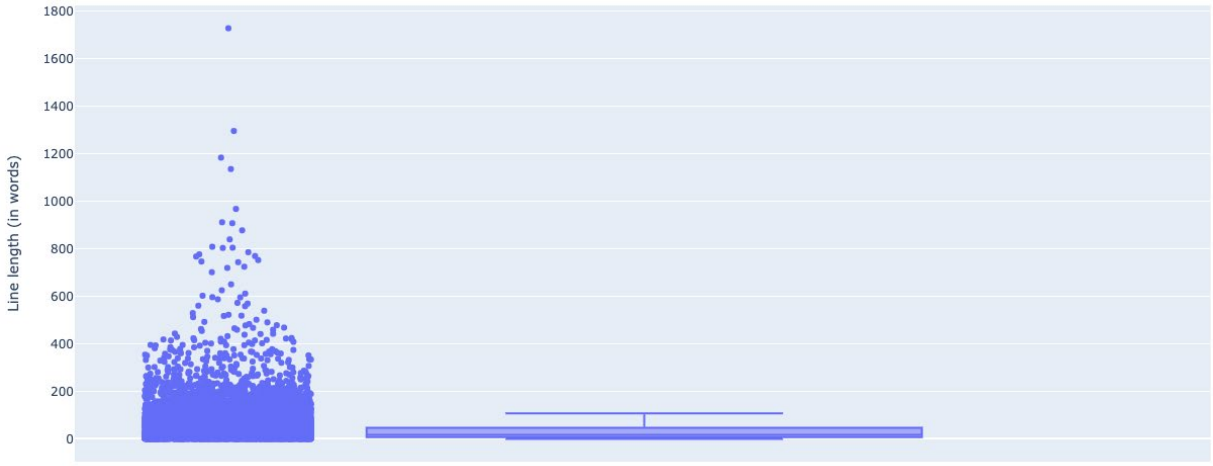
You can also option to split data by category, for instance, Latitude by type of NER location tags (e.g., COUNTRY, CITY).

By selecting “Split data by category,” you can group the boxplots by a selected CATEGORY field. By selecting the first extra CATEGORICAL csv file field, you can add in an X-axis on which the points will be graphed. By selecting a second extra CATEGORICAL csv file field, you can add color to your plots.

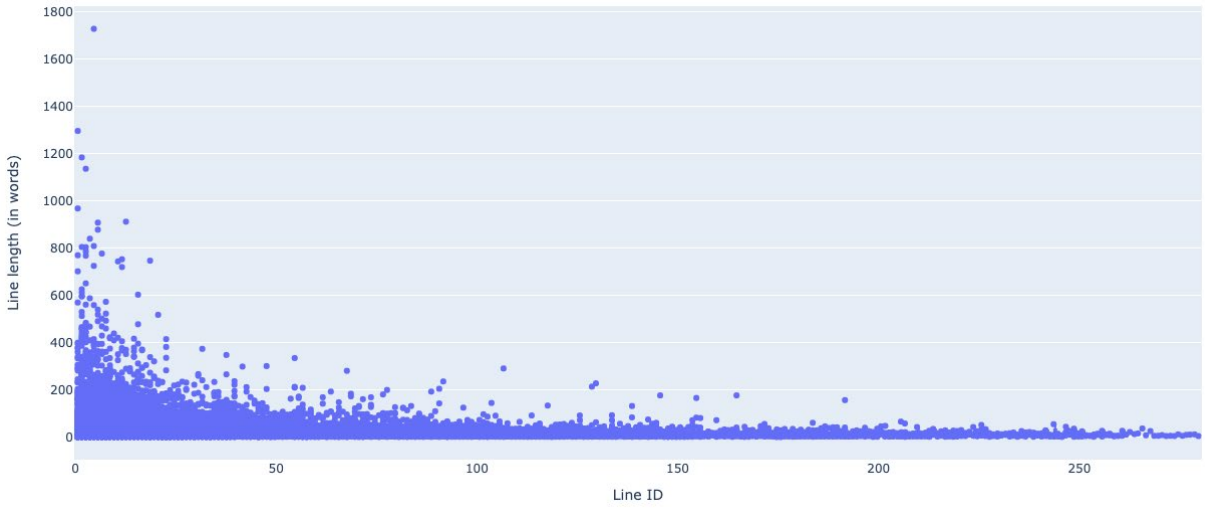




The following example compares line length (in words) in the HIV in Africa Corpus. All data points were graphed, and the data was not split by category.



The user can observe that there is a clear outlier. To identify the source of this outlier, you could split the data by Line ID.

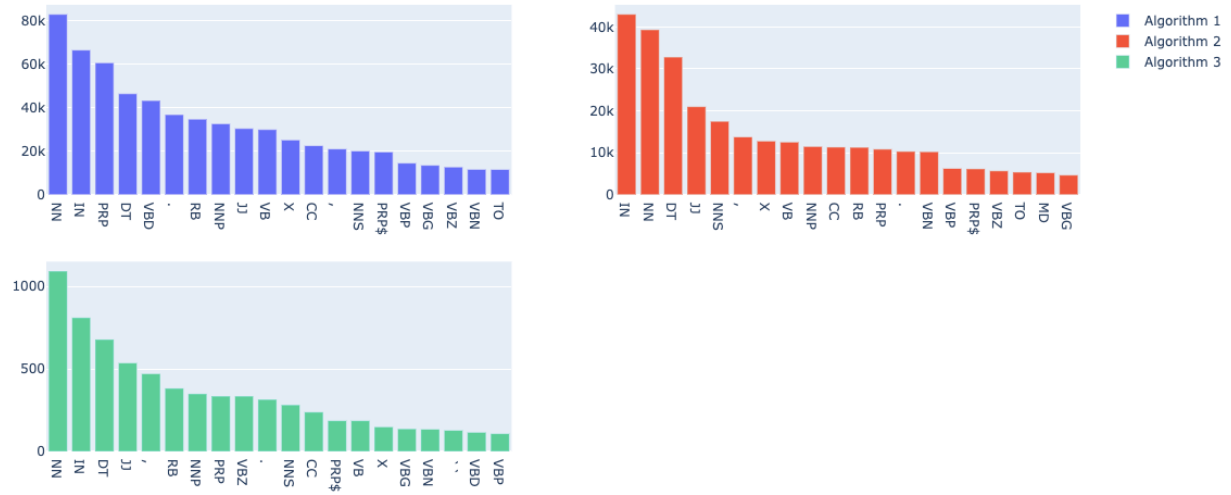


Multiple bar charts

The Multiple bar charts tool is used to compare different datasets across the same categorical measure. It does this by producing different bar charts in a side-by-side frame. To use this tool, the user must have at least two csv files that can be compared to each other. The csv files must have a data column, such as POS, that is shared between them and can be graphed on the bar charts for comparison purposes. You must first start by uploading one of the csv files. By clicking the csv file field for visualization widget, all the columns in your csv field will appear as an option to select, and you can choose what you want on the x-axis of your bar charts. Under the Visualization options widget, you can then click “Comparative bar charts.” After that, all you need to adjust is the Multiple Bar Charts Parameters section. This is where you can upload the other csv files that you wish to compare to your input csv. To do this, simply click the plus

button, which will lead you directly to your file finder. When your files are all selected, the tool can be run. The output is an html file. It is important to note that the tool will not automatically title the bar charts for you. Instead, the bar charts come with a key; each one is labeled "Algorithm 1, Algorithm 2" and so on. Algorithm 1 corresponds to the csv file that you added first, Algorithm 2 corresponds to the csv file that you added second, and so on.

There are many applications of this tool, such as comparing CoNLL tables from various datasets. The following example compares the CoNLL tables from a POTUS State of the Union corpus, the sample newspaper articles corpus, and an HIV in Africa corpus. To do this, the initial csv file was uploaded (HIV in Africa CoNLL table), and then the two other csv files were added under the Multiple Bar Charts Parameters Section. The order in which the graphs are graphed is related to the order in which the csv files are inputted. Thus, the Algorithm 1 bar chart that was produced corresponds to the HIV in Africa dataset.



Time mapper

The Time mapper 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: