Aggregate Data

Description

This algorithm aggregates/summarizes the input table based on values in a "Grouped On" column provided by the user. The types of aggregation performed on each column can be selected by the user from a drop-down box.

Currently "Sum", "Difference", "Average", "Min", "Max" aggregations are available for numerical column types.

Non-numerical column types are treated as Text and hence a user can select appropriate text delimiters for each. These text delimiters are used as a separator when appending values of 2 separate text table cell contents.

Pros & Cons

The Plugin has a very simple interface & provides basic aggregation function for numerical & non-numerical column types. In the future version we would be adding more functionalities so that Date & other data types can be handled better. Also more intelligent aggregation of Text content will be looked into.

Applications

The plugin can be used in workflows where aggregation of data is to be performed based on certain unique values. By grouping the rows based on values in a column we can essentially collapse the table. For example in a dataset where there is information about amount of grant money awarded, we can easily find out the total grant money awarded in each mainstream sciences like biology, chemistry etc just by grouping on *Science_Information* column.

Implementation Details

The user chooses which column should have its values "grouped together". Let's call this the "grouped" column. If two or more rows have the same value in the grouped column, we consider these rows to be duplicates of each other. Rows with duplicate values in the grouped column will be merged together to form one row. When we merge duplicate rows together, the values for each column will be aggregated together according to some function that the user selects (sum, count, etc...). The resulting merged row will have the original value in its "grouped" column, and aggregated values for all its other columns. A new column "Count" is added to the output table which contains frequency of the values in the "Grouped On" column.

Usage Hints

The user has to provide the following inputs; a file containing the data to be aggregated, name of the column on which the grouping is based off and type of aggregation for each column/attribute present in the original dataset. For Numerical columns user can select from None, Sum, Difference, Average, Min, Max aggregation options. The default option for this is "None". For Non-numerical columns the user can specify a text delimiter using which the data in that columns will be aggregated. The default text delimiter is "| ".

In case that user selects "None" aggregation function or empty text delimiter, that particular column will not be included in the output table. No aggregation will be performed on the column on which the grouping is based off of.

Links

- Source Code

Acknowledgments

The aggregate data plugin was authored, implemented, integrated and documented by Chintan Tank. Many thanks to Micah Linnemeier for providing guidance during the implementation phase of this plugin.

See Also