**Feature Selection Pipeline Parameters**

**Input Parameters:**

-ab *Use if you want to run an A/B comparison \*\* not completed*

-f1 First file to read in

Default: none

-f2 Second file to read in (to use for A/B comparison)

Default: none

-m Metric to use

Default: none

Valid Values: graph\_centrality, pca\_importance

-e Evaluation type to use

Default: none

Valid Values: kl\_divergence

-min Features with counts below this number will be removed

Default:3

-c Conditioning type to use on the data

default: none

Valid Values: add\_one, hellinger

-st Character value for number of features to select in total. Value should either be ‘all’ or a number value.

Default: none

-si Character value for number of features to select each time the

metric is called. Value should either be ‘all’ or a number value.

Default: none

-p Integer number of pca components to use (if using a pca metric)

default: 4

-sm Type of smoothing to be used to remove noise (if using pca

metric)

Default: sliding\_window

Valid Values: sliding\_window

-w If smoothing type is a sliding window, this is the size of the window

Default: 3

-cent If using the graph\_centrality metric, this specifies the type of centrality to use

Default: none

Valid Values: betweenness, closeness, degree, eigenvector

-th If graph\_centrality is the metric type, this is the threshold values to use to remove weak edges

Default: 0.5

-cor If graph\_centrality is the metric type, this is the type of correlation to use to build the graph

Default: none

Valid Values: spearman, pearson, kendall, MIC

-wt If graph\_centrality is the metric type, this specifies if the graph should be created as weighted or unweighted. If included, this parameter is set to true.

Default: unweighted

-cp If graph\_centrality it the metric type, the specifies if we should include only negative correlations, positive correlations, or both when creating the graph

Default: both

Valid Values: negative, positive, both

-pm Boolean value for if the metric plot should be created. If included, this parameter is set to true.

Default: False

-cg Boolean value for if the network graph should be created. Only applicable if graph\_centrality is the metric. If included, this parameter is set to true.

Default: False

-pp Boolean value for if the pca plot should be displayed. This scatter plot plots the first principle component against the second principle component. If included, this parameter is set to true.

Default: False

**Usage example:**

python pipeline.py -f1 bromeA\_all.csv -m graph\_centrality -e kl\_divergence -min 3 -c add\_one -st 25 -si 25 -cent betweenness –th 0.5 –cor spearman –wt –cp both –pm -cg

**Notes**:

This code is still quite rough and unfinished. There are a few parameters that aren’t applicable/useful at this point: evaluation, smoothing, window size, ab comparison and file2.

During our initial sensitivity analysis, we focused mainly on changing the centrality type, correlation type, total selected and number selected per iteration. We haven’t tested the pca\_importance metric yet, but it should run.