

# Clustering operator

This operator clusters a set of tuples.

## Parameter

- ATTRIBUTES: The attributes of the incoming tuples that should be recognized for clustering by the distance/similarity function. Notice, not all kinds of attribute types work here
- learner: The clustering algorithm that should be used
  - Currently implemented: kMeans, Weka (which in turn has further algorithms)
- ALGORITHM: A set of options to describe the algorithm

## Example

This example uses the weka-clusterer. The weka-clusterer should use the "simplekmeans" algorithm. the arguments to set up the weka-simplekmeans is "-N 3".

### Operator

```
clustered = CLUSTERING({
    attributes=['source'],
    learner='weka',
    algorithm='SimpleKMeans',
    options = [
        ['arguments', '-N 3 -I 500 -S 10 -O']]
    }, input)
```

For weka, there are currently the following algorithms that can be used as the "model". Further details and possible arguments can be found in the Weka Docs

- SIMPLEKMEANS (<http://weka.sourceforge.net/doc.dev/weka/clusterers/SimpleKMeans.html>)
- EM (<http://weka.sourceforge.net/doc.dev/weka/clusterers/EM.html>)
- COBWEB (<http://weka.sourceforge.net/doc.dev/weka/clusterers/Cobweb.html>)
- FARTHESTFIRST (<http://weka.sourceforge.net/doc.dev/weka/clusterers/FarthestFirst.html>)
- DENSITY\_KMEANS (<http://weka.sourceforge.net/doc.dev/weka/clusterers/MakeDensityBasedClusterer.html>)
- HIERARCHICAL (<http://weka.sourceforge.net/doc.dev/weka/clusterers/HierarchicalClusterer.html>)