## 1. Basic Settings

inputSampleName = r'S8' (Assumes sample is delimited file type .txt)
delimiter=','     (Specified the delimiter)
hasHeader = True or False      (True means first line is header and will skipped)
tokenizerType = 'Splitter' or 'Compress'

- If the method is "Compress," the non-word characters are replaced by a null character. For example, if a field has the value "123-456", then the after replacing the hyphen character with a null character is becomes the single string "123456".
- If the method is "Splitter," each non-word character is replaced by a blank character. The same example "123-456" becomes two strings (tokens), "123" and "345".

removeDuplicateTokens = True or False

If True, any duplicates of tokens within the same reference are removed, otherwise duplicates are left in the reference. In the end, the cleaned tokens from each reference are reassembled into a blank delimited string and written to the tokenized reference file.

## 2. Setting for global replacement DWM2 (runReplacement):

runReplacement = True or False (If True, it runs DWM2, otherwise this code is skipped)

These setting only apply if runReplacement = True
The "standard token" is the one kept, and "error token" is the token replaced by the standard token
minFreqStdToken = 5
The minimum frequency of a standard token
minLenStdToken = 3
The minimum string length of a standard token
maxFreqErrToken = 3
The maximum frequency of an error token

## 3. Fixed for All Iterations

muIterate = 0.1

The amount to increase match threshold (mu) each iteration. This must be a positive value. If and when mu become greater than 1.0, the iterations stop.

epsilonIterate = 0.0

The amount to increase or decrease the entropy threshold (epsilon) each iteration. We have usually set this value to 0.0 leaving the entropy threshold the same for each iteration. However, it could be either a positive or negative value to vary the threshold each iteration

runClusterMetrics = True or False

If True, this calculates the actual precision, recall, and F-measure of each cluster during each iteration. This allows you to compare the quality of the cluster to its entropy value. This can only be done when running on of the synthetic dataset where the truth set is available. Also, this slows the process considerably.

runFinalMetrics  = True or False

If True, this calculates the actual precision, recall, and F-measure of the final set of clusters. This can only be done when running on of the synthetic dataset where the truth set is available.

createFinalJoin = True or False

If True, it generates a join file (S8-ResultsFinalJoin.txt) of the format X:Y:Z where

X is the original Reference ID from the input source
Y is the cluster ID generated by the program
Z is the tokenized version of the original source record

If False, the program still produces a Link File (S8-ResultsLinkFile.txt) of the format X, Y where

X is the original Reference ID form the input source
Y is the cluster ID generated by the program

## 4. Cluster Robot Grid Search Settings

The robot will run the sample with the fixed settings listed above, but will

muStart = 0.60

The parameter mu is the match threshold for reference comparison. The value of muStart indicates the first value of mu in grid search. It must be a value between 0.0 and 1.0. This value is increased by the value of muIncr for the next grid search cycle.

muEnd = 0.70

The last starting value of mu in the grid search. Must be greater than muStart, but less or equal to 1.0

muIncr = 0.05

The amount to increase the starting match threshold (mu) in each grid search cycle.

betaStart = 14

The starting value of the blocking size parameter (beta) in the grid search. Must be a positive integer value.

betaEnd = 16

The ending value of the blocking size (beta) in the grid search. If betaEnd equals betaStart, the value of beta stays fixed for all grid search cycles.

betaIncr = 1

The amount to increment beta for the next grid search cycle.

sigmaStart = 143

The starting value of the stop word threshold (sigma) in the grid search. Must be a positive integer value.

sigmaEnd = 150

The ending stop word threshold (sigma) in the grid search. If sigmaEnd equals sigmaStart, the value of sigma stays fixed for all grid search cycles.

sigmaIncr = 1

The amount to increment beta for the next grid search cycle. Must be a positive integer value

epsilonStart = 35

The starting value of the entropy threshold (epsilon) in the grid search. Must be a positive value.

epsilonEnd = 36

The ending entropy threshold (epsilon) in the grid search. If epsilonEnd equals epsilonStart, the value of epsilon stays fixed for all grid search cycles.

sigmaIncr = 1

The amount to increment beta for the next grid search cycle. Must be a positive value