README

Requirements:

- Python Modules:
  - numpy
  - matplotlib
  - subprocess
  - shlex
  - os
  - sys
- Misc:
  - Weka
  - *weka.jar* must be added to the CLASSPATH environment variable

normalize.py:
To run: python3 normalize.py <countyData.arff>
The program reads the original dataset *countyData.arff* and creates a new file named *norm_countyData.arff* with normalized feature values between 0 and 1.

wekaCL.py:
To run: python3 wekaCL.py
This script automates WEKA through the command line using the *os.system()* and *subprocess.call()* commands. It can add *weka.jar* to the CLASSPATH environment variable if this has not already been done, and then gives the option to run three clustering algorithms:

1. Simple K-means:
   - Parameters:
     - numClusters = 2, 3, 4
     - distanceFunction: EuclideanDistance or ManhattanDistance
   - Output files:
     - with EuclideanDistance: *kmeansEuclid2.arff*, *kmeansEuclid3.arff*, *kmeansEuclid4.arff*
     - with ManhattanDistance: *kmeansMan2.arff, kmeansMan3.arff, kmeansMan4.arff*
2. Expectation Maximization:
   - Parameters: default
   - Output files: *EM.arff*
3. Hierarchical Clusterer:
   - Parameters:
     - numClusters = 2, 3, 4
     - distanceFunction: EuclideanDistance or ManhattanDistance
     - linkType: AVERAGE
   - Output files:
     - With EuclideanDistance: *hierarchicalEuclid2.arff, hierarchicalEuclid3.arff, hierarchicalEuclid4.arff*
     - With ManhattanDistance: *hierarchicalMan2.arff, hierarchicalMan3.arff, hierarchicalMan4.arff*

Note: It is recommended that weka.jar is added to the CLASSPATH environment variable manually rather than through the script, which has only been tested to work on the lab computers. The procedures for manually running the clustering algorithms in WEKA Explorer are provided below if there are issues with using wekaCL.py.

analyze.py:

To run: python3 analyze.py <result.arff>

This program reads a dataset file created by *wekaCL.py* and analyzes the clustering assignments. It determines the priorities by obtaining the average feature values of each cluster and uses these averages to calculate respective magnitudes. The results are displayed on a histogram, with darker red bars indicating higher priority. The clusters in the legend are ordered from highest priority to lowest priority.

STEPS TO RUN CLUSTERING ALGORITHMS MANUALLY IN WEKA:

1) Start the Explorer and load in the *norm_countyData.arff* dataset
2) In the Cluster tab, click on ignore attributes and select *county*
3) Choose one of the three clustering algorithms mentioned above with the specified parameters, then run the algorithm
4) To write the clustering results to an .arff file, go to the Preprocess tab
5) In the Filter section, click on Choose and select the *AddCluster* option under: *weka/filters/unsupervised/attribute/AddCluster*
6) Click on the options box to edit the *AddCluster* filter
7) In the cluster section, choose the algorithm that was run in step 3 with the same parameters
8) Close the editor and click apply; a new feature called *cluster* should appear under the existing features
9) Save the file; it should contain a new column of cluster assignments for each county (do not overwrite *norm_countyData.arff*)
10) If more algorithms will be run, click Undo to remove the cluster labels from the dataset