Poblete, Hector

(A04620462)

March 22, 2018

**Clustering thru K-Means**

**FORWARD**:

I might be a little obvious with my report, as I am not quite sure what to write. Please bear with me.

**-How to run:**

Make sure folder contains:

kmeans.py

\*fileWithValues\*.txt

\*.txt file must contain datasets in a format of:

x1 y1

x2 y2

x3 y3…. And so on.

To then run from zeus, command line:

~$ python kmeans.py \*fileWithValues.txt\* \*#of clusters\*

**ie: python kmeans.py values2.txt 3**

The program will run and then output a values\_output.txt file

The format of the output file will be

x1 y1 cluster#

x2 y2 cluster#

x3 y3 cluster#…. And so on.

**-Challenges:**

I have no python experience before this class. I have been trying to do my best with learning the language to conquer the tasks. Unfortunately, this has left my code looking quite disorganized. I made small programs to test simple function implementations such as printing to file and reading from file. These are included in an “misc” folder for documentation. Also, in this folder I have a program that outputs to console as to make troubleshooting the basic logic easier and quicker. I have not used linux too much and moving between the folders and opening and closing the text files is cumbersome for me.

The program seems to have some trouble dealing with smaller numbers of clusters, and in some testing, there were strange clusterings that led to clusters within clusters. Was not able to pin down the error. There is very little input validation. Such as errors if there are extra lines at the end of txt files.

I have a bad habit of not did not commenting my code and realized quite late, as I may have finished debugging in class.

**-Random information:** (that I didn’t know where to put)

Program was created completely through SSH into Zeus from my Chromebook and desktop at home. Entire program was written directly into the folder with VIM because I don’t really know better.

My program has some commented out functions that were leftover from before when I was making the output in the format:

Initial data sets used as centroids:

Data set

Data set

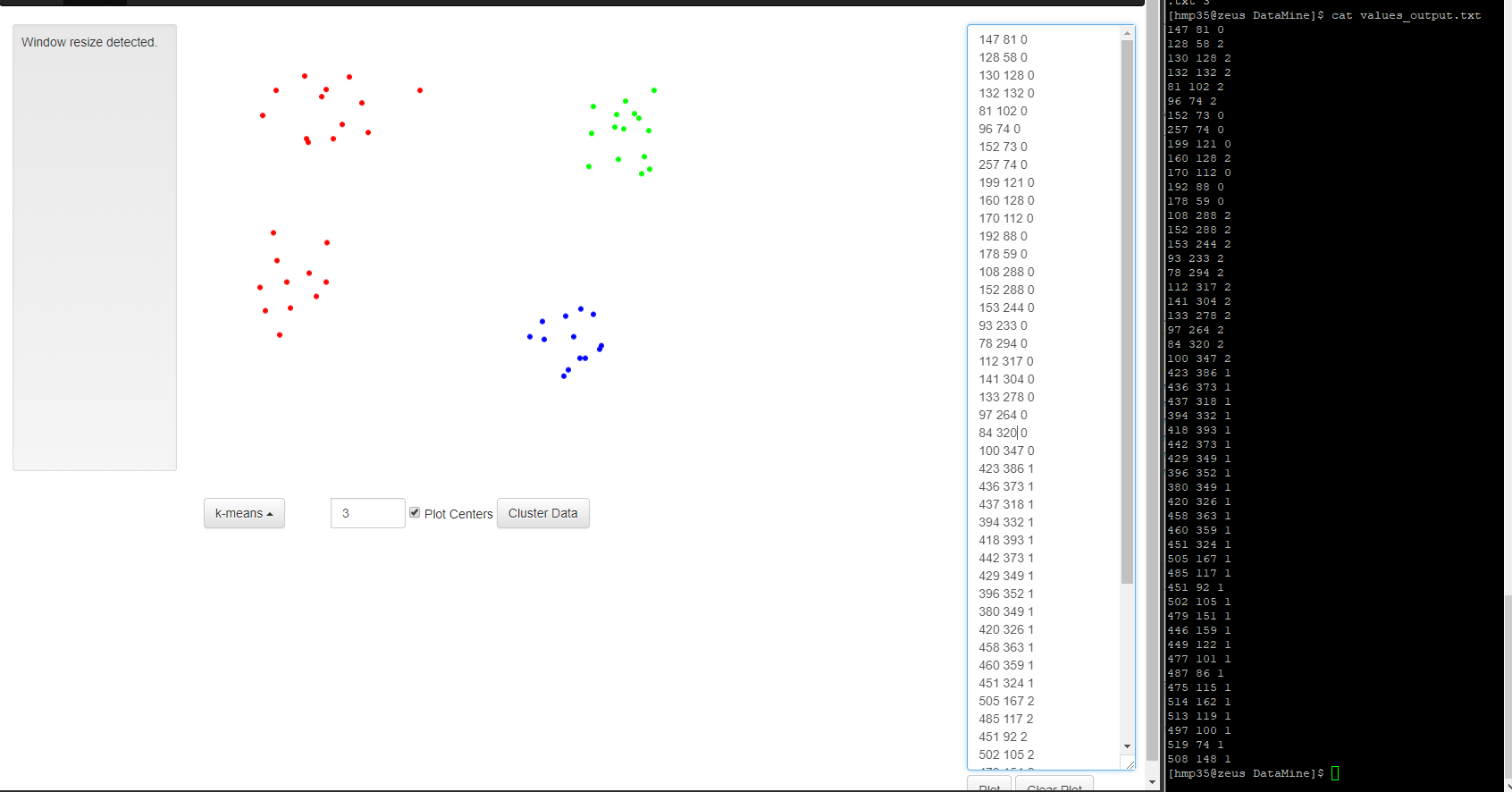
Cluster 0:

Data sets in cluster 0

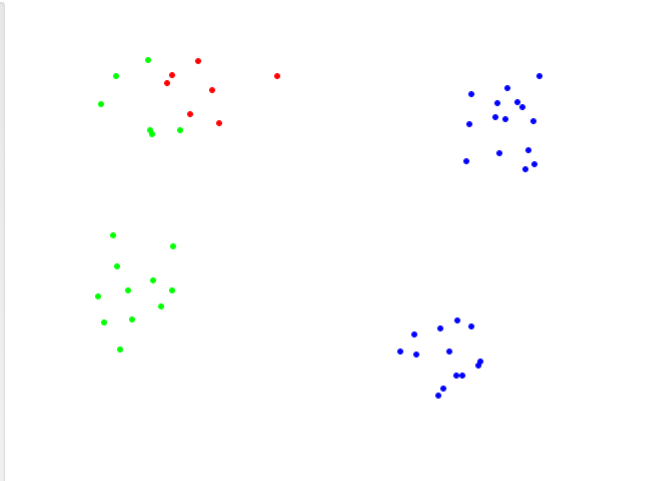
Cluster 1:

Data sets in cluster 1

**-Screenshots:** of a comparison of clustering of values2.txt on clusteringdemos site and the output of my kmeans clustering project.



A correct clustering of input file & my program output on right



Plot of my output file