CSI 531 HW 4 :

I have made 2 folders for Q-1,2,3 and Q-4 .

Q-1,2,3 contains answers for q1,2 and 3 in a pdf file.

For Q-4,

I have taken two different data sets, One contains tweets (collected in hw3) with its sentiment value in an excel sheet.

And another from UIC, The file name "servo" in a csv format. The description of the data set is also included in "servo_info".

The word cloud in stored in an image which I made using my tweets collected in hw 3, Topic was crime related tweets.

The file "new book" is a data set of my tweets with their respective sentiment value. I collected 501 total tweets.

I used my twitter data set, for box plot and finding SD,MAD,MEAN,MEDIAN,MAX AND MIN

For others I used the "servo" data set.

Basic Observation of Q-4

1. Summary of stats: I used my tweet collection as a data set, and used sentiment values of each tweet to calculate mean, median, mode , SD etc. Since sentiment values are between -1 and 1, I got all stats between these values. And got median as 0 because there were too may 0 sentiment values .

2. Histogram: I used "servo" data set, and from that I counted occurrence of the vgain values. The vgain values are 1,2,3,4 and 5

Most occurred was :2

Least occurred: 3

3. Scatter Plot : Used "servo" as the data set, used pgain and class to plot the scatter plot.

Because values in pgain are frequent, and in class all values are distinct and varies from 1 to 7.1

y-axis – class

x-axis – pgain

Most scatter can be seen at below 2.0 in y axis.

4. Box plot : Used tweet collection for the box plot, since sentiment values are from -1 to 1.0 , it plot looks small. And apart from that it has so many values around 0, which can be seen from the plot.

5. Density Map: Used “servo” as data set, Map comparing values of pgain and vgain.

The red part shows the highest occurrence of the pgain value , which is 6 and The blue part shows the highest occurrence of the vgain value, which is 4.

6. Parallel co-ordinate plot: Used “servo” as the data set, used pgain,vgain and class as attributes. All the values lie between 0 to 6 .

7. Co relation matrix : I used “servo” as data set and displayed the matrix plot with all the numerical values in my data set ; which are pgain,vgain and class.

8. Word cloud: Used crime topic as input. As seen from the image, the majority of the keywords are attack,murder,crime etc.