CS550: Massive Data Mining and Learning Spring 2020
Problem Set 2
Due 11:59pm Monday, March 23, 2019

Only one late period is allowed for this homework (11:59pm Tuesday 3/24)

Submission Instructions

Assignment Submission: Include a signed agreement to the Honor Code with this assignment. Assignments are due at 11:59pm. All students must submit their homework via Sakai. Students can typeset or scan their homework. Students also need to include their code in the final submission zip file. Put all the code for a single question into a single file.

Late Day Policy: Each student will have a total of **two** free late days, and for each homework only one late day can be used. If a late day is used, the due date is 11:59pm on the next day.

Honor Code: Students may discuss and work on homework problems in groups. This is encouraged. However, each student must write down their solutions independently to show they understand the solution well enough in order to reconstruct it by themselves. Students should clearly mention the names of all the other students who were part of their discussion group. Using code or solutions obtained from the web is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code seriously and expect students to do the same.

Discussion Group (People with whom you discussed ideas used in your answers):

On-line or hardcopy documents used as part of your answers:

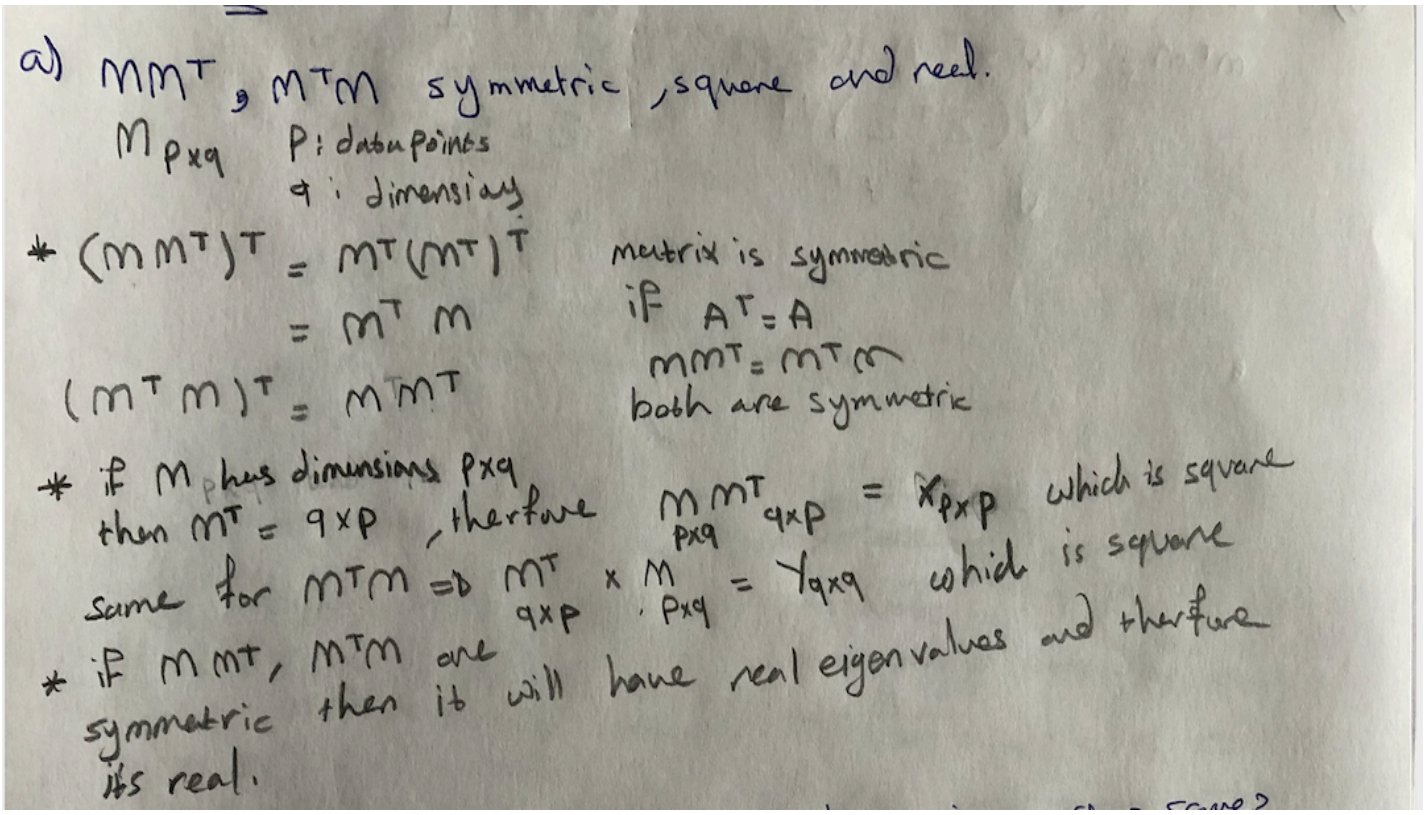
https://sites.cs.ucsb.edu/~veronika/MAE/parallelkmeansmapreduce_zhao.pdf
https://www.tutorialspoint.com/jfreechart/jfreechart_line_chart.htm

I acknowledge and accept the Honor Code.

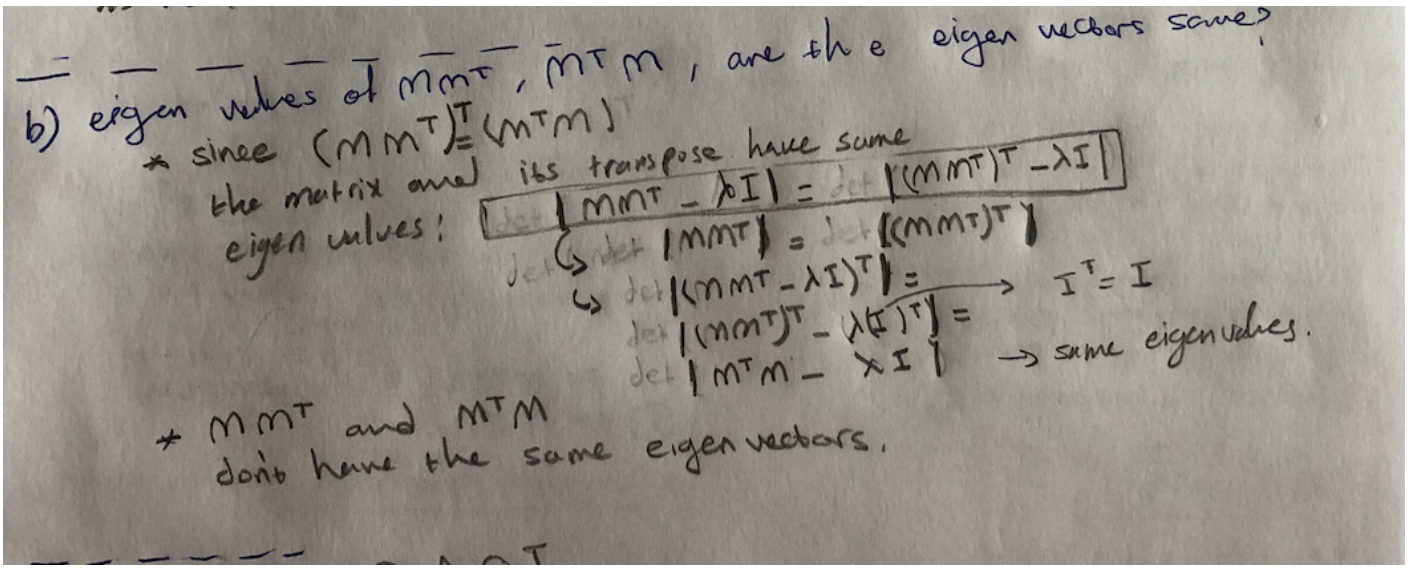
(Signed)__Fatima AlSaadeh_____

If you are not printing this document out, please type your initials above.

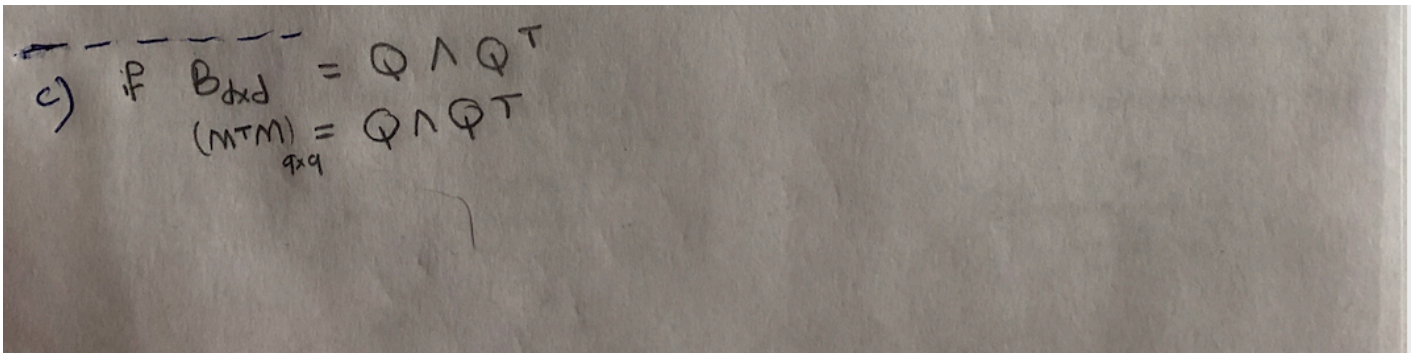
Answer to Question 1(a)



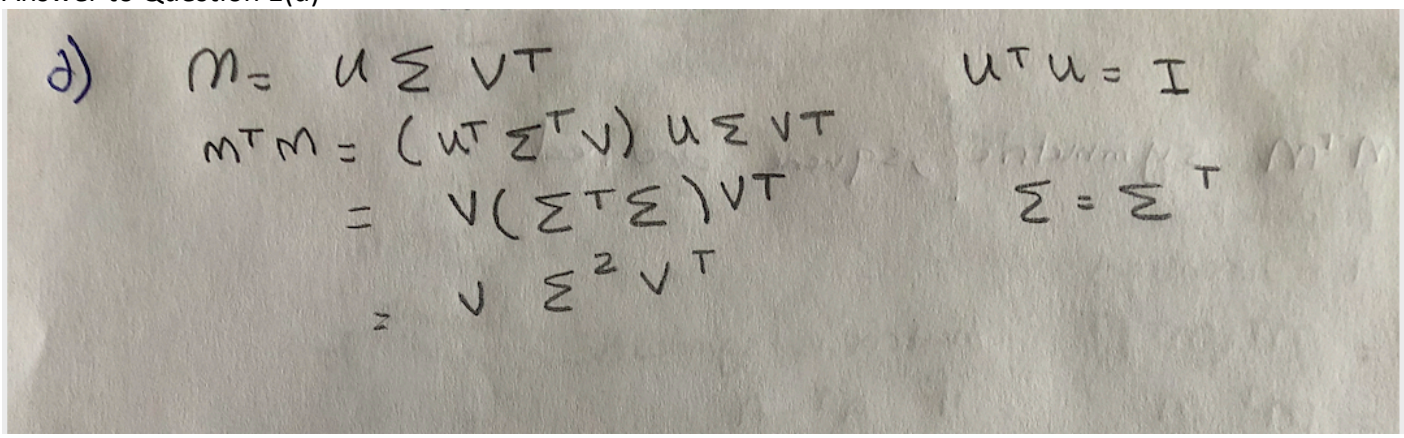
Answer to Question 1(b)



Answer to Question 1(c)



Answer to Question 1(d)



Answer to Question 1(e)(a)

U = [
[-0.27854301 0.5]
[-0.27854301 -0.5]
[-0.64993368 0.5]
[-0.64993368 -0.5]
]

Sigma= [7.61577311 1.41421356]

V^T = [
[-0.70710678 -0.70710678]
[-0.70710678 0.70710678]
]

Answer to Question 1(e)(b)

Sorted Evals=
[58. 2.]

Sorted Evecs=
[
[0.70710678 -0.70710678]
[0.70710678 0.70710678]
]

Answer to Question 1(e)(c)

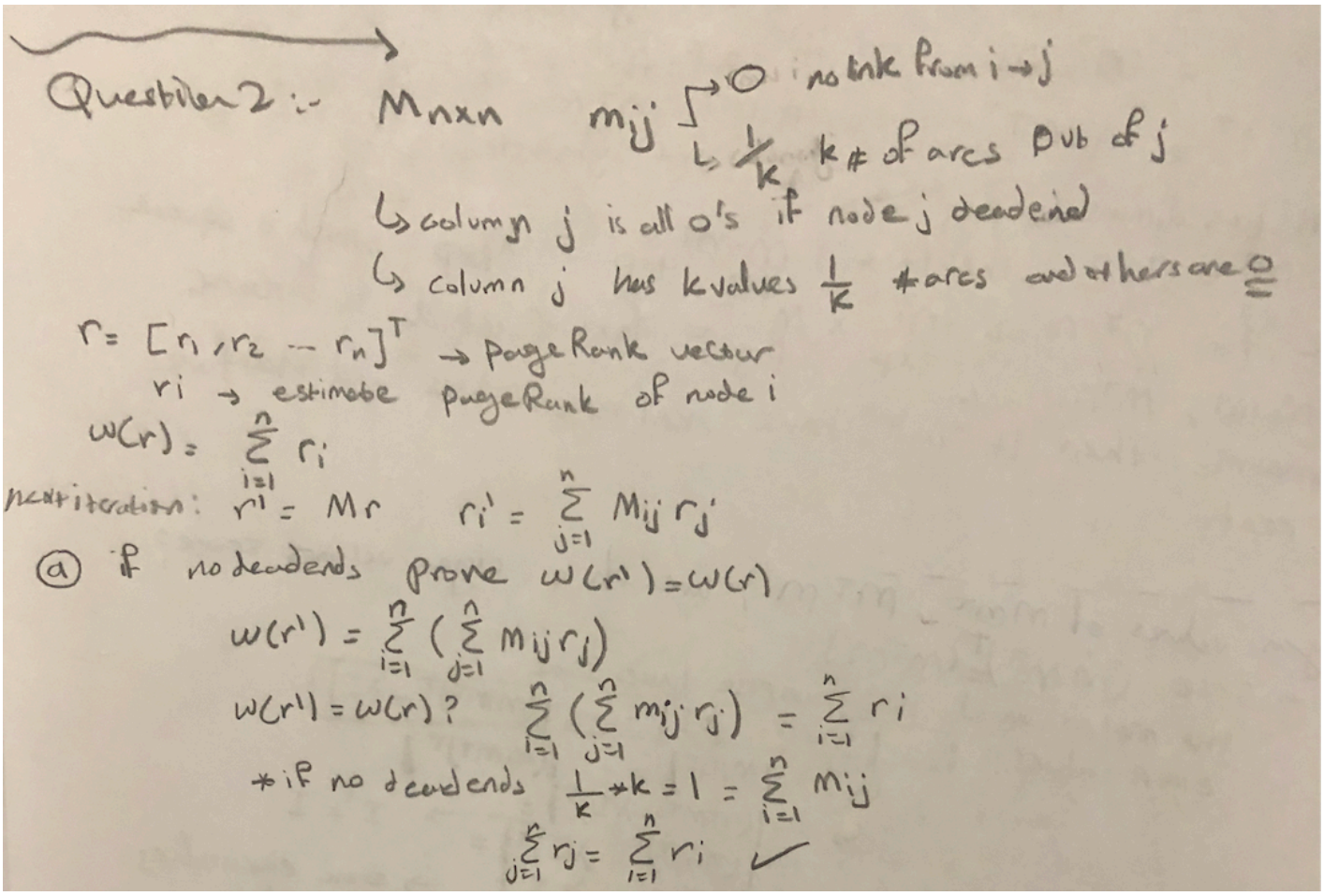
V= [
[-0.70710678 -0.70710678]
[-0.70710678 0.70710678]
]

In V the columns are the eigenvectors
first column is the eigen vector for eigen value 58 multiplied by -1
second column is the eigen vector for eigen value 2

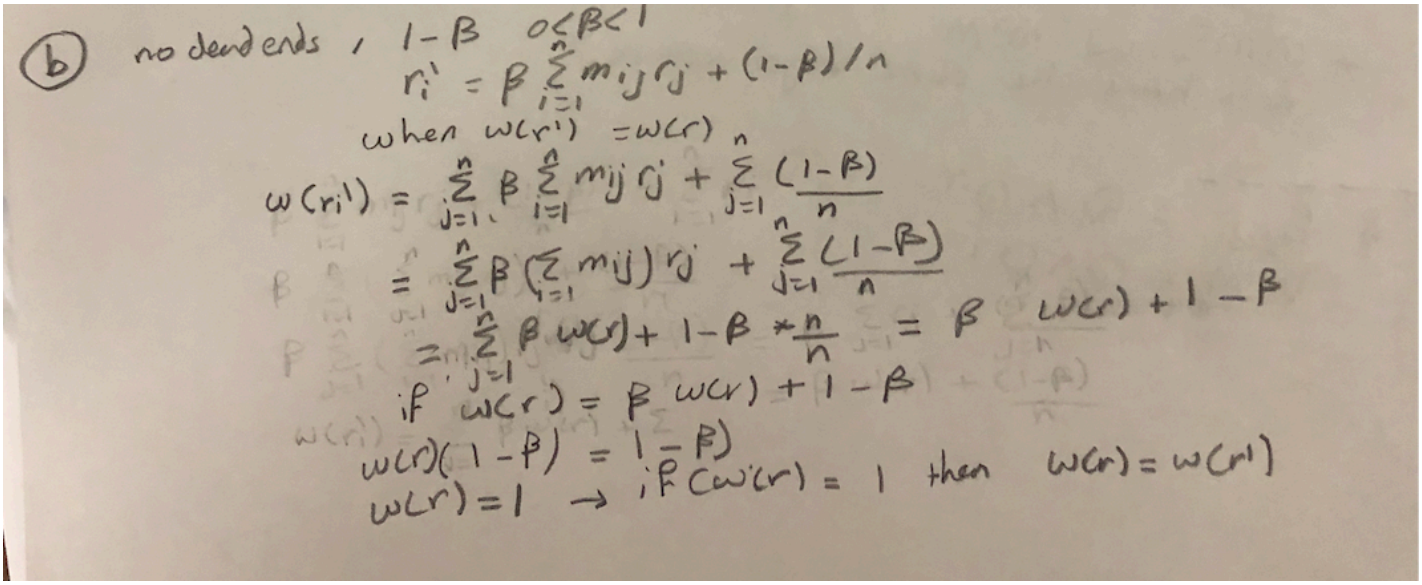
Answer to Question 1(e)(d)

sort_Evals = Sigma^2 = [58. 2.]

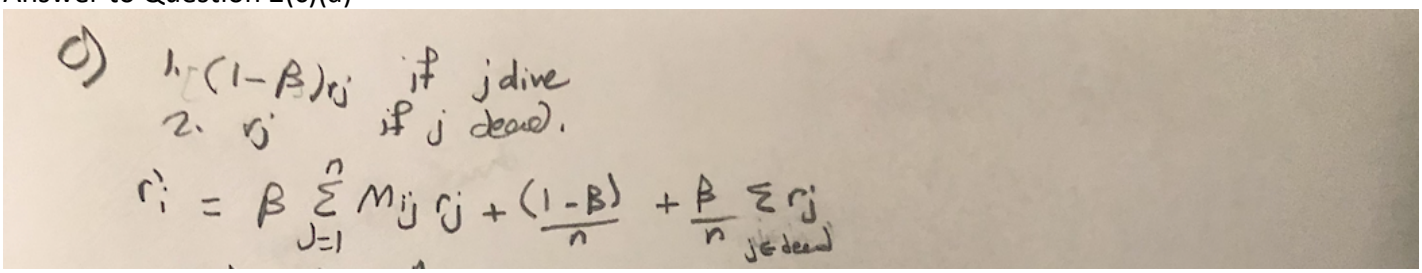
Answer to Question 2(a)



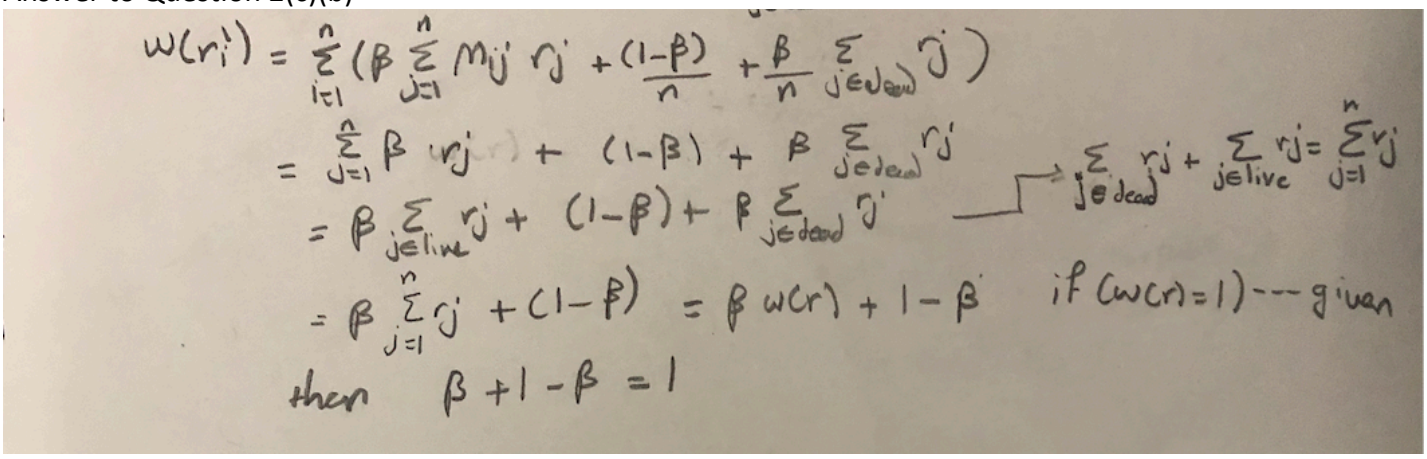
Answer to Question 2(b)



Answer to Question 2(c)(a)



Answer to Question 2(c)(b)



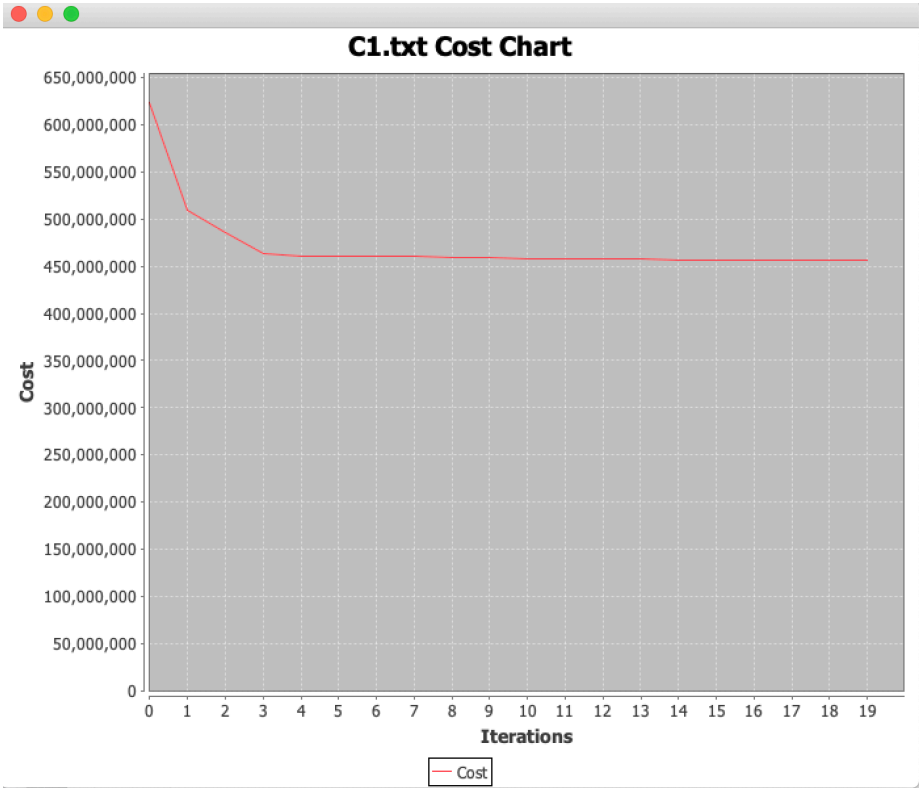
Answer to Question 3(a)

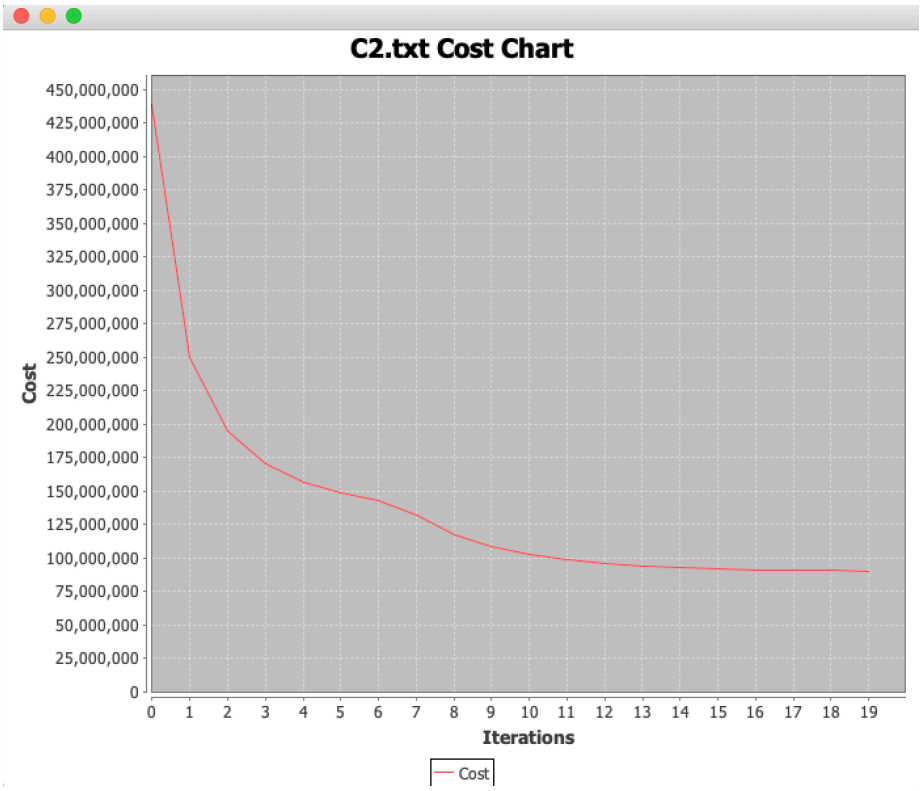
53	0.0357312022326716
14	0.034170906972591376
40	0.0336300871897439
1	0.030005979479788617
27	0.02972014420140539

Answer to Question 3(b)

85	0.0034096940774028216
59	0.003669860660127284
81	0.0036953517493609916
37	0.003808204291611451
89	0.003922466019802268

Answer to Question 4(a)





Answer to Question 4(b)
c1.txt:
percentage change in cost after 10 iterations: 26.398863292044183%
c2.txt :
percentage change in cost after 10 iterations: 75.25973243724756%

Random initialization isn't better as picking the centroids which are as far apart as possible is a better technique because it'll give us a better insight on where the points should be clustered resulting a minimum cost.