Φραγκαθούλας Χρήστος

Ραφαήλ Μουτζούρης

K-means Algorithm:

Η υλοποίηση του Kmeans έχει γίνει με τρεις κλάσεις.

Κλάση Cluster: Περιέχει τα σημεία που ανήκουν σε αυτό το cluster, το centroid του cluster, και ένα μοναδικό ακέραιο id για το cluster αυτό. Σημαντική μέθοδος σε αυτή τη κλάση είναι η compute SSE η οποία υπολογίζει για όλα τα σημεία του cluster τη απόστασή τους από το centroid χρησιμοποιώντας μία μέθοδο μιας άλλης κλάσης που θα δούμε στη συνέχεια.

Εάν κατά τη διάρκεια που τρέχει το πρόγραμμα επιθυμούμε να δούμε ποια σημεία ανήκουν σε ποιο cluster τότε βγάζουμε τα σχόλια που υπάρχουν στη μέθοδο printCluster.

Κλάση Point: Περιέχει τον ορισμό ενός σημείου στο χώρο. Έχει ως πεδία τις συντεταγμένες του σημείο καθώς και σε ποιο Cluster ανήκει το σημείο αυτό. Η μέθοδος distance είναι αυτή που έχει τη πράξη για την ευκλείδεια απόσταση δύο σημείων σε 2 διαστάσεις. Η μέθοδος producePoints είναι υπεύθυνη να κατασκευάσει ένα αρχείο με τυχαίες συντεταγμένες με βάση τα ορίσματά της και η μέθοδος loadPoints να φορτώσει αυτό το αρχείο και να δημιουργήσει τα σημεία.

Κλάση Kmeans: Είναι η βασική κλάση. Περιέχει ως πεδία της τον αριθμό των clusters που διμιουργούνται, το πλήθος των σημείων, τα όρια των συντεταγμένων αλλά και τα ίδια τα σημεία και τα clusters. Στην αρχικοποίηση της κλάσης με τη μέθοδο init, γίνεται μόνο δημιουργία νέου αρχείου με νέα random points(αν το επιθυμούμε, δηλαδή πρέπει να βγουν τα σχόλια αλλιώς θα χρησιμοποιεί το υπάρχων αρχείο point.txt). Στη συνέχεια φορτώνεται το αρχείο και γίνονται ανάθεση στα clusters τυχαία centroids από το πλήθος των point και παρουσιάζει την αρχική κατάσταση του αλγορίθμου.

Η μέθοδος assignPointsToClusters αναθέτει κάθε σημείο σε ένα Cluster. Για κάθε σημείο κρατάει βρίσκει σε πιο cluster έχει τη κοντινότερη απόσταση και ανατίθεται σε αυτό.

Η μέθοδος rellocateCentroids δημιουργεί νέο point και το αναθέτει ως centroid για κάθε Cluster χρησιμοποιώντας για συντεταγμένες το μέσο όρο των συντεταγμένων των σημείων που υπάρχουν ήδη στο Clyster.

Η μέθοδος computeSSE χρησιμοποιεί τις μεθόδους των δύο παραπάνω κλάσεων για τον υπολογισμό του αθροίσματος των

Η μέθοδος run είναι αυτή που ‘κινεί’ τον αλγόριθμο. Τρέχει τις δύο βασικές μεθόδους του αλγορίθμου έως ότου η απόσταση των centroids (πριν εκτελεστούν οι μέθοδοι assignPointsToClusters, rellocateCentroid) και των centroids αφού εκτελεστούν αυτές, να μην αλλάξει κατά 0.01.

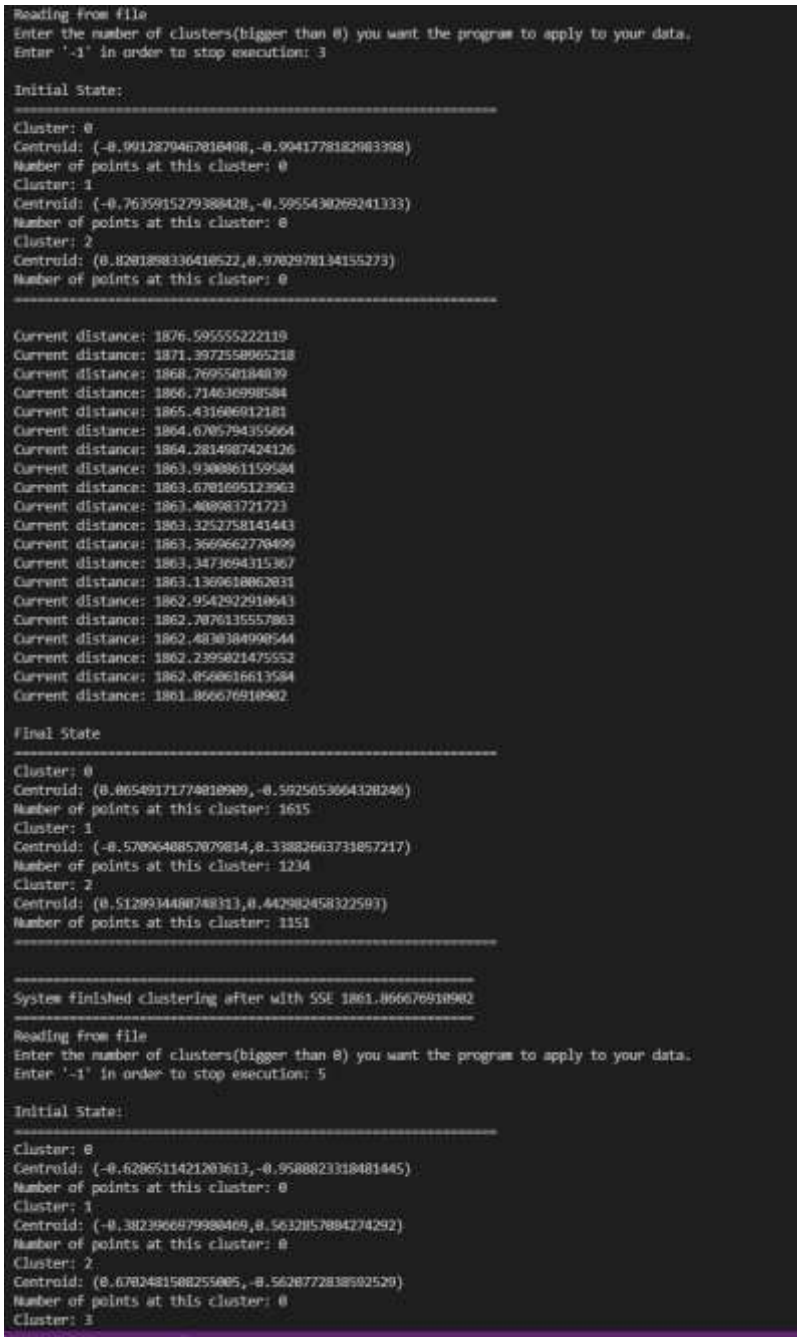
Η μέθοδος logClusters είναι μία μέθοδος που εκτελείται στο τέλος του αλγορίθμου και έχει σκοπό τη δημιουργία αρχείων με τη μορφή Clusters”αριθμός από Clusters””σε ποιο cluster είναι εκείνη τη στιγμή”. Με λίγα λόγια δημιουργεί αρχεία που απλά περιέχουν τα σημεία που ανατίθενται σε κάθε Cluster για κάθε αριθμό M που τρέχουμε το πρόγραμμα.

Η μέθοδος produceLogFile ανάλογα με τα ορίσματά της, δημιουργεί είτε ένα αρχείο που περιέχει τα SSE και τα M (πλήθος Cluster), είτε τα κέντρα των Clusters για κάθε M.

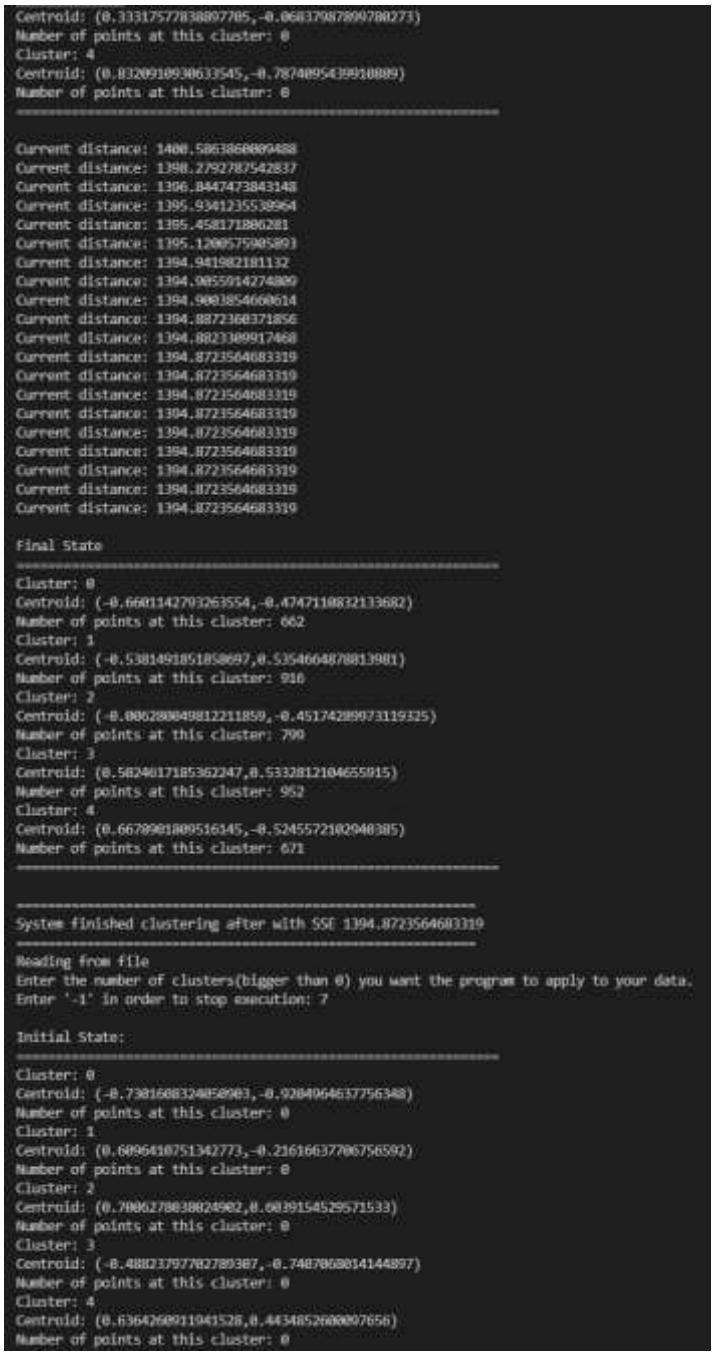
Η μέθοδος main έχει υλοποιηθεί δυναμικά ώστε να μπορεί ο χρήστης να ελέγξει για οποιοδήποτε αριθμό από Clusters το αποτέλεσμα. Για κάθε είσοδο του χρήστη ‘τρέχει’ ο αλγόριθμο 20 φορές και κρατάει το καλύτερο αποτέλεσμα. Στο τέλος καλούνται οι μέθοδοι που δημιουργούν τα αρχεία για να έχουμε και οπτικό αποτέλεσμα.

Οι παρακάτω 8 φωτογραφίες είναι τα αποτελέσματα του terminal για εισόδους: 3,5,7,9,11,13.

1



2



3

4

```
Centroid: (-0.39635777473449707,-0.3865245580673218)
Number of points at this cluster: 0
Cluster: 3
Centroid: (0.2771197557449341,0.7419818639755249)
Number of points at this cluster: 0
Cluster: 4
Centroid: (-0.7590712308883667,0.8163819313049316)
Number of points at this cluster: 0
Cluster: 5
Centroid: (0.03767716884613037,0.9549295902252197)
Number of points at this cluster: 0
Cluster: 6
Centroid: (-0.1043778657913208,-0.009638071060180664)
Number of points at this cluster: 0
Cluster: 7
Centroid: (0.16022086143493652,-0.8695952892303467)
Number of points at this cluster: 0
Cluster: 8
Centroid: (-0.12416934967041016,0.8281183242797852)
Number of points at this cluster: 0
============================================================

Current distance: 1004.8309304959926
Current distance: 1003.1127878061524
Current distance: 1002.1146350400471
Current distance: 1001.6404864860568
Current distance: 1001.2692006098255
Current distance: 1000.9479834070922
Current distance: 1000.8664840854645
Current distance: 1000.8536498917956
Current distance: 1000.8348406361688
Current distance: 1000.811590221497
Current distance: 1000.7823360699512
Current distance: 1000.778905349543
Current distance: 1000.7626801392506
Current distance: 1000.7513022205375
Current distance: 1000.7483365280248
Current distance: 1000.7659817182503
Current distance: 1000.7469051637394
Current distance: 1000.7324880798668
Current distance: 1000.7256047287398
Current distance: 1000.7256047287398

Final State
============================================================
Cluster: 0
Centroid: (-0.6414059653426661,-0.03907827716885191)
Number of points at this cluster: 462
Cluster: 1
Centroid: (0.6788513931219634,-0.02445264211211441)
Number of points at this cluster: 383
Cluster: 2
Centroid: (-0.6667485258966216,-0.6900458710355909)
Number of points at this cluster: 382
Cluster: 3
Centroid: (0.039308877271764416,-0.02588208142448874)
Number of points at this cluster: 425
Cluster: 4
Centroid: (-0.7014423887940902,0.650272059289715)
Number of points at this cluster: 474
Cluster: 5
Centroid: (0.6590552901124356,0.6741322553307442)
Number of points at this cluster: 478
Cluster: 6
Centroid: (-0.011804290042644869,-0.6719594983080281)
Number of points at this cluster: 509
Cluster: 7
Centroid: (0.6737522629842366,-0.6882588681020693)
Number of points at this cluster: 438
Cluster: 8
Centroid: (-0.012870615468524346,0.6370363947012376)
```


```
Centroid: (-0.39635777473449707,-0.3865245580673218)
Number of points at this cluster: 0
Cluster: 3
Centroid: (0.2771197557449341,0.7419818639755249)
Number of points at this cluster: 0
Cluster: 4
Centroid: (-0.7590712308883667,0.8163819313049316)
Number of points at this cluster: 0
Cluster: 5
Centroid: (0.03767716884613037,0.9549295902252197)
Number of points at this cluster: 0
Cluster: 6
Centroid: (-0.1043778657913208,-0.009638071060180664)
Number of points at this cluster: 0
Cluster: 7
Centroid: (0.16022086143493652,-0.8695952892303467)
Number of points at this cluster: 0
Cluster: 8
Centroid: (-0.12416934967041016,0.8281183242797852)
Number of points at this cluster: 0
============================================================

Current distance: 1004.8309304959926
Current distance: 1003.1127878061524
Current distance: 1002.1146350400471
Current distance: 1001.6404864860568
Current distance: 1001.2692006098255
Current distance: 1000.9479834070922
Current distance: 1000.8664840854645
Current distance: 1000.8536498917956
Current distance: 1000.8348406361688
Current distance: 1000.811590221497
Current distance: 1000.7823360699512
Current distance: 1000.778905349543
Current distance: 1000.7626801392506
Current distance: 1000.7513022205375
Current distance: 1000.7483365280248
Current distance: 1000.7659817182503
Current distance: 1000.7469051637394
Current distance: 1000.7324880798668
Current distance: 1000.7256047287398
Current distance: 1000.7256047287398

Final State
============================================================
Cluster: 0
Centroid: (-0.6414059653426661,-0.03907827716885191)
Number of points at this cluster: 462
Cluster: 1
Centroid: (0.6788513931219634,-0.02445264211211441)
Number of points at this cluster: 383
Cluster: 2
Centroid: (-0.6667485258966216,-0.6900458710355909)
Number of points at this cluster: 382
Cluster: 3
Centroid: (0.039308877271764416,-0.02588208142448874)
Number of points at this cluster: 425
Cluster: 4
Centroid: (-0.7014423887940902,0.650272059289715)
Number of points at this cluster: 474
Cluster: 5
Centroid: (0.6590552901124356,0.6741322553307442)
Number of points at this cluster: 478
Cluster: 6
Centroid: (-0.011804290042644869,-0.6719594983080281)
Number of points at this cluster: 509
Cluster: 7
Centroid: (0.6737522629842366,-0.6882588681020693)
Number of points at this cluster: 438
Cluster: 8
Centroid: (-0.012870615468524346,0.6370363947012376)
```

```
Number of points at this cluster: 449
============================================================


============================================================
System finished clustering after with SSE 1000.7256047287398
============================================================
Reading from file
Enter the number of clusters(bigger than 0) you want the program to apply to your data.
Enter '-1' in order to stop execution: 11

Initial State:
============================================================
Cluster: 0
Centroid: (0.5018041133880615,-0.30083680152893066)
Number of points at this cluster: 0
Cluster: 1
Centroid: (0.4665839672088623,-0.39106738567352295)
Number of points at this cluster: 0
Cluster: 2
Centroid: (-0.05644345283508301,0.05767810344696045)
Number of points at this cluster: 0
Cluster: 3
Centroid: (-0.16707539558410645,0.8330187797546387)
Number of points at this cluster: 0
Cluster: 4
Centroid: (-0.7368882894515991,-0.9723618030548096)
Number of points at this cluster: 0
Cluster: 5
Centroid: (0.6619813442230225,0.13296246528625488)
Number of points at this cluster: 0
Cluster: 6
Centroid: (0.3339540958404541,-0.927528977394104)
Number of points at this cluster: 0
Cluster: 7
Centroid: (0.4010787010192871,0.8591305017471313)
Number of points at this cluster: 0
Cluster: 8
Centroid: (0.31485949592590033,-0.9438143968582153)
Number of points at this cluster: 0
Cluster: 9
Centroid: (0.5512828826904297,-0.6786551475524902)
Number of points at this cluster: 0
Cluster: 10
Centroid: (-0.6647669076919556,-0.9293550252914429)
Number of points at this cluster: 0
============================================================

Current distance: 924.1921697946478
Current distance: 923.8882470442596
Current distance: 923.6422163327989
Current distance: 923.5511561563388
Current distance: 923.5406798965564
Current distance: 923.5366274888414
Current distance: 923.5281431858955
Current distance: 923.4787255542535
Current distance: 923.3995847700636
Current distance: 923.3578384011213
Current distance: 923.3008720104533
Current distance: 923.3085569261773
Current distance: 923.26528703719344
Current distance: 923.204914364743
Current distance: 923.1778391673888
Current distance: 923.0760053635482
Current distance: 922.8572951049988
Current distance: 922.4939666723192
Current distance: 921.9151490264053
Current distance: 921.2818330467643

Final State
============================================================
```

```
Cluster: 0
Centroid: (0.6651930818765625,-0.005012232211370234)
Number of points at this cluster: 367
Cluster: 1
Centroid: (0.03689453471683095,-0.05257668270366659)
Number of points at this cluster: 403
Cluster: 2
Centroid: (-0.2829315418633409,0.5482950519068173)
Number of points at this cluster: 313
Cluster: 3
Centroid: (-0.7540608876599902,0.6752282583371155)
Number of points at this cluster: 383
Cluster: 4
Centroid: (-0.7312871601837274,-0.6986978398219069)
Number of points at this cluster: 311
Cluster: 5
Centroid: (0.726358185338852,0.6690822648052096)
Number of points at this cluster: 391
Cluster: 6
Centroid: (0.2846140407708586,-0.7227122272763934)
Number of points at this cluster: 322
Cluster: 7
Centroid: (0.195252693436119,0.6611403990327642)
Number of points at this cluster: 356
Cluster: 8
Centroid: (-0.19540022388316435,-0.6309851725303118)
Number of points at this cluster: 402
Cluster: 9
Centroid: (0.7675469966093219,-0.6580619556305992)
Number of points at this cluster: 331
Cluster: 10
Centroid: (-0.6729876168266894,-0.8336348520605128)
Number of points at this cluster: 421
============================================================


============================================================
System finished clustering after with SSE 921.2818330467643
============================================================
Reading from file
Enter the number of clusters(bigger than 0) you want the program to apply to your data.
Enter '-1' in order to stop execution: 13

Initial State:
============================================================
Cluster: 0
Centroid: (-0.5647205114364624,-0.8244592428203974G)
Number of points at this cluster: 0
Cluster: 1
Centroid: (0.32155323028564453,-0.7541853189468384)
Number of points at this cluster: 0
Cluster: 2
Centroid: (0.6806519451141357,-0.9828335046768188)
Number of points at this cluster: 0
Cluster: 3
Centroid: (-0.5914775133132935,-0.8723737478256256)
Number of points at this cluster: 0
Cluster: 4
Centroid: (0.6974006513977051,0.341341495513916)
Number of points at this cluster: 0
Cluster: 5
Centroid: (-0.4790897369384765G,0.5150036811828613)
Number of points at this cluster: 0
Cluster: 6
Centroid: (0.5123419761657715,0.8712958097457886)
Number of points at this cluster: 0
Cluster: 7
Centroid: (-0.3621821403503418,-0.7902261018753052)
Number of points at this cluster: 0
Cluster: 8
Centroid: (-0.7391767501831055,-0.0926971435546875)
Number of points at this cluster: 0
```

7

```
Cluster: 9
Centroid: (0.06282436847686768,0.6887653636932373)
Number of points at this cluster: 0
Cluster: 10
Centroid: (0.4799131155014038,0.10227549076080322)
Number of points at this cluster: 0
Cluster: 11
Centroid: (-0.9315271377563477,-0.8782726526260376)
Number of points at this cluster: 0
Cluster: 12
Centroid: (0.28301310539245605,0.20629835128784418)
Number of points at this cluster: 0
============================================================

Current distance: 851.6737694197986
Current distance: 851.1317318969352
Current distance: 850.748443212269
Current distance: 850.4476252280949
Current distance: 850.1672500160273
Current distance: 850.0889805294648
Current distance: 849.8998987230426
Current distance: 849.8068706912217
Current distance: 849.7702257774468
Current distance: 849.7815998927831
Current distance: 849.7571667592241
Current distance: 849.7252711352443
Current distance: 849.694353628365
Current distance: 849.6793509008946
Current distance: 849.6654618563623
Current distance: 849.6654618563623
Current distance: 849.6654618563623
Current distance: 849.6654618563623
Current distance: 849.6654618563623
Current distance: 849.6654618563623

Final State
============================================================
Cluster: 0
Centroid: (-0.3526567332446575,0.33680485328659415)
Number of points at this cluster: 256
Cluster: 1
Centroid: (0.20616777221877852,-0.657710174848507)
Number of points at this cluster: 308
Cluster: 2
Centroid: (0.7206399975424985,-0.7593188927419162)
Number of points at this cluster: 301
Cluster: 3
Centroid: (-0.28173151754197623,-0.2005262883017663)
Number of points at this cluster: 294
Cluster: 4
Centroid: (0.7085477885553393,0.34535447419699077)
Number of points at this cluster: 295
Cluster: 5
Centroid: (-0.7251787397795751,0.6974468138582426)
Number of points at this cluster: 399
Cluster: 6
Centroid: (0.6659810718122896,0.8068593446651538)
Number of points at this cluster: 286
Cluster: 7
Centroid: (-0.21007476933300495,-0.7870845515280962)
Number of points at this cluster: 256
Cluster: 8
Centroid: (-0.7510526301684203,-0.02869726257559694)
Number of points at this cluster: 324
Cluster: 9
Centroid: (0.05635173583593894,0.6813263933802527)
Number of points at this cluster: 381
Cluster: 10
Centroid: (0.7019937601582757,-0.22655932656649885)
Number of points at this cluster: 290
Cluster: 11
```

8

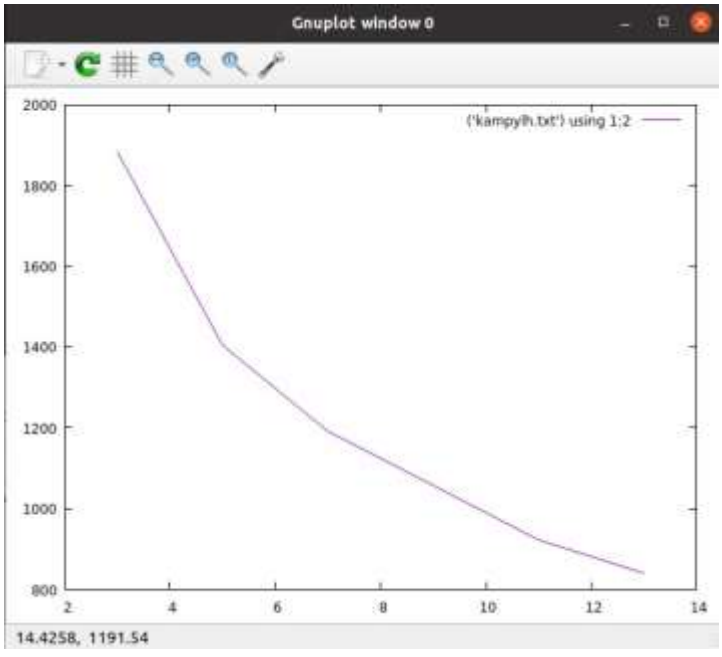
```
Centroid: (-0.7539337564715378,-0.6973043452921531)
Number of points at this cluster: 278
Cluster: 12
Centroid: (0.1679409232484289,0.004154861332422279)
Number of points at this cluster: 332
============================================================


============================================================
System finished clustering after with SSE 849.6654618563623
============================================================
Reading from file
Enter the number of clusters(bigger than 0) you want the program to apply to your data.
Enter '-1' in order to stop execution: -1
For 3 cluster -> SSE: 1861.866676910902
For 5 cluster -> SSE: 1394.8723564683319
For 7 cluster -> SSE: 1173.8886997779084
For 9 cluster -> SSE: 1000.7256047287398
For 11 cluster -> SSE: 921.2818330467643
For 13 cluster -> SSE: 849.6654618563623
PS D:\UNIVERSITY\SEMESTER_9\Ypologistikh_Nohmosynh\ask1>
```

Επίσης υπάρχουν ιδιαίτερες περιπτώσεις:

```
Reading from file
Enter the number of clusters(bigger than 0) you want the program to apply to your data.
Enter '-1' in order to stop execution: -1
PS D:\UNIVERSITY\SEMESTER_9\Ypologistikh_Nohmosynh\ask1> java Kmeans
Reading from file
Enter the number of clusters(bigger than 0) you want the program to apply to your data.
Enter '-1' in order to stop execution: 0
Enter the number of clusters(bigger than 0) you want the program to apply to your data.
Enter '-1' in order to stop execution: -1
PS D:\UNIVERSITY\SEMESTER_9\Ypologistikh_Nohmosynh\ask1>
```

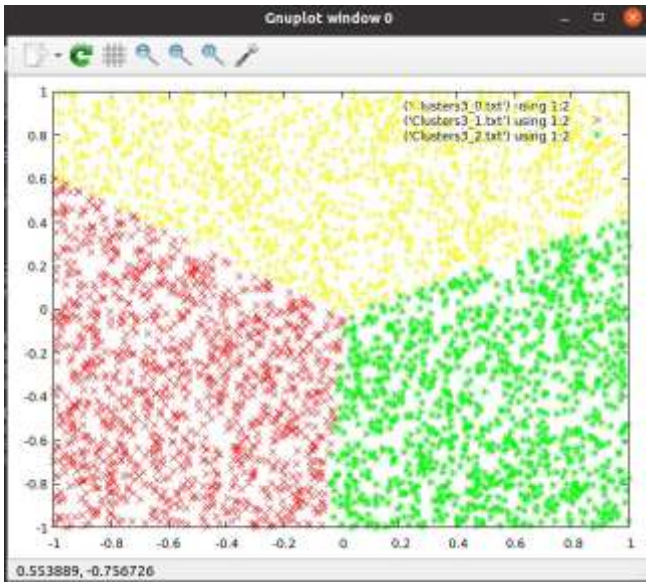
Η καμπύλη με το SSE και τα Μ:



Θα μπορούσαμε να επιλέξουμε και το 5 ως καλό αριθμό από clusters, αλλά το 9 παρουσιάζει επίσης μία μορφή γονάτου στη γραφική, έχει και χαμηλότερο SSE, οπότε το 9 είναι ο καλύτερος αριθμός για τη δημιουργία των clusters. Η καμπύλη παρουσιάζεται μέσω του gnuplot στα ubuntu με την εντολή: plot (‘kampylh.txt’) using 1:2 with line

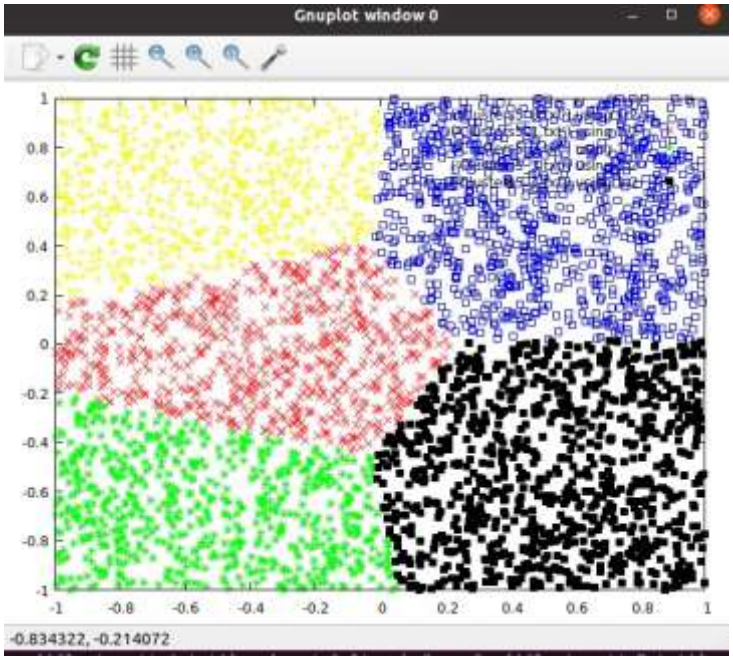
Τα αποτελέσματα του κάθε Cluster για κάθε αριθμό από Clusters:

plot('Clusters3_0.txt') using 1:2 lt rgb "yellow", ('Clusters3_1.txt') using 1:2 lt rgb "red", ('Clusters3_2.txt') using 1:2 lt rgb "green"



plot('Clusters5_0.txt') using 1:2 lt rgb "yellow", ('Clusters5_1.txt') using 1:2 lt rgb "red", ('Clusters5_2.txt') using 1:2 lt rgb "green", \

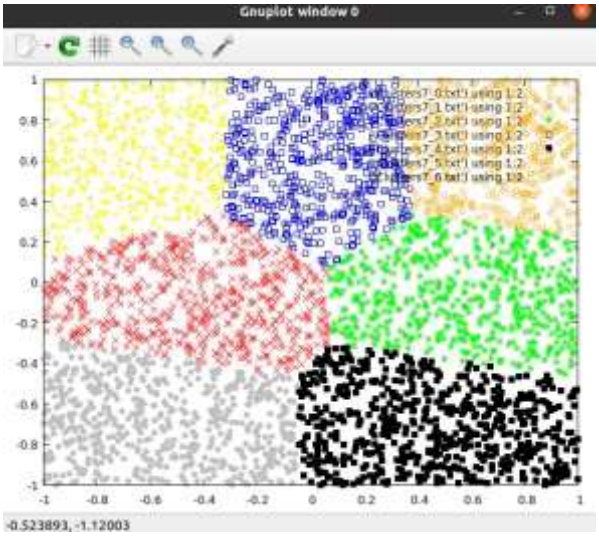
('Clusters5_3.txt') using 1:2 lt rgb "blue", ('Clusters5_4.txt') using 1:2 lt rgb "black"



plot('Clusters7_0.txt') using 1:2 lt rgb "yellow", ('Clusters7_1.txt') using 1:2 lt rgb "red", ('Clusters7_2.txt') using 1:2 lt rgb "green", \

('Clusters7_3.txt') using 1:2 lt rgb "blue", ('Clusters7_4.txt') using 1:2 lt rgb "black", ('Clusters7_5.txt') using 1:2 lt rgb "orange"\

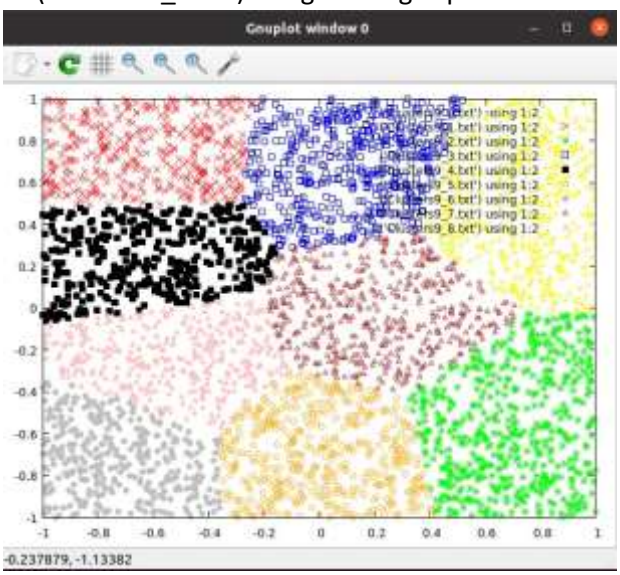
, ('Clusters7_6.txt') using 1:2 lt rgb "grey"



plot('Clusters9_0.txt') using 1:2 lt rgb "yellow", ('Clusters9_1.txt') using 1:2 lt rgb "red", ('Clusters9_2.txt') using 1:2 lt rgb "green", \

('Clusters9_3.txt') using 1:2 lt rgb "blue", ('Clusters9_4.txt') using 1:2 lt rgb "black", ('Clusters9_5.txt') using 1:2 lt rgb "orange"\

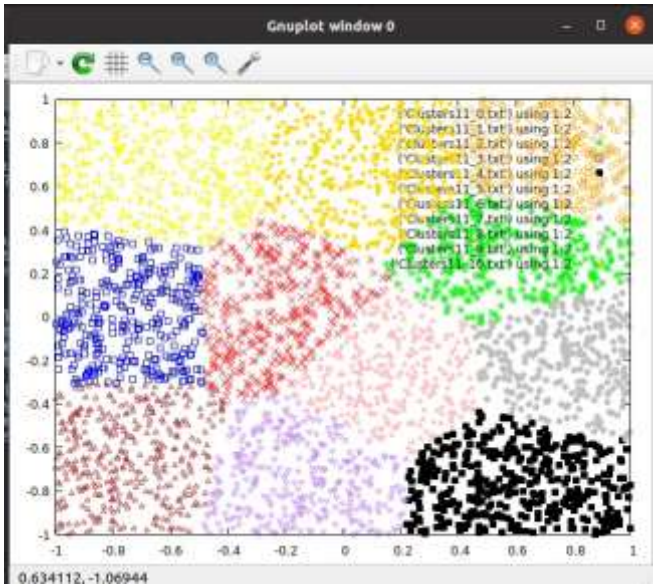
, ('Clusters9_6.txt') using 1:2 lt rgb "grey", ('Clusters9_7.txt') using 1:2 lt rgb "brown", ('Clusters9_8.txt') using 1:2 lt rgb "pink"



plot('Clusters11_0.txt') using 1:2 lt rgb "yellow", ('Clusters11_1.txt') using 1:2 lt rgb "red", ('Clusters11_2.txt') using 1:2 lt rgb "green", \('Clusters11_3.txt') using 1:2 lt rgb "blue", ('Clusters11_4.txt') using 1:2 lt rgb "black", ('Clusters11_5.txt') using 1:2 lt rgb "orange"\

, ('Clusters11_6.txt') using 1:2 lt rgb "grey", ('Clusters11_7.txt') using 1:2 lt rgb "brown", ('Clusters11_8.txt') using 1:2 lt rgb "pink"\

, ('Clusters11_9.txt') using 1:2 lt rgb "purple", ('Clusters11_10.txt') using 1:2 lt rgb "gold"



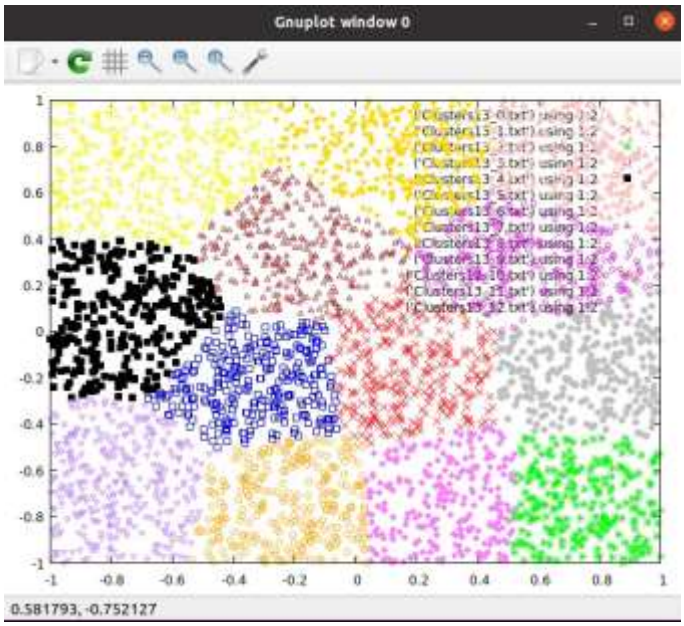
plot('Clusters13_0.txt') using 1:2 lt rgb "yellow", ('Clusters13_1.txt') using 1:2 lt rgb "red", ('Clusters13_2.txt') using 1:2 lt rgb "green", \

('Clusters13_3.txt') using 1:2 lt rgb "blue", ('Clusters13_4.txt') using 1:2 lt rgb "black", ('Clusters13_5.txt') using 1:2 lt rgb "orange"\

, ('Clusters13_6.txt') using 1:2 lt rgb "grey", ('Clusters13_7.txt') using 1:2 lt rgb "brown", ('Clusters13_8.txt') using 1:2 lt rgb "pink"\

, ('Clusters13_9.txt') using 1:2 lt rgb "purple", ('Clusters13_10.txt') using 1:2 lt rgb "gold", ('Clusters13_11.txt') using 1:2 lt rgb "magenta"\

, ('Clusters13_12.txt') using 1:2 lt rgb "orchid"



Εντολή μεταγλώττισης: javac ./Kmeans.java

Εντολή εκτέλεσης: java Kmeans