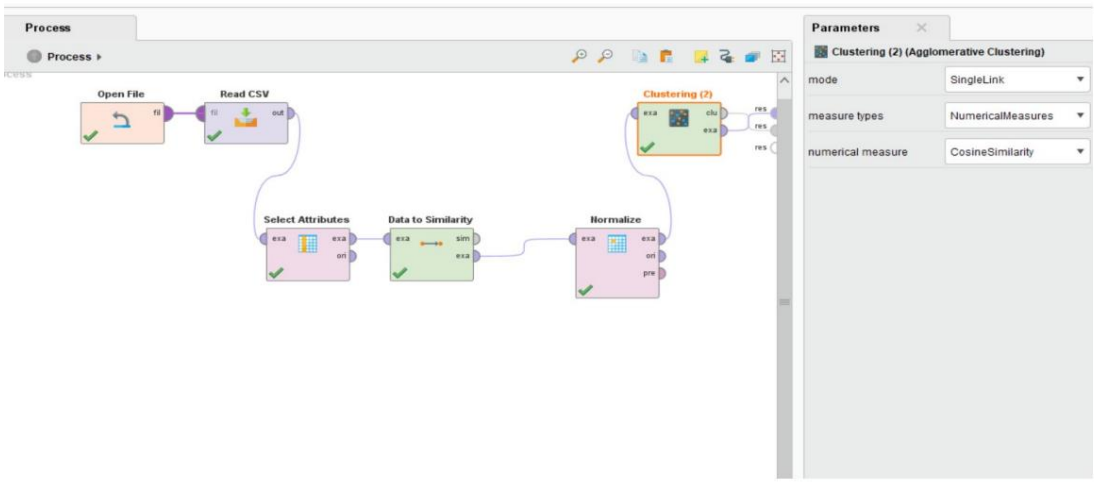
UNIVERSITY OF WESTERN ATTICA
SCHOOL OF TECHNOLOGICAL
APPLICATIONS DEPARTMENT OF INFORMATION
AND COMPUTER ENGINEERING

VASILIS CHRISTODOULOU
AM:161028

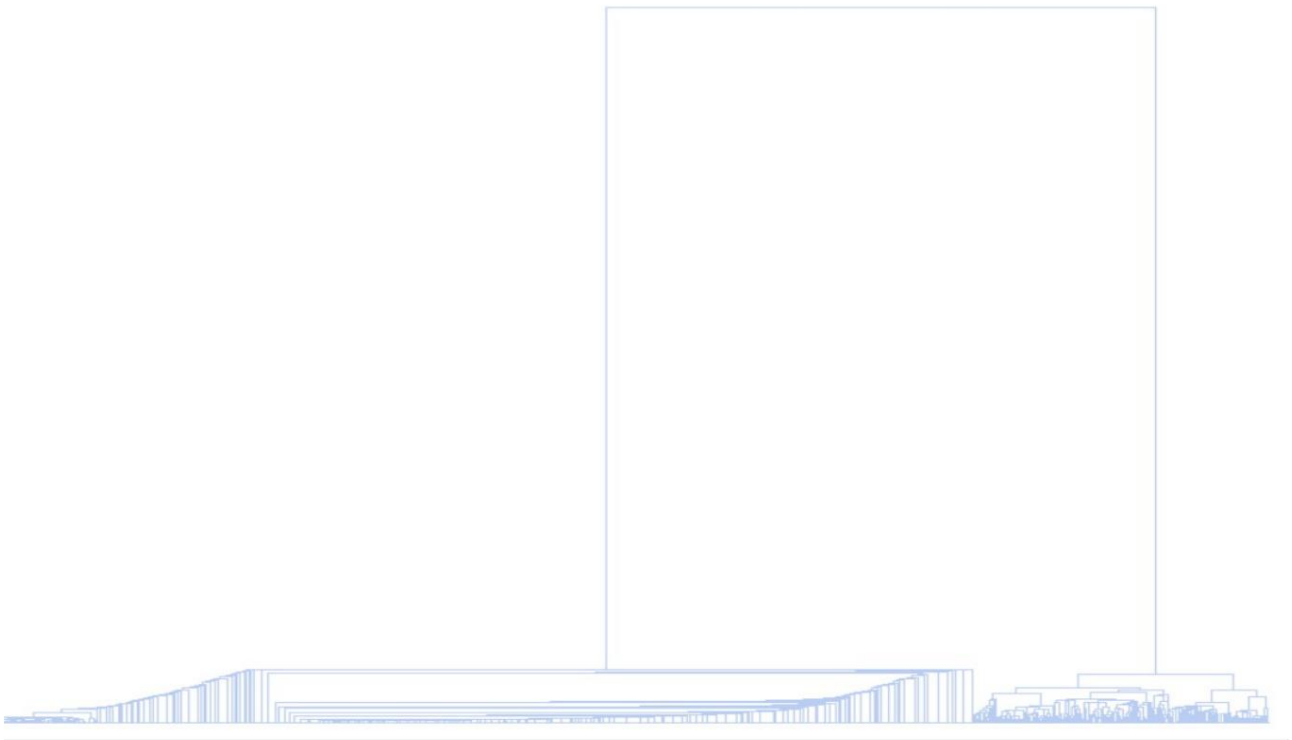
DATA MINING TASK 1
DATA COLLECTION

PART 1

1.1



Dendrogram:



1.2

Dendrogram:

1.3

Dendrogram:

z-score:

1.4

Based on what results from the above questions
1.1, 1.2 and 1.3 we see through the dendrograms
that were produced that in the case of 1.2 with the Cosine
similarity index and also with the method of the average
bond in the second and third columns of the table
enron100 we have better results in the
dendrogram. In particular, much more information is
captured and the final image is sharper for study and
analysis. Unlike the other two methods of simple linkage and
Cosine similarity index
and Euclidean distance where the results are not easy to
read. Also a lot of great information is captured in the z-
score transformation.

Below are images of the operators used with their parameters

PART 2 2.1

Scatters:

Numerical statistics

2.2
Manhattan scatter

Manhattan Numerical statistics

Cosine scatter

Cosine Numerical statistics

2.3

Observing the above results we come to the final conclusion that with all three similarity measures used Euclidean distance, Manhattan and Cosine the results are identical and no obvious differences can be distinguished as we use the same

partitional k-means clustering method.

2.4

Scatters:

Numerical statistics

2.5
Scatters

2.6

Scatter

Numerical statistics

2.7

By performing the above steps 2.4, 2.5 and 2.6 we arrive to the conclusion that in the case where
all the characteristics of the xV table were used the result in the graph is more complete and captures more information than the other two. While in the other two cases the

graphs are sparser and do not have a high density
of elements as in 2.6, because only two features are used
at a time. Regarding the measure we notice that the
similarity measure
Cosine is the one that performs better as seen in the
questions above giving more detailed graphs as opposed
to Euclidean distance measures
and Manhattan.

Below are images of the operators used with
their parameters

PART 3 3.1

1st chart

2nd chart

3.2

In the two clustering questions 3.1 and 3.2 we notice in the scatterplots that with the two methods DBSCAN and k-means quite similar graphs are produced with the Euclidean distance, but we notice that in the second graph there is a visual difference as it seems to display more information than in the first making it more complete.

3.3

3.4

3.5

In the previous steps 3.3 and 3.4, DBSCAN clustering methods were used in the iris data and in 3.4, the data was normalized with the z score method. Thus, observing the results, the opinion emerges that in both methods we have a great similarity
in the end and it is not easy to see some differences

either in the scatter plot or the cluster plot.

3.6

3.7

I chose for parameter values 0.9 and 40 respectively for ÿ and MinPts

3.8

In the above two examples 3.6 and 3.7 we see that in the scatterplots there are no big differences to the point where we could say they are the same. This is because there was no big change in the parameter values (from 0.5 to 0.9 and from 50 to 40). In another case with larger parameters we would observe a more noticeable difference in the two graphs

Below are images of the operators used with their parameters