Assignment 2 – Rajesh Reddy Vanga

Problem description:

1.Find 100 CNN news articles online (try to find them in different categories, e.g., sports and finance). You need to find the new article by yourself. Pls ignore picture or other non-text data in the new article.

The 100 CNN articles can be found in the documents.zip file. Articles are collected from 5 different categories, Entertainment , Health, Politics, Sports, Technology.

The entire index of the articles can be found in the file CNN Articles.xlsx along with the category that a document belongs to and also the URL links and headings of the documents.

2.Convert them to data matrix (each row is an article and each column is a unique term).

R Programming with the help of tm package is used to generate the document term matrix. It can be achieved from the following script:

Set Working Directory to the folder that has all the documents:

setwd("D:\\KDD\\Rajesh\_Assign2\\Totalset")

cname <- file.path("D:\\KDD\\Rajesh\_Assign2", "Totalset")

dir(cname);

Invoke the tm library:

library(tm)

Create a corpus of the docs:

docs <- Corpus(DirSource(cname))

Remove punctuations:

docs <- tm\_map(docs, removePunctuation)

Remove Numbers:

docs <- tm\_map(docs, removeNumbers) ;

Convert to lower case:

docs <- tm\_map(docs, content\_transformer(tolower)) ;

Remove English Stopwords:

docs <- tm\_map(docs, removeWords, stopwords("english")) ;

Strip whitecases:

docs <- tm\_map(docs, stripWhitespace);

Generate the Document Term matrix:

dtm <- DocumentTermMatrix(docs)

Convert it as a matrix:

m <- as.matrix(dtm)

Write the DTM to a csv file:

write.csv(m, file="DocumentTermMatrix.csv")

The DocumentTermMatrix.csv file can be found in the root assignment folder

3. Run K-means clustering with Euclidean, Cosine and Jaccard similarity. (Specify K as the number of categories of your 100 CNN news articles)

#install and load libraries

install.packages("tm")

install.packages("fpc")

install.packages("proxy")

library(proxy)

library(tm)

library(fpc)

# input the document term matrix as input

dtmfile <-read.csv("D:\\KDD\\Rajesh\_Assign2\\TotalSet\\dtm.csv");

# parse the input as a matrix

matrix<-data.matrix(dtmfile)

#Calculate Euclidean

ekcinput <- proxy::dist(matrix, method = "euclidean")

#Calculate Cosine distance

ckcinput <- proxy::dist(matrix, method = "cosine")

#Calculate jaccard distance

jkcinput <- proxy::dist(matrix, method = "jaccard")

#Calculate kmeans on Euclidean

ekcoutput <-kmeans(ekcinput,centers=5)

#Calculate kmeans on cosine

ckcoutput <-kmeans(ckcinput,centers=5)

#Calculate kmeans on Jaccard

jkcoutput <-kmeans(jkcinput,centers=5)

# Print the outputs and evaluate the results

print(ekcoutput)

print(ckcoutput)

print(jkcoutput)

The EuclideanMatrix.csv, JaccardMatrix.csv and CosineMatrix.csv files can be found in the root assignment folder.

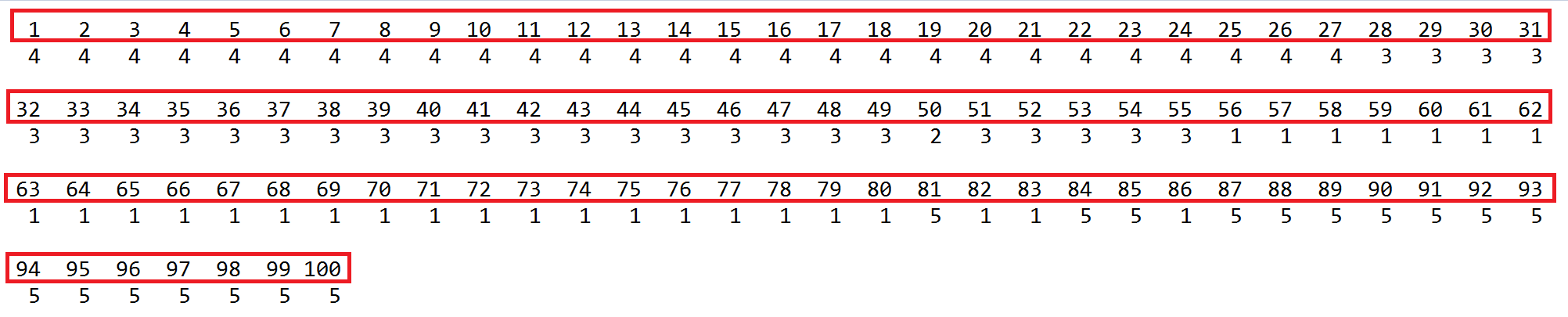
The K-means clustering outputs with Euclidean, Cosine and Jaccard measures (Euclidean\_Kmeans.txt, Cosine\_Kmeans.txt and Jaccard\_kmeans.txt) can be found in the root assignment folder along with the cluster graph plots (Euclideanplot.jpeg, Cosineplot.jpeg and Jaccardplot.jpeg)

4. Evaluate K-means clustering results with SSE

* K-means Clustering with Euclidean measures:

The Cluster means and document vectors can be found in the file Euclidean\_Kmeans.txt file.

Clustering vector:



Intra cluster sum of squares by cluster:

Cluster 1 - 143071.1

Cluster 2 - 0.0

Cluster 3 - 222072.0

Cluster 4 - 227300.7

Cluster 5 - 144162.2

Since documents are chosen from a batch of 20 documents from each category, the first 20 belong to the same category, the next 20 belong to another and so on..

1- 20 : Politics

21-40 : Sports

41-60 : Health

61-80 : Science & Technology

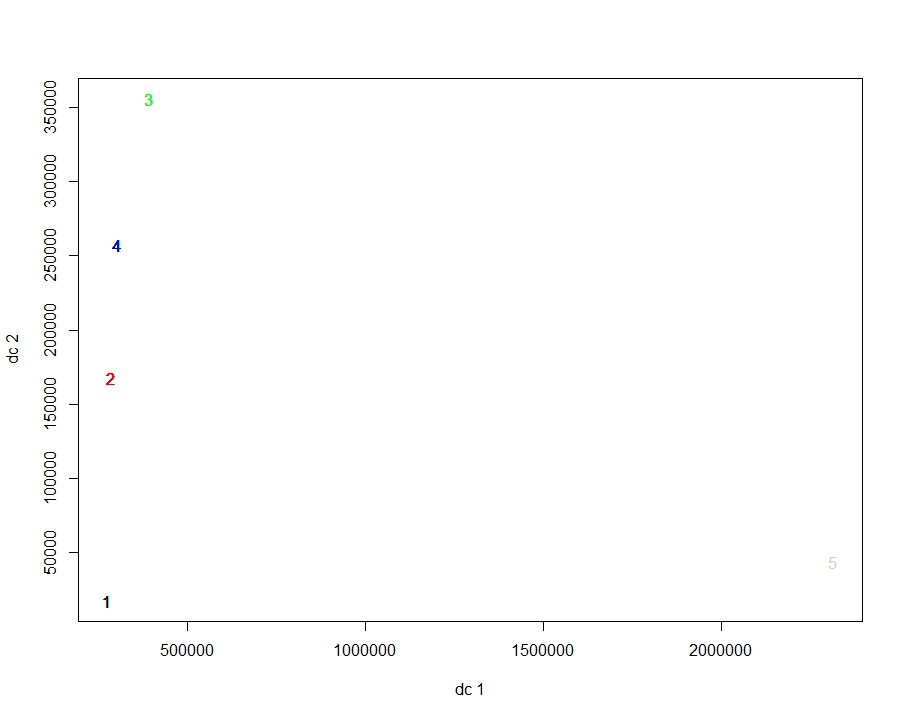
81-100 : Entertainment

From the Clusters formed it can be interpreted that 1 to 20 have been clustered into a single cluster, however, even 21-27 have also been included in this cluster which belong to a different class. 28-40 have been grouped under cluster 3.

Since only the document 50 is added under cluster 2, It becomes the only element in that cluster. It effectively becomes the centroid of the cluster. Hence, the SE is 0 and the SSE of cluster 2 can be seen as 0.

Barring Cluster 2, the SSE of cluster 1 is the least. This explains that the data points in the cluster are the nearest to the centroid of the cluster and hence are more similar, it can be seen from the results as well and most of the cluster points belong to category 1.

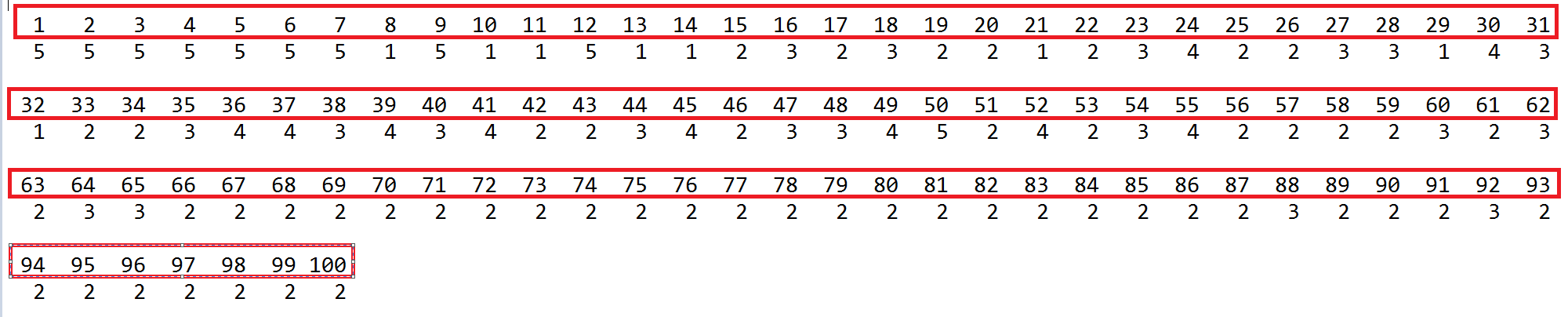
The Cluster Plots can be observed from the following image:



* K-means Clustering with Cosine measures:

The Cluster means and document vectors can be found in the file Cosine\_Kmeans.txt file.

Clustering vector:



Intra cluster sum of squares by cluster:

Cluster 1 - 5.849463

Cluster 2 - 2.936238

Cluster 3 - 2.776550

Cluster 4 - 2.909846

Cluster 5 - 13.227957

Since documents are chosen from a batch of 20 documents from each category, the first 20 belong to the same category, the next 20 belong to another and so on..

1- 20 : Politics

21-40 : Sports

41-60 : Health

61-80 : Science & Technology

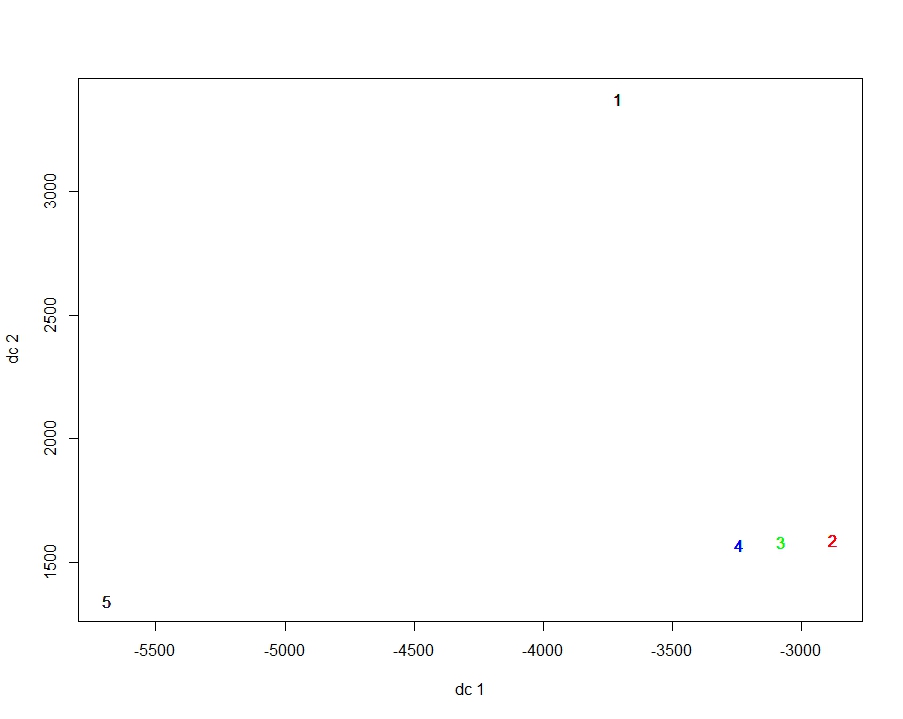
81-100 : Entertainment

From the Clusters formed it can be interpreted that the clusters have documents from different clusters under them.

The documents from category politics have been added to the clusters 5, 1, 2 and 3 with the majority being added to cluster 5.

Cluster 2 is more pure with most of the documents are from the category Science & Technology. The Same can be said about the category Entertainment where most of the documents have been added to cluster 2. But the documents from both these categories have been mostly added to a single cluster, in this case cluster 2.

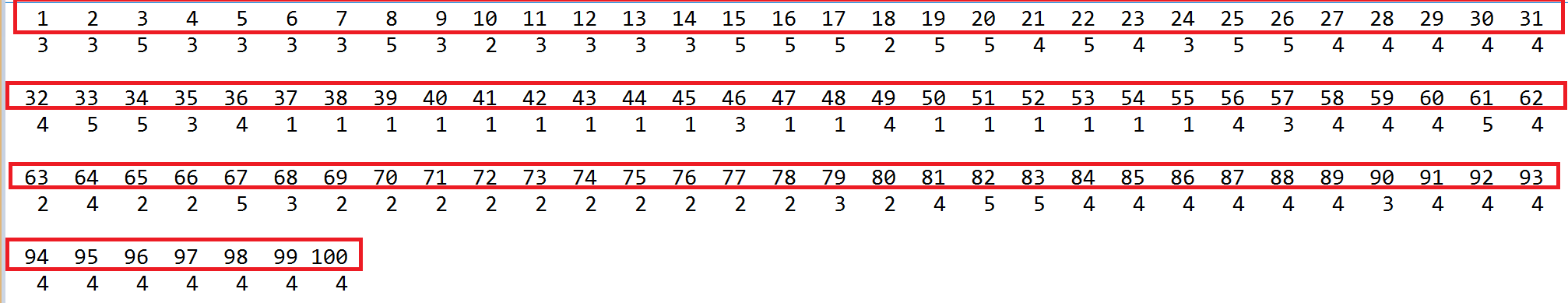
The Cluster Plots can be observed from the following image:



* K-means Clustering with Jaccard measures:

The Cluster means and document vectors can be found in the file Jaccard\_Kmeans.txt file.

Clustering vector:



Intra cluster sum of squares by cluster:

Cluster 1 - 13.88746

Cluster 2 - 13.49479

Cluster 3 - 15.11011

Cluster 4 - 29.35407

Cluster 5 - 14.48461

Since documents are chosen from a batch of 20 documents from each category, the first 20 belong to the same category, the next 20 belong to another and so on..

1- 20 : Politics

21-40 : Sports

41-60 : Health

61-80 : Science & Technology

81-100 : Entertainment

From the Clusters formed we can analyze that most of the documents under category Health have been clustered under the Cluster 1.

Most of the documents under category Entertainment have been clustered under the Cluster 4. The same can be said about the category Science & Technology which are clustered into the cluster 3.

The Cluster Plots can be observed from the following image:

