Assignment 1 – 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.

1. 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("C:\\Users\\rajes\\Desktop\\Rajesh\\Total set")

cname <- file.path("C:\\Users\\rajes\\Desktop\\Rajesh", "Total set")

dir(cname);

Invoke the tm library:

library(tm)

Create a corpus of the docs:

docs <- Corpus(DirSource(cname))

docs <- Corpus(VectorSource(docs))

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="dtm.csv")

1. Compute the similarity between each pair of articles with Euclidean distance (you need to convert the Euclidean distance to similarity), Cosine and Jaccard.

Euclidean Distance:

dist(dtm, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)

cosine\_dtm <- as.matrix(dist(t(dtm), method = "cosine"))

install.packages("proxy")

library("proxy")

Cosine Dissimilarity Function:

proxy::dist(m, method = "cosine")

Jaccard Dissimilarity Function:

proxy::dist(m, method = "jaccard")

The Document Term Matrix, Cosine dissimilarity matrix, Jaccard dissimilarity matrix and Euclidean Distance matrix can be found in the following files:

dtm.csv

cosine.csv

jaccard.csv

euclidean.csv

1. Sort all pairs from most similar to least similar based on each of the three types of similarity measurements

The sorted pairs of the respective similarity measurements can be found in the 2nd sheet in the following csv files

cosine.csv

jaccard.csv

euclidean.csv

1. Compare the sorted pairs and discuss which one is more accurate (i.e., close to your own judgment)

The collaboration of the Similarity measurements can be found in the excel sheet Collaborated Measurements.xls file. As it can be inferred from the file/below image, The results obtained from the Jaccard and Cosine measurements are identical and that of the Euclidean don't quite match that of the other two.

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 an other and so on..

1- 20 : Politics

21-40 : Sports

41-60 : Health

61-80 : Science & Technology

81-100 : Entertainment

From the above pairs and similarity measurements, it can be seen that jaccard and cosine are more accurate since the similar pairs (D7,D6) ; (D99,D87) ; (D87,100) ; (D87,D88) all belong to the same categories majorly. The same cannot be said about that of the pairs obtained by euclidean distances since more than a handful of pairs have been wrongly paired like (D19,D22) , (D25,D19) , (D83,D19) (D26,D19), (D67, D19).

From the data, It can be inferred that both Jaccard and Cosine similarity measurements are more accurate for the set of documents than that of euclidean distance.

