Arun Sangar

Homework 4

1)

a. Term-Document Count Matrix

	Document 1	Document 2	Document 3
see	1	1	0
spot	1	1	1
run	0	1	2

b. Term-Document Matrix weighted by tf

	Document 1	Document 2	Document 3
see	1	1	0
spot	1	1	1
run	0	1	1.3

c. Term IDF Values

	idf
see	.1761
spot	0
run	.1761

d. Term-Document Matrix weighted by tf-idf

	Document 1	Document 2	Document 3
see	.1761	.1761	0
spot	0	0	0
run	0	.1761	.2289

e. “run” has the most information according to the tf-idf matrix.
“spot” has the least information according to the tf-idf matrix.

2)

a. inspect(mycorpus2[[1]]) (removed html tags)

   One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this is exactly what happened with me. The first thing that struck me about Oz was its brutality and unflinching scenes of violence, which set in right from the word GO. Trust me, this is not a show for the faint hearted or timid. This show pulls no punches with regards to drugs, sex or violence. Its is hardcore, in the classic use of the word. It is called OZ as that is the nickname given to the Oswald Maximum Security State Penitentary. It focuses mainly on Emerald City, an experimental section of the prison where all the cells have glass fronts and face inwards, so privacy is not high on the agenda. Em City is home to many..Aryans, Muslims, gangstas, Latinos, Christians, Italians, Irish and more....so scuffles, death stares, dodgy dealings and shady agreements are never far away. I would say the main appeal of the show is due to the fact that it goes where other shows wouldn't dare. Forget pretty pictures painted for mainstream audiences, forget charm, forget romance...OZ doesn't mess around. The first episode I ever saw struck me as so nasty it was surreal, I couldn't say I was ready for it, but as I watched more, I developed a taste for Oz, and got accustomed to the high levels of graphic violence. Not just violence, but injustice (crooked guards who'll be sold out for a nickel, inmates who'll kill on order and get away with it, well mannered, middle class inmates being turned into prison bitches due to their lack of street skills or prison experience) Watching Oz, you may become comfortable with what is uncomfortable viewing....thats if you can get in touch with your darker side.

b. reviewdata <- read.csv("C:\\Users\\aruns\\Desktop\\Fall 2018\\CPSC 375\\Homework 4\\movie_reviews.csv")

```
review_corpus <- Corpus(VectorSource(reviewdata[,1]))
review_corpus2 <- Corpus(VectorSource(reviewdata[,1]))
for(i in seq(review_corpus))
   review_corpus2[[i]] <- gsub("[<][b][r][ ][/][>][<][b][r][ ][/][>]", " ", review_corpus[[i]])
review_corpus3 <- tm_map(review_corpus2,tolower)
review_corpus4 <- tm_map(review_corpus3,removeNumbers)
review_corpus5 <- tm_map(review_corpus4,removePunctuation)
review_corpus6 <- tm_map(review_corpus5,removeWords, stopwords("english"))
review_corpus7 <- tm_map(review_corpus6,stripWhitespace)
```

c. 162752 terms

```
dtm_tfidf <- DocumentTermMatrix(review_corpus7, control = list(weighting =
weightTfIdf))
```

d. 1596 terms

```
dtm_tfidf2 <- removeSparseTerms(dtm_tfidf,.99)
```

e.

i.
```
cosinesim <- function(doc1,doc2)
{
        sumab <- 0
        sumasq <- 0
        sumbsq <- 0
        retcosim <- 0

        for(i in 1:length(doc1))
        {
                sumab = sumab + (doc1[i] * doc2[i])
                sumasq = sumasq + (doc1[i]^2)
                sumbsq = sumbsq + (doc2[i]^2)
        }

        if(sumbsq > 0) retcosim = sumab / sqrt(sumasq * sumbsq)
        return (head(unname(retcosim)))
}
```

ii.
```
mymatrix <- as.matrix(dtm_tfidf2)
```

iii.
```
cosim <- 0
maxcosim <- 0
maxindex <- 2

for(i in 2:nrow(mymatrix))
{
        cosim = cosinesim(mymatrix[1,],mymatrix[i,])
        if(cosim > maxcosim)
        {
                maxcosim = cosim
                maxindex = i
        }
}
```

Index of review with maximum cosine similarity: 42868

f. Inspect(mycorpus2[[42868]])

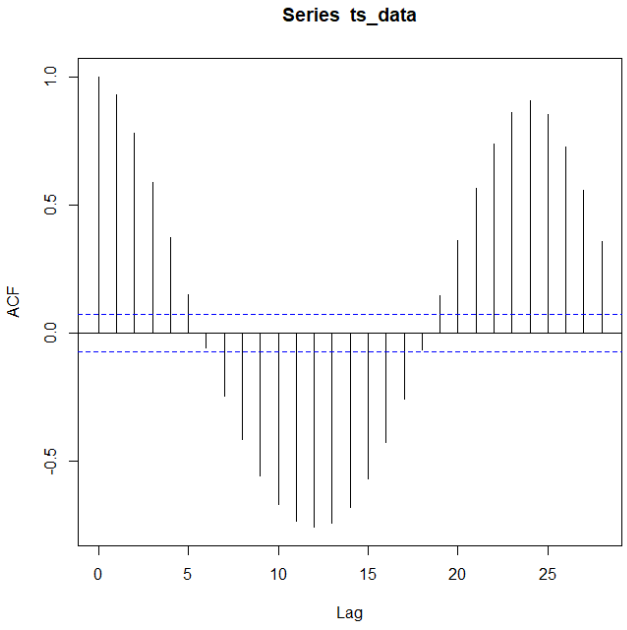
I admit that I almost gave up on watching TV shows. Why? Because most of them are about doctors, forensics or some girl/boy, who can predict a murder by sitting on the toilet seat. But I must say I was wrong when I watched the first episode of Oz, one of the most brilliant shows on TV until today. Oz is a show about maximum secure prison and shows episodes in life of every cell - mate inside the prison and their frustrations. What I like about this show is the fact it has the courage to show the real life in prison without any annoying characters, stereotypes and unrealistic dialogs. Even thought the characters in the show are not supposed to be heroes, you became quickly attached to them and their life. Sure, their is a lot of violence in this show and some disturbing images of rape and taking drugs, but that's the point of this show; to present how brutal and frustrating can be life in prison... heck, you can't show life in prison like a "clean environment" with "likeable characters", right? Oz is perfect in every way, the actors are doing an excellent job and the whole hour of this show is going threw really really fast. The only thing that I didn't like in this show was the last season of the show, I think it became too weird and brutal and the finale was not satisfying. But overall Oz is still my favorite show on TV so far, it surpasses Sopranos and million CSI and Dr. House crap with just a main title. Brilliant show which has the balls to show what other shows are taking away from us; realism.

g. Yes!

3)

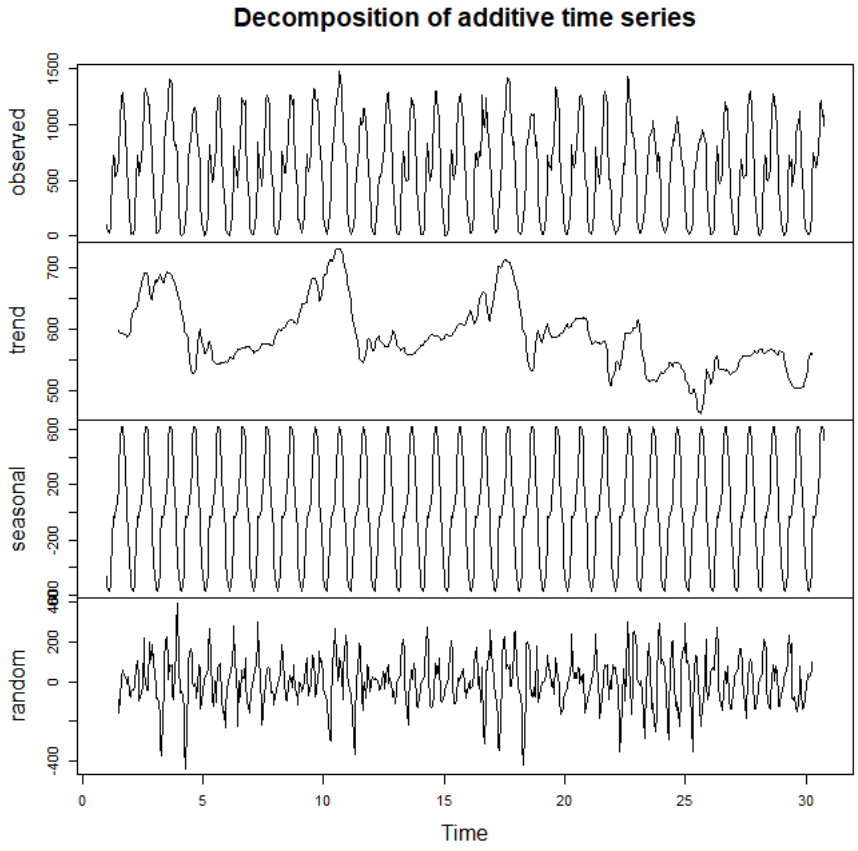
a. raw_data <- scan('C:\\Users\\aruns\\Desktop\\Fall 2018\\CPSC 375\\Homework 4\\pems_output.txt')

b. acf(raw_data)



c. ts_data <- ts(data=raw_data,frequency=24)

d. plot(decompose(ts_data))



4)

a. Cost matrix

(1-1)^2 = 0	(1-2)^2 = 1	(1-5)^2 = 16
(2-1)^2 = 1	(2-2)^2 = 0	(2-5)^2 = 9
(4-1)^2 = 9	(4-2)^2 = 4	(4-5)^2 = 1
(5-1)^2 = 16	(5-2)^2 = 9	(5-5)&2 = 0

b. C matrix

0	0 + 1 = 1	1 + 16 = 17
0 + 1 = 1	Min(1,1,0) + 0 = 0	Min(17,0,1) + 9 = 9
1 + 9 = 10	Min(0,10,1) + 4 = 4	Min(9,4,0) + 1 = 1
10 + 16 = 26	Min(4,26,10) + 9 = 13	Min(1,13,4) + 0 = 1

Dynamic Time Warping Distance: 1

5)

a. Sliding Window
b. Standing Query
c. Fixed Size Sampling
d. Proportional Sampling