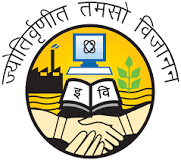
MINOR PROJECT

ON

User Future Request Prediction Using FCM in Web Usage Mining

Submitted in partial fulfilment of requirement of B.Tech CSE (Session 2014-18)



Under the guidance of: Submitted by:

Dr. ANJU SAHA SUJIT KUMAR

En. 05216403214

B.Tech CSE 7th sem

**University School of Information and Communication Technology**

**GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY**

**DWARKA SEC-16C, DELHI 110078**

**Content**

1. Introduction
2. Strategy
3. Cleaning Process
4. Reference

Introduction

Web mining is an application of data mining that uses data mining techniques to extract useful information from web documents. Web mining is further divided into three types Web Usage Mining, Web Content Mining and Web Structure Mining. Web usage mining is a process of mining useful information from server logs. When user use the internet and open different websites then browsing behaviour of the user automatically save into log file on server sites. Web usage mining analyses these log files and extracts information about user browsing behaviour on internet. This information is used in Personalization, Improving the website design, Business intelligence and predicting the user future requests.

Strategy

Web Log Mining is subject of studying Data Mining. Mining is very tedious job. Mining means extracting information from raw data in which there is some hidden knowledge. Raw data can be structured and unstructured. Structured data are organised in table form or tree hierarchical form like table in Database. Unstructured data are unorganised data like text files, images with information shown in it, or images of location. Here we have Log File in common format of apache server. Firstly, we need to know what are information can be obtained from web server log file and if it is sufficient to obtain information for determining for useful knowledge. After analysing and considering answers of questions, we need to know how to organise data. Web usage mining have broadly three steps for discovering knowledge, 1.) cleaning 2.) processing and 3.) creating model on data from which we can get information which is our objective. Cleaning and organising process is very huge task which is around 80% of whole task. In this process removing unnecessary data. Organising data into well-formed. Now organised data divides into sample test and training data. So that test data can be used to test the model. To Make model need to create modules. Each Module have its own job for discovering the knowledge. When all modules get together to make analysis on well-formed data and then discovers knowledge. Model takes input from test-data-set so that it could predict and recommend reference for future request pages so that we can browse on it. It helps to improve our search requests.

Cleaning Process

It is first phase of web log mining. We take data log file in common format of apache. Import log file named as logData into R studio where we implementing the model. Removing unnecessary rows which have invalid ip address, request that have image media link, create data frame on status of successful request and converting the date time into standard format. Adding column of domain name from request field.

#To import log file

library(readxl)

access\_log2 <- read\_excel("E:/sujeet sharma/R code/access\_log/access\_log2.xlsx")

View(access\_log2)

#Extracting subset of log file which have successfull request

logData<-subset(access\_log2, access\_log2$Status==200)

#Extracting subset of log file which have Valid IP address

logData<-logData[grep("(([0-9]|[1-9][0-9]|1[0-9]{2}|2[0-4][0-9]|25[0-5]).){3}([0-9]|[1-9][0-9]|1[0-9]{2}|2[0-4][0-9]|25[0-5])",logData$`IP address`),]

head(logData)

#Creating data frame in which Request field have Image Media Link

logImg<-logData[grep(c(".png|.gif|.jpeg|.jpg|.ai|.bmp|.ico|.ps|.psd|.svg|.tif|.tiff"),logData$Request),]

#Removing rows in which Request field have image media link

logData1<-logData[-grep(c(".png|.gif|.jpeg|.jpg|.ai|.bmp|.ico|.ps|.psd|.svg|.tif|.tiff"),logData$Request),]

#Converting date time field in standard format

logData1$Date<-strptime(logData1$Date,"[%d/%B/%Y:%H:%M:%S")

#Removing rows in which Request made without domain name

logData1<-logData1[-grep("GET / HTTP/1.1",logData1$Request),]

#Removing Status and Size Comlumn

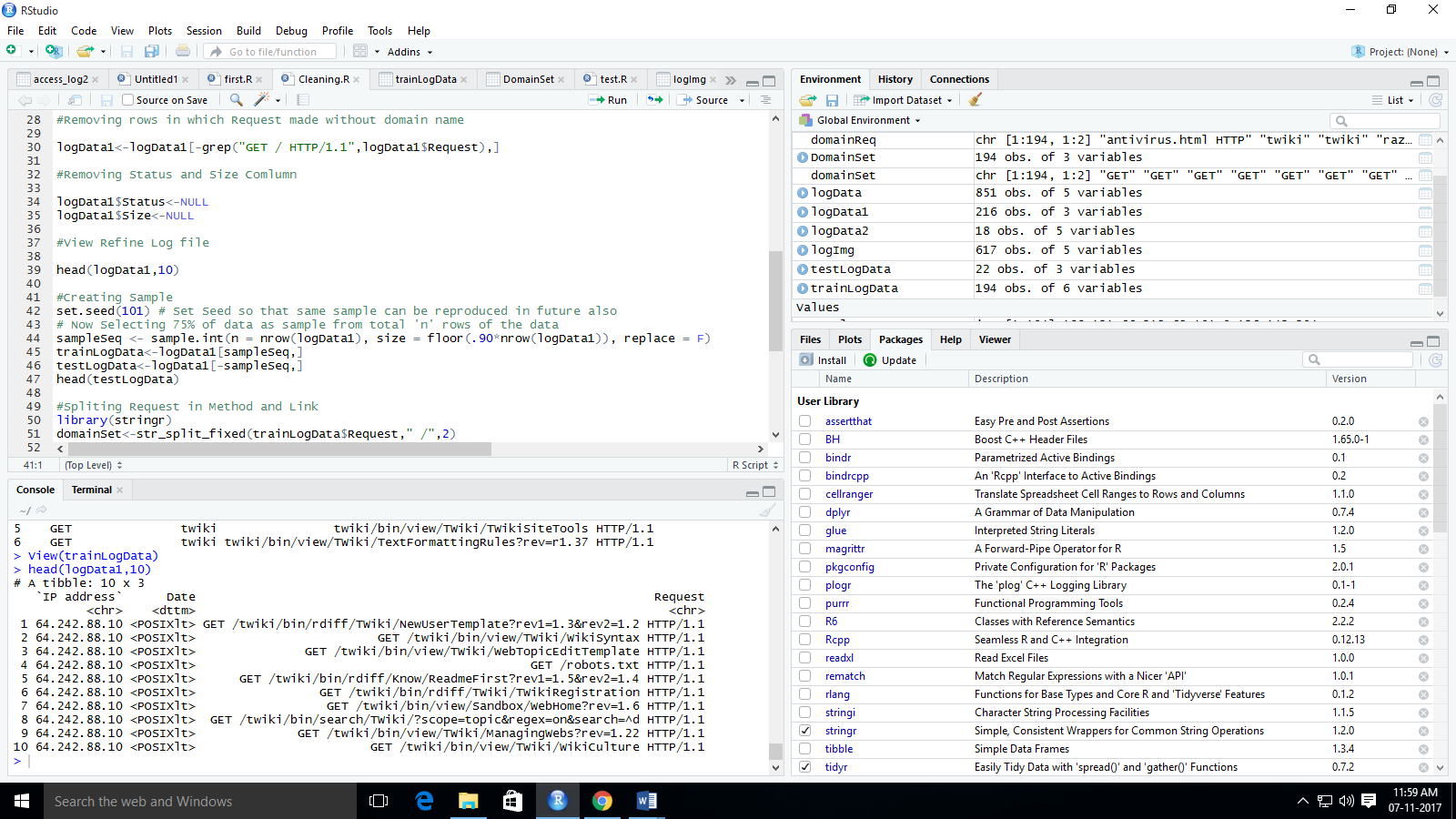
logData1$Status<-NULL

logData1$Size<-NULL

#View Refine Log file

head(logData1,10)

#output



#Creating Sample

set.seed(101) # Set Seed so that same sample can be reproduced in future also

# Now Selecting 75% of data as sample from total 'n' rows of the data

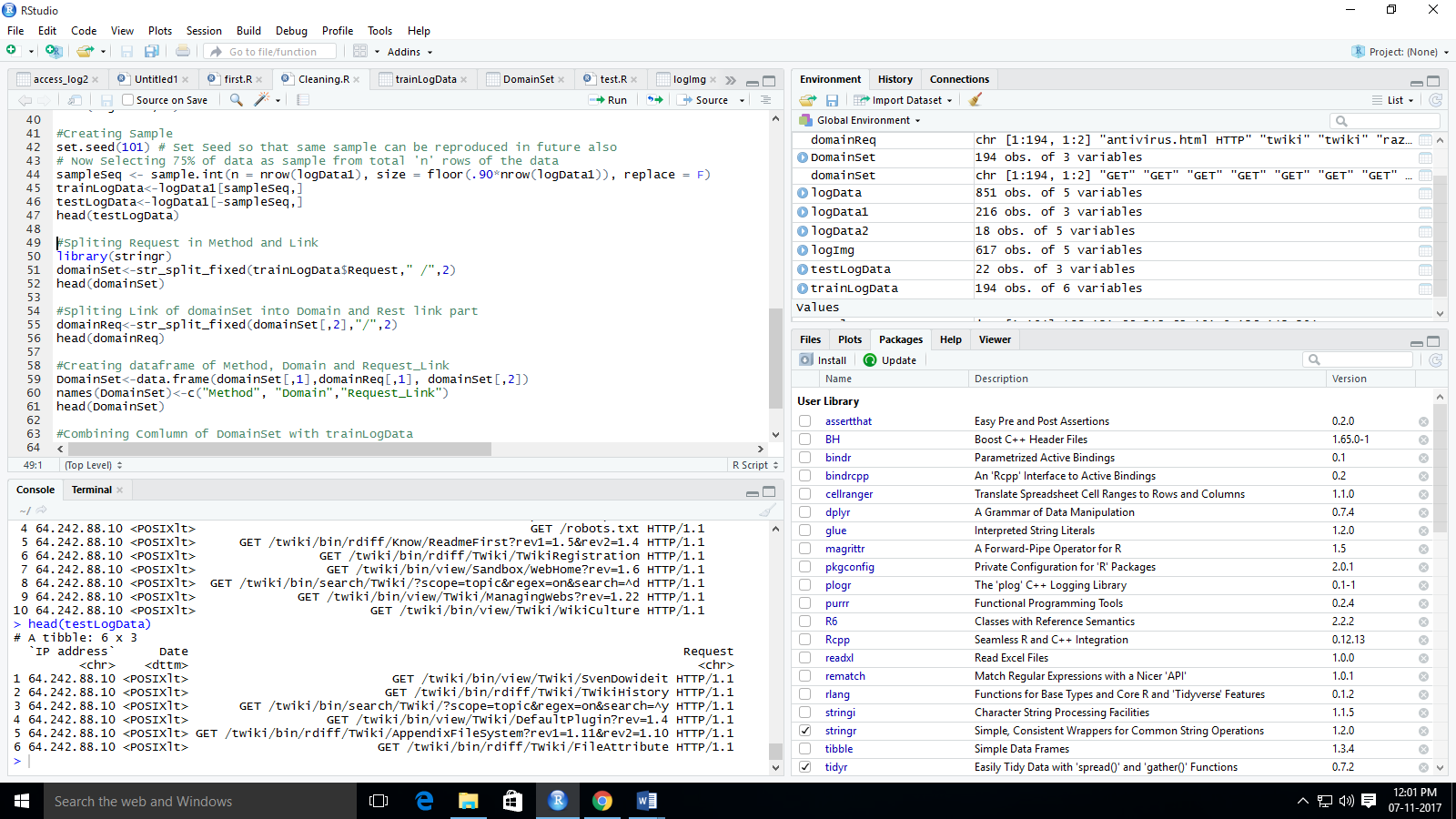
sampleSeq <- sample.int(n = nrow(logData1), size = floor(.90\*nrow(logData1)), replace = F)

trainLogData<-logData1[sampleSeq,]

testLogData<-logData1[-sampleSeq,]

head(testLogData)

#output



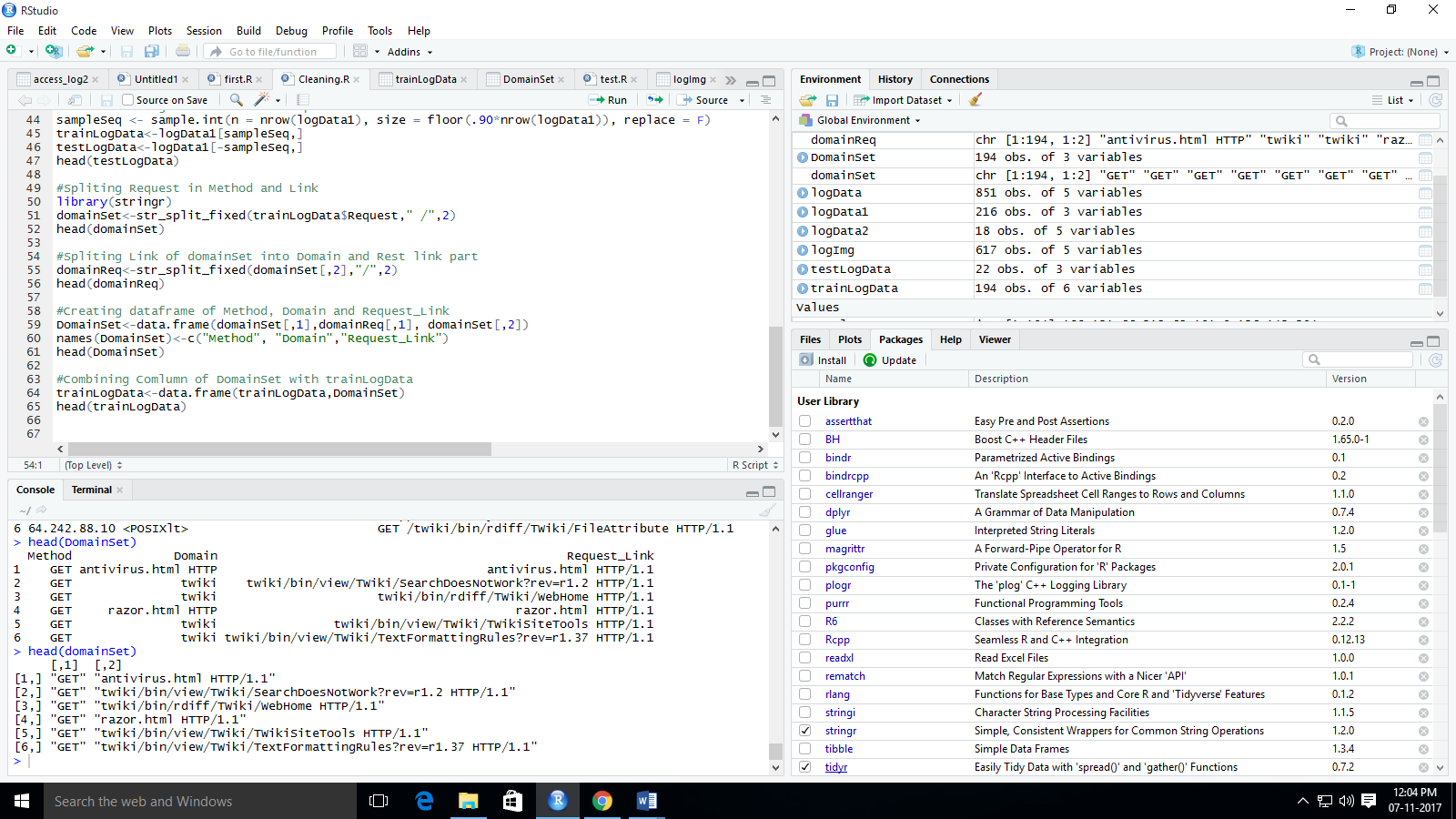
#Spliting Request in Method and Link

library(stringr)

domainSet<-str\_split\_fixed(trainLogData$Request," /",2)

head(domainSet)

#output

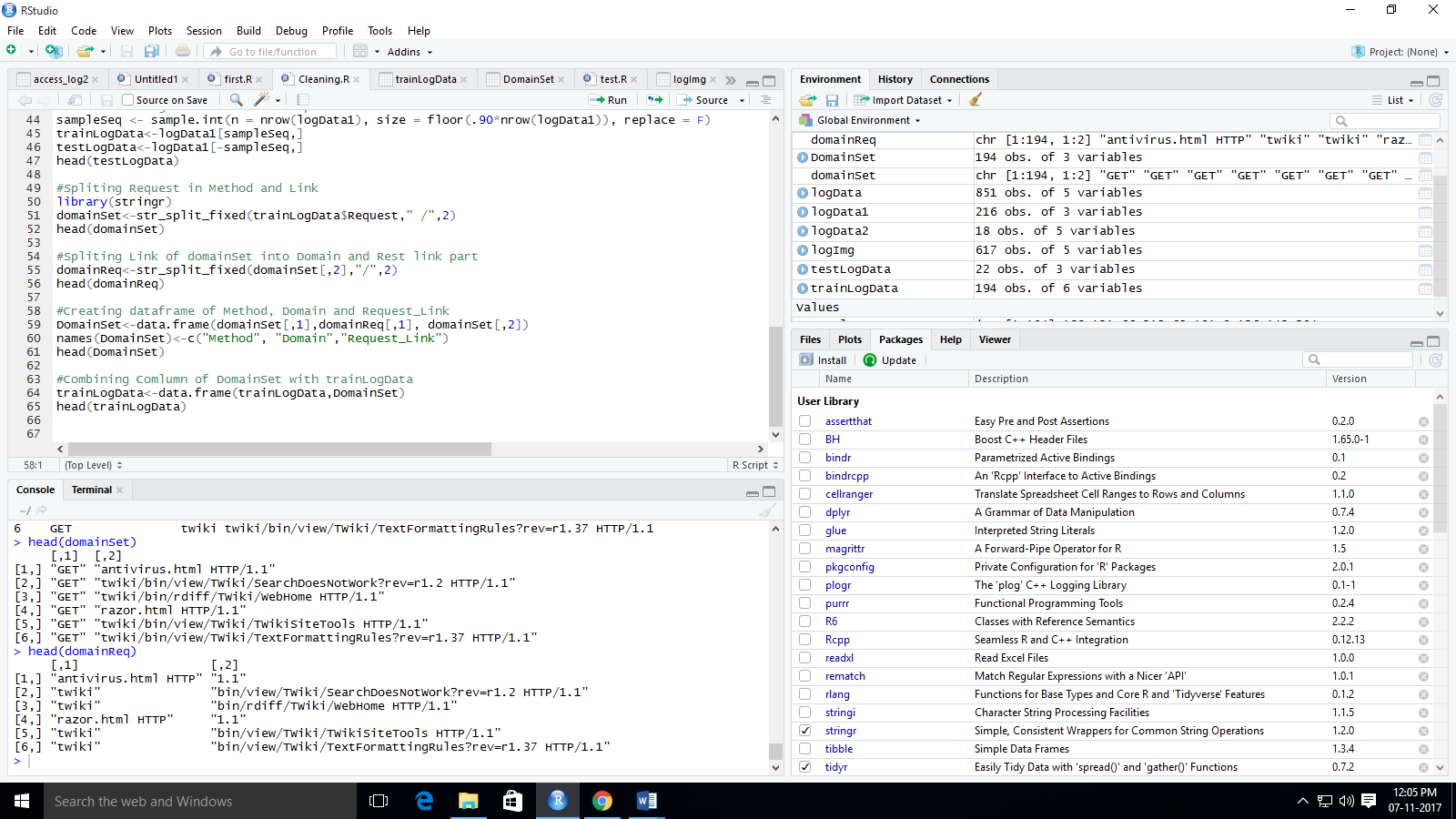


#Spliting Link of domainSet into Domain and Rest link part

domainReq<-str\_split\_fixed(domainSet[,2],"/",2)

head(domainReq)

#output



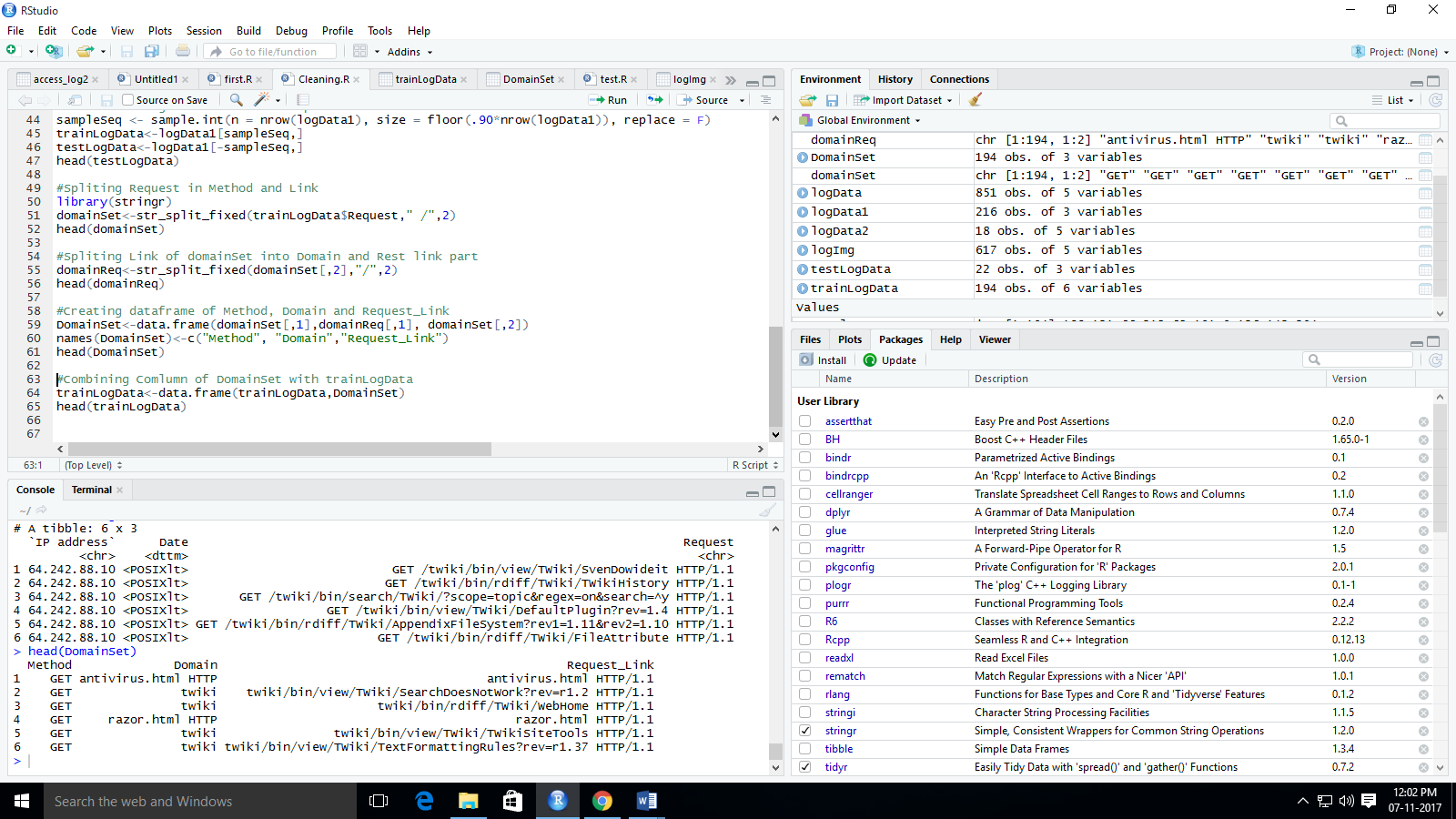
#Creating dataframe of Method, Domain and Request\_Link

DomainSet<-data.frame(domainSet[,1],domainReq[,1], domainSet[,2])

names(DomainSet)<-c("Method", "Domain","Request\_Link")

head(DomainSet)

#ouput

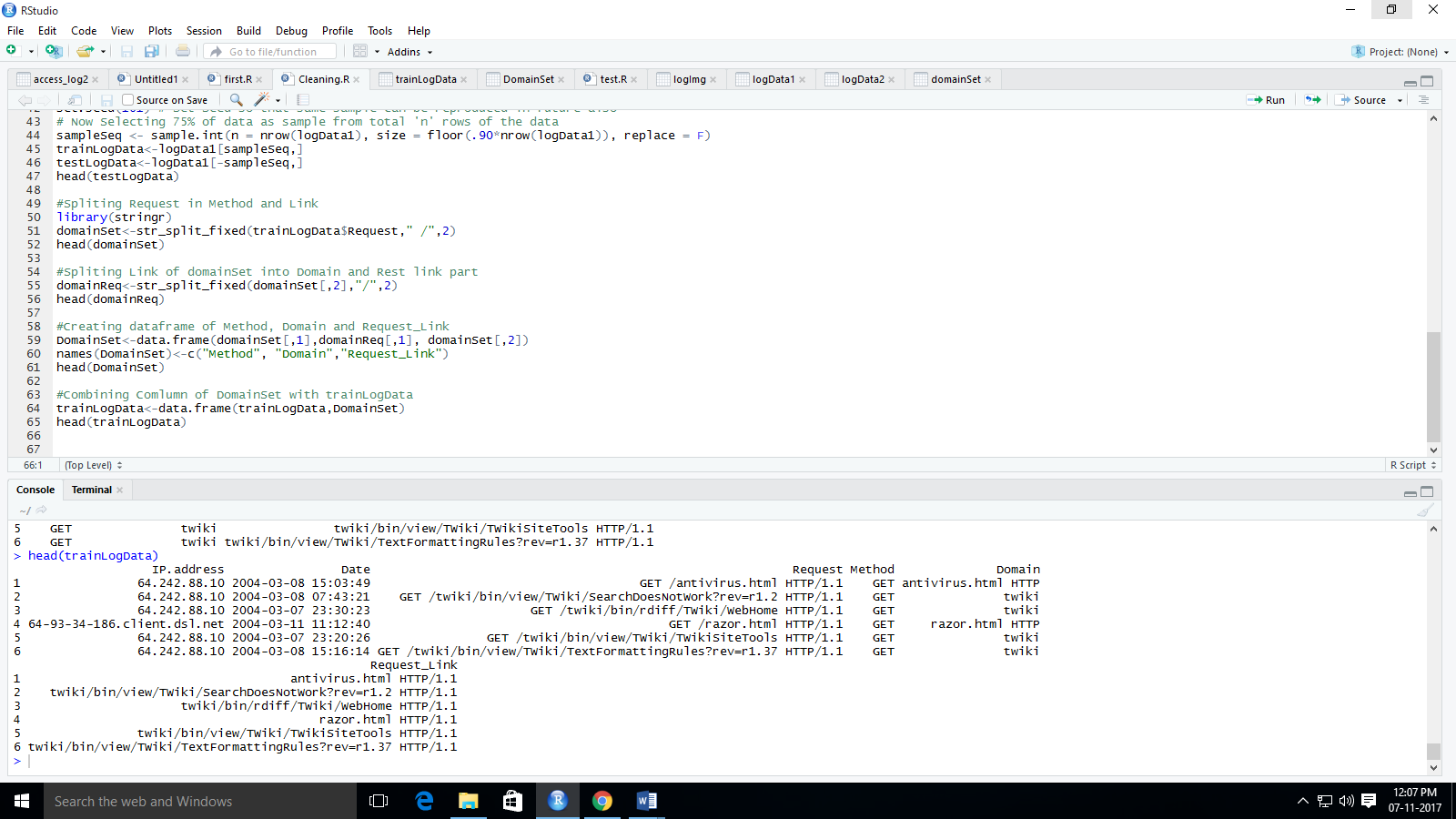


#Combining Comlumn of DomainSet with trainLogData

trainLogData<-data.frame(trainLogData,DomainSet)

head(trainLogData)

#output



References :

International Journals

1. Evaluation of Models for Predicting User’s Next Request in Web Usage Mining.
2. Predicting Web User's Next Access Based on Log Data
3. User Future Request Prediction Using KFCM in Web Usage Mining
4. Web Usage mining for Predicting User Access Behaviour

IEEE and ScienceDirect sites for studying web usage mining and web log mining.

DataCamp for R programming.

Data Mining courses from data camp, youtube.