HW_TechCrunch

John Antony

2/14/2017

Question 1: Extract all articles from the TechCrunch main page of their web site. Create a data frame of these articles.

```
library(rvest)
```

```
## Loading required package: xml2
```

```
url = "https://techcrunch.com/"

doc.html = read_html(url)
text = doc.html %>% html_nodes(".excerpt") %>% html_text()

text = gsub("[\t\n]","",text)

df = as.data.frame(text)
dim(df)
```

```
## [1] 19  1
```

```
head(df$text)
```

```
## [1] Microsoft CVP Yusuf Mehdi isn't shy about mentioning the competition. As we discuss the company's recent l
aunch of Windows 10 S, he refers to the pared down operating system as a "Chromebook compete." The company knows
 that Google is eating its lunch in K-12 here in the States, so it answered with its own version, and it's not re
ally making any bones about it.Gone… Read More
## [2] People may well remember 2017 as the year that blockchain broke.After years of development and flickering
 just outside of mainstream consciousness and acceptance, record high prices for the most popular blockchain-base
d cryptocurrencies Bitcoin and newcomer Ethereum, and an embrace of the technology's core principles by some of t
he world's largest institutions may mean that… Read More
## [3] Google's Jamboard is not a kitchen app for curating PB&J recipes — it's a 55-inch digital whiteboard, with
pen and touch input, companion iOS and Android apps, an Nvidia Jetson TX1 processor on board and 4K resolution. T
he behemoth is an enterprise-focused collaboration tool, that comes in three fun colors and has a stand that look
s ripped from a Herman Miller catalog, and… Read More
## [4] Printing static photos from your phone was so last year. Now, it's all about AR.Prynt, a startup that make
s a small mobile printer that connects to your phone, has just released their newest device, called Prynt Pocket.
And the device puts a renewed focus on the company's AR play, which essentially makes each printed photo a Harry
 Potter-style moving image.When you select a photo… Read More
## [5] There are inefficiencies in the giant self-storage market, and startups like MakeSpace and Clutter have al
ready been gaining traction by offering a more consumer-friendly approach.Now there's another startup entering th
e picture, with Trove formally launching in the San Francisco Bay Area after a few months of testing the concept
 in beta. They are also securing $8 million in a… Read More
## [6] Thanks to James Corden, CBS will now join a number of other major media publishers in creating content for
Snapchat's Shows — the social service's smattering of short-form original video series found in the app's "Discov
er" section. The network announced this morning that James Corden will star in a new show on Snapchat called "Jam
es Corden's… Read More
## 19 Levels: Amazon this morning staked another claim in the billion-dollar wedding industry with the launch of
 a new Wedding Shop dedicated to sales of handcrafted items, including décor, invitations, gifts, jewelry, access
ories and more. The shop is an offshoot from the retailer's less than two-year old Etsy competitor, Handmade at A
mazon, which aims to offer the same sort of original… Read More ...
```

Question 2: Create a corpus of these articles, and then clean it of numbers, punctuation, stopwords, and stem the documents as well.

```
library(stringr)
library(tm)
```

```
## Loading required package: NLP
```

```
ctext = Corpus(VectorSource(df))
print(ctext)
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 1
```

```
clean_web_page = function(text,cstem=0,cstop=0,ccase=0,cpunc=0,cflat=0,cNumber=1) {
    #text = readLines(url)
    text = str_replace_all(text,"[<>{}()&;,.\n]"," ")
    #text = str_replace_all(text,"/(Read More)/g"," ")

    text = text[setdiff(seq(1,length(text)),grep("<",text))]
    text = text[setdiff(seq(1,length(text)),grep(">",text))]
    text = text[setdiff(seq(1,length(text)),grep("]",text))]
    text = text[setdiff(seq(1,length(text)),grep("}",text))]
    text = text[setdiff(seq(1,length(text)),grep("_",text))]
    text = text[setdiff(seq(1,length(text)),grep("\\/",text))]
    ctext = Corpus(VectorSource(text))

    if (cstop==1) { ctext = tm_map(ctext, removeWords, stopwords("english"))
                  }
    if (cpunc==1) { ctext = tm_map(ctext, removePunctuation) }
    if (cstem==1) { ctext = tm_map(ctext, stemDocument) }
    if (ccase==1) { ctext = tm_map(ctext, tolower) }
    if (ccase==2) { ctext = tm_map(ctext, toupper) }
    if (cNumber==1){ctext = tm_map(ctext, removeNumbers)}
    ctext = tm_map(ctext,removeWords,"Read More")
    text = ctext
    #CONVERT FROM CORPUS IF NEEDED
    if (cflat>0) {
        text = NULL
        for (j in 1:length(ctext)) {
            temp = ctext[[j]]$content
            if (temp!="") { text = c(text,temp) }
        }
        text = as.array(text)
    }
    if (cflat==1) {
        text = paste(text,collapse="\n")
        text = str_replace_all(text, "[\r\n]" , " ")
    }
    result = text
}

webpage = clean_web_page(text,cstem=1,cstop=1,ccase=0,cpunc=1,cflat=0,cNumber = 1)
```

Question 3: Create a term document matrix from the corpus and print the top 50 lines. What is the dimension of the TDM?

```
tdm = TermDocumentMatrix(webpage,control=list(minWordLength=1))
#inspect(tdm[1:30,1:10])
print(lapply(webpage[1:5],as.character))
```

```
## $`1`
## [1] "Microsoft CVP Yusuf Mehdi isnt shi  mention  competit  As  discuss  compani recent launch  Window  S   re
fer   pare  oper system   Chromebook compet  The compani know  Googl  eat  lunch  K    State    answer    version
   s  realli make  bone   Gone "
##
## $`2`
## [1] "Peopl may well rememb    year  blockchain broke After year  develop  flicker just outsid  mainstream cons
cious  accept  record high price    popular blockchainbas cryptocurr Bitcoin  newcom Ethereum    embrac   technol
og core principl     world largest institut may mean  "
##
## $`3`
## [1] "Googl Jamboard    kitchen app  curat PB J recip  s  inch digit whiteboard   pen  touch input  companion i
OS  Android app   Nvidia Jetson TX processor  board  K resolut  The behemoth   enterprisefocus collabor tool   co
me  three fun color    stand  look rip   Herman Miller catalog   "
##
## $`4`
## [1] "Print static photo   phone   last year  Now  s   AR Prynt   startup  make  small mobil printer  connect
phone   just releas  newest devic  call Prynt Pocket  And  devic put  renew focus   compani AR play   essenti mak
e  print photo  Harri Potterstyl move imag When  select  photo "
##
## $`5`
## [1] "There  ineffici   giant selfstorag market   startup like MakeSpac  Clutter  alreadi  gain traction  offer
   consumerfriend approach Now s anoth startup enter  pictur   Trove formal launch   San Francisco Bay Area    mo
nth  test  concept  beta  They  also secur  million   "
```

```
tdm
```

```
## <<TermDocumentMatrix (terms: 505, documents: 19)>>
## Non-/sparse entries: 634/8961
## Sparsity           : 93%
## Maximal term length: 15
## Weighting          : term frequency (tf)
```

Question 4: From the TDM, make a network adjacency matrix of words. Assume two words are linked once if they appear in the same document. If they co-occur in say, three documents, then they are connected by strength 3. Based on co-occurrence withim documents, create the adjacency matrix of words, where the words are nodes, and their co-occurrences provide the data for the links.

```
dtdm <- as.matrix(tdm)
dtdm[1:10,1:10]
```

```
##             Docs
## Terms        1 2 3 4 5 6 7 8 9 10
##   abil       0 0 0 0 0 0 0 0 0  0
##   accept     0 1 0 0 0 0 0 0 0  0
##   accessori  0 0 0 0 0 0 0 0 0  0
##   account    0 0 0 0 0 0 0 0 0  0
##   acquir     0 0 0 0 0 0 0 0 1  0
##   addit      0 0 0 0 0 0 0 0 0  0
##   advertis   0 0 0 0 0 0 0 0 0  1
##   afoot      0 0 0 0 0 0 0 0 1  0
##   after      0 1 0 0 0 0 0 0 0  0
##   aftermarket 0 0 0 0 0 0 0 0 0  0
```

```
# change it to a Boolean matrix
dtdm[dtdm>=1] <- 1
# transform into a term-term adjacency matrix (for finding the adjacent words)
termMatrix  <- dtdm %*% t(dtdm)
# inspect terms numbered 1 to 10
termMatrix[10:20,10:20]
```

```
##              Terms
## Terms         aftermarket aim allow alreadi also alum amazon and android
##   aftermarket           1   0     0       0    1    0      0   0       0
##   aim                   0   1     0       0    0    0      1   0       0
##   allow                 0   0     2       0    0    0      0   1       0
##   alreadi               0   0     0       2    1    0      0   0       0
##   also                  1   0     0       1    2    0      0   0       0
##   alum                  0   0     0       0    0    1      0   0       0
##   amazon                0   1     0       0    0    0      1   0       0
##   and                   0   0     1       0    0    0      0   4       0
##   android               0   0     0       0    0    0      0   0       1
##   announc               0   0     1       1    0    1      0   0       0
##   anoth                 0   1     0       1    1    0      1   0       0
##              Terms
## Terms         announc anoth
##   aftermarket       0     0
##   aim               0     1
##   allow             1     0
##   alreadi           1     1
##   also              0     1
##   alum              1     0
##   amazon            0     1
##   and               0     0
##   android           0     0
##   announc           5     0
##   anoth             0     2
```

Question 5: Convert the Adjacency Matrix into a Edge List (This is just a two column listing of nodes from and nodes to.)

```
library(igraph)
```

```
##
## Attaching package: 'igraph'
```

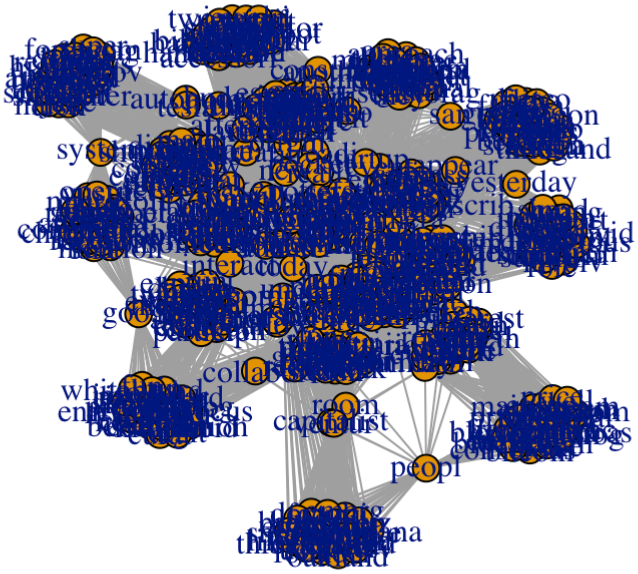
```
## The following object is masked from 'package:stringr':
##
##     %>%
```

```
## The following object is masked from 'package:rvest':
##
##     %>%
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

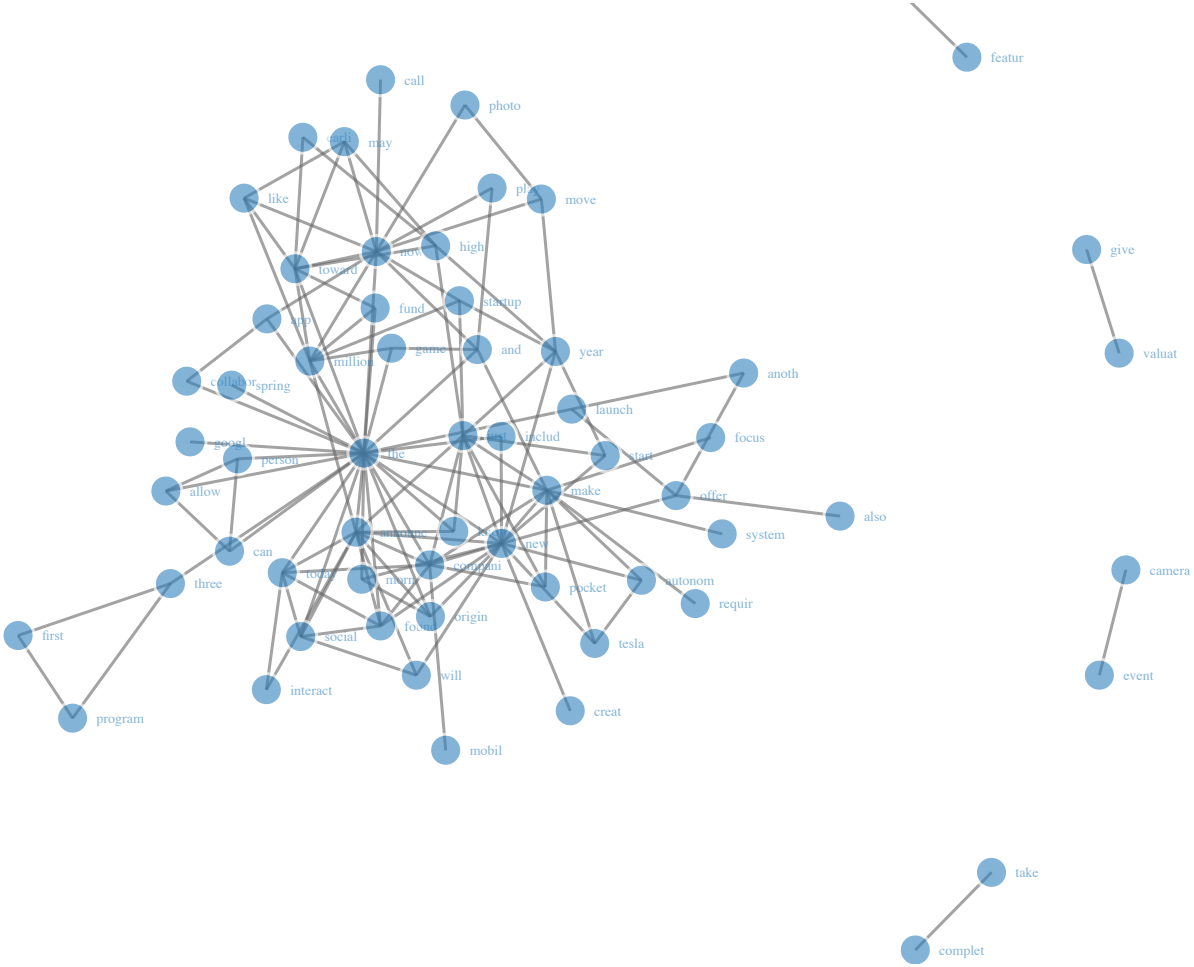
```
g <- graph.adjacency(termMatrix,weighted=T, mode = "undirected")
g <- simplify(g)

V(g)$label <- V(g)$name
V(g)$degree <- degree(g)

set.seed(70000)
plot.igraph(g,layout=layout.fruchterman.reingold,edge.arrow.size=0.5,vertex.size=10)
```



```
edge_list = get.edgelist(g)
```

Question 6: Using the edge list, create a spring force plot using D3. Redo the same plot, but zero out all edges which have value 1, and keep all edges with values 2 or greater. How different are the two plots, describe the difference.

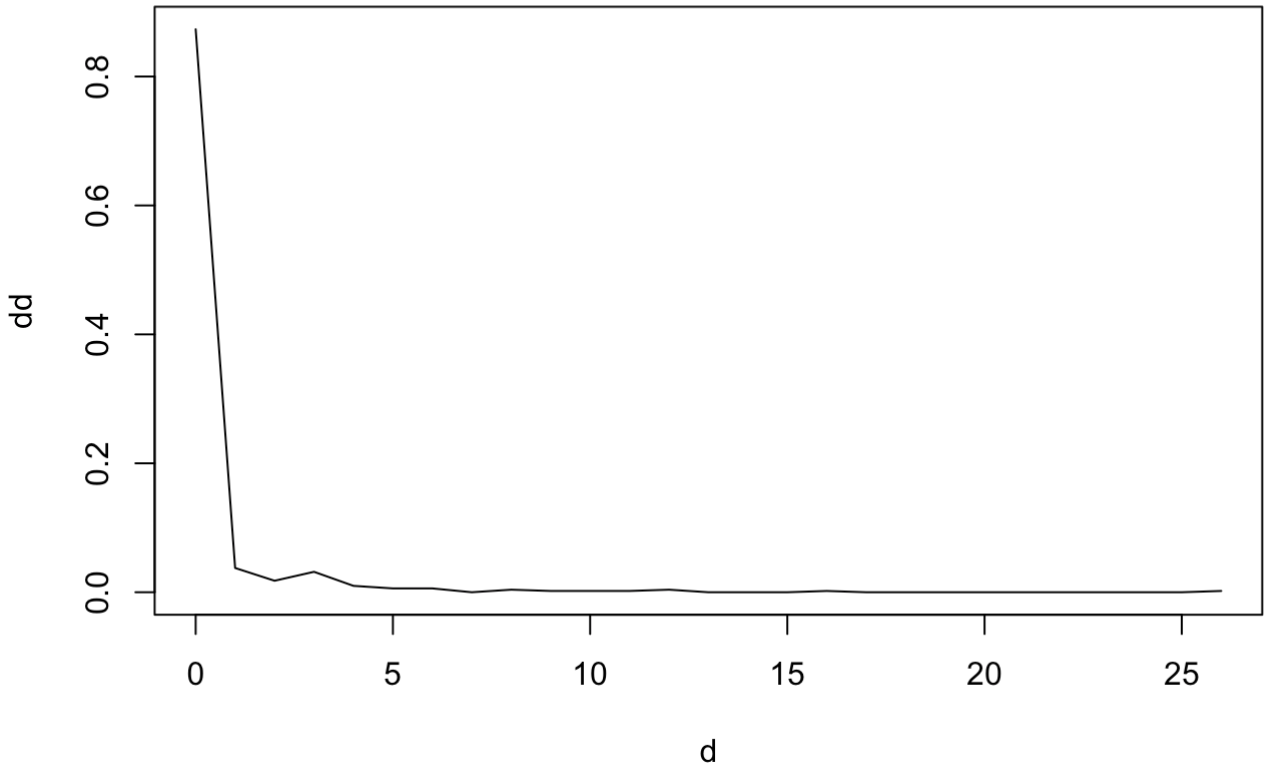
```
g2 = delete_edges(g, which(E(g)$weight < 2))
#plot(g2)
edge_list2 = get.edgelist(g2)
library(networkD3)
networkData = data.frame(edge_list2[,1], edge_list2[,2])
simpleNetwork(networkData)
```



As we can see here, the nodes have decreased significantly also expanding to the

Question 7: Plot the degree distribution of your word network.

```
#COMPUTE DEGREE DISTRIBUTION
dd = degree.distribution(g2)
dd = as.matrix(dd)
d = as.matrix(seq(0,max(degree(g2))))
plot(d,dd,type="l")
```



Question 8: Calculate the centrality of the words in the network. What are the top 10 central words? What conclusions can you state from this?

```
res = evcent(g2,scale = FALSE)
print(names(res))
```

```
## [1] "vector"  "value"   "options"
```

```
vect = res$vector

top10centralwords = tail(sort(vect),10)
print(names(top10centralwords))
```

```
##  [1] "million" "just"    "make"    "origin"  "today"   "found"   "compani"
##  [8] "new"     "announc" "the"
```

Question 9: Form “communities” of words, and state any regularities you may see from these communities. You may decide on the setting for the size of these communities as you prefer.

```
wtc = walktrap.community(g2)
res=membership(wtc)
print(res)
```

```
##           abil         accept      accessori        account
##             14             15             16             17
##         acquir          addit       advertis         afoot
##             18             19             20             21
##          after     aftermarket            aim         allow
##             22             23             24              8
##        alreadi           also           alum         amazon
##             25              3             26             27
##            and        android        announc          anoth
##              1             28              2              3
##         answer          ape            app         appear
##             29             30              2             31
##          apple        appoint    appointment       approach
##             32             33             34             35
##           area         around       autonom      autopilot
##             36             37              4             38
##           back          bad           bank           base
##             39             40             41             42
##        basebal          bath            bay          begin
##             43             44             45             46
##       behemoth         belong            bet           beta
##             47             48             49             50
##   billiondollar        bitcoin     blockchain   blockchainbas
##             51             52             53             54
##           blue         board           bone            bot
##             55             56             57             58
##          brand     breathalyz         bring          broke
##             59             60             61             62
##          buddi   budgetfriend          buoy           busi
##             63             64             65             66
##           but          calif           call         camera
##             67             68              1              9
##           can          candi          capit     capitalist
##              8             69             70             13
##         captur           car         career          cash
##             71             72             73             74
##        catalog          cbs            ceo          chang
##             75             76             77             78
##          chat        chatbot     chromebook          claim
##             79             80             81             82
```

```
##           clean           close         clutter        cocktail
##              83              84              85              86
##        collabor           color            come       communiti
##               2              87              88              89
##         compani       companion          compet         competit
##               4              90              91              92
##      competitor         complet         concept         confirm
##              93              12              94              95
##         connect       conscious        consolid          consum
##              96              97              98              99
##  consumerfriend         content          corden            core
##             100             101             102             103
##           coupl           cours           creat      cryptocurr
##             104             105               2             106
##           curat             cvp            deal           décor
##             107             108             109             110
##           dedic            deep          delici          denis
##             111             112             113             114
##          deputi          design          detail         develop
##             115             116             117             118
##           devic           diego           digit          direct
##             119             120             121             122
##         disagre          discov         discuss            dish
##             123             124             125             126
##          divers          django          doctor          dollar
##             127             128             129             130
##            down           drive           drone           earli
##             131             132             133               6
##             eat          effect            elon          embrac
##             134             135             136             137
##           emerg           engag           enjoy           enter
##             138             139             140             141
## enterprisefocus           entir          equiti          especi
##             142             143             144             145
##         essenti        ethereum            etsi           event
##             146             147             148               9
##           everi          exampl          expens          experi
##             149             150             151             152
##         explain        facebook         faculti             fan
##             153             154             155             156
##        fdicinsur          featur          financ            find
```

```
##             157              10             158             159
##           first         firstev            flew             fli
##               5             160             161             162
##         flicker           focus             for          formal
##             163               4             164             165
##          former        forthcom           found         founder
##             166             167               2             168
##       francisco          friend             fun            fund
##             169             170             171               6
##            gain            game            gear           giant
##             172               1             173             174
##            gift         gimmick            give            gone
##             175             176              11             177
##            good           googl           gopro            grow
##             178               2             179             180
##       handcraft         handmad           harri            head
##             181             182             183             184
##            help          herman            high         highway
##             185             186               6             187
##             hit          hoagi           hound            hour
##             188             189             190             191
##           human           idea           illeg            imag
##             192             193             194             195
##          improv           inch          includ          inclus
##             196             197               2             198
##           indic        industri        ineffici         ingredi
##             199             200             201             202
##         innoviz          input        institut        interact
##             203             204             205               2
##           invit         involv             ios            isnt
##             206             207             208             209
##            item        jamboard            jame          jetson
##             210             211             212             213
##         jewelri            join         journey            june
##             214             215             216             217
##            just           karma             key         kitchen
##               7             218             219             220
##            know             lab         largest            last
##               2             221             222             223
##          launch             len            less           lidar
##               3             224             225             226
```

```
##         lifetim             like             live           london
##             227                6              228              229
##             long         longterm             look         lotteri
##             230              231              232              233
##           lunch             lynn              mac       mainstream
##             234              235              236              237
##            major             make         makespac             mani
##             238                4              239              240
##        marijuana           market              may             mean
##             241              242                6              243
##            media             meet            mehdi           member
##             244              245              246              247
##          mention           messag       microsoft           midsiz
##             248              249              250              251
##            might             mike           miller          million
##             252              253              254                6
##             mind              mlb            mobil           monday
##             255              256              257              258
##            month             morn             move             musk
##             259                2                1              260
##             must          network              new           newcom
##             261              262                2              263
##           newest             news             next            night
##             264              265              266              267
##              now           number           nvidia          oakland
##                1              268              269              270
##            offer           offici         offshoot              old
##                3              271              272              273
##             once              one             oper           origin
##             274              275              276                2
##           outsid           overal     oversubscrib           packag
##             277              278              279              280
##             padr             pare             park             part
##             281              282              283              284
##         particip      partnership              pen          pentech
##             285              286              287              288
##            peopl          percent          perform           person
##             289              290              291                8
##            petco            phone            photo           pictur
##             292              293                1              294
##        pinterest             plan         platform             play
```

```
##          295          296          297            1
##     playfish       pocket      popular   potterstyl
##          298            4          299          300
##     previous        price     principl        print
##          301          302          303          304
##      printer       prize      probabl    processor
##          305          306          307          308
##      program     prohibit       promot       prynt
##            5          309          310          311
##          psa      publish         pull          put
##          312          313          314          315
##        quick        rais     reaction      realiti
##          316          317          318          319
##       realli       recent        recip      recogn
##          320          321          322          323
##       record        refer      regard       relat
##          324          325          326          327
##       releas       rememb       renew      report
##          328          329          330          331
##      request       requir    research      reserv
##          332            4          333          334
##      resolut      resourc      restaur      retail
##          335          336          337          338
##        ride         ridg      ridicul         rip
##          339          340          341          342
##       ritual         role         roll        room
##          343          344          345          346
##          run         sale          san        save
##          347          348          349          350
##          say       search      section       secur
##          351          352          353          354
##          see         seem       select  selfstorag
##          355          356          357          358
##         sens         seri         serv      servic
##          359          360          361          362
##        sever      sheriff         shi         shop
##          363          364          365          366
##    shortform         show         sinc         size
##          367          368          369          370
##        small      smatter        smith     smoother
##          371          372          373          374
```

```
##        snapchat          social         softwar            some
##             375               2             376             377
##         someon         someth         sometim            sort
##             378             379             380             381
##         sound          space         spring        stadium
##             382             383               2             384
##         stage          stake         stand           star
##             385             386             387             388
##         start        startup         state         static
##               7               7             389             390
##          step        strateg         studio         super
##             391             392             393             394
##       supercel        surpass        surpris       surround
##             395             396             397             398
##        system           take         target           team
##               4              12             399             400
##     techcrunch       technolog            ten          tesla
##             401             402             403               4
##          test          thank         that           the
##             404             405             406               2
##          then          there         they         third
##             407             408             409             410
##         though        thousand         three    threeyearold
##             411             412               5             413
##         thrive           tie          time         today
##             414             415             416               2
##         togeth          tool         total         touch
##             417             418             419             420
##         toward       traction         trend         trial
##               6             421             422             423
##         trove         tweet         twist        twitter
##             424             425             426             427
##        twoyear          ucsf     understand          unit
##             428             429             430             431
##         updat           use          user         valuat
##             432             433             434              11
##        ventur        version         video       virginia
##              13             435             436             437
##          want           way          wed          well
##             438              10             439             440
##          when          while     whiteboard          will
```

```
##         441         442         443           2
##         win      window      within        work
##         444         445         446         447
##       world    worldwid       write        year
##         448         449         450           7
##         yes   yesterday         you       young
##         451         452         453         454
##       yusuf
##         455
```

Question 10: What is the diameter of the network? Why is this interesting?

```
print(diameter(g))
```

```
## [1] 3
```

Diameter is the longest shortest distance from a node to any other noder. It useful to know how quickly something might spread.