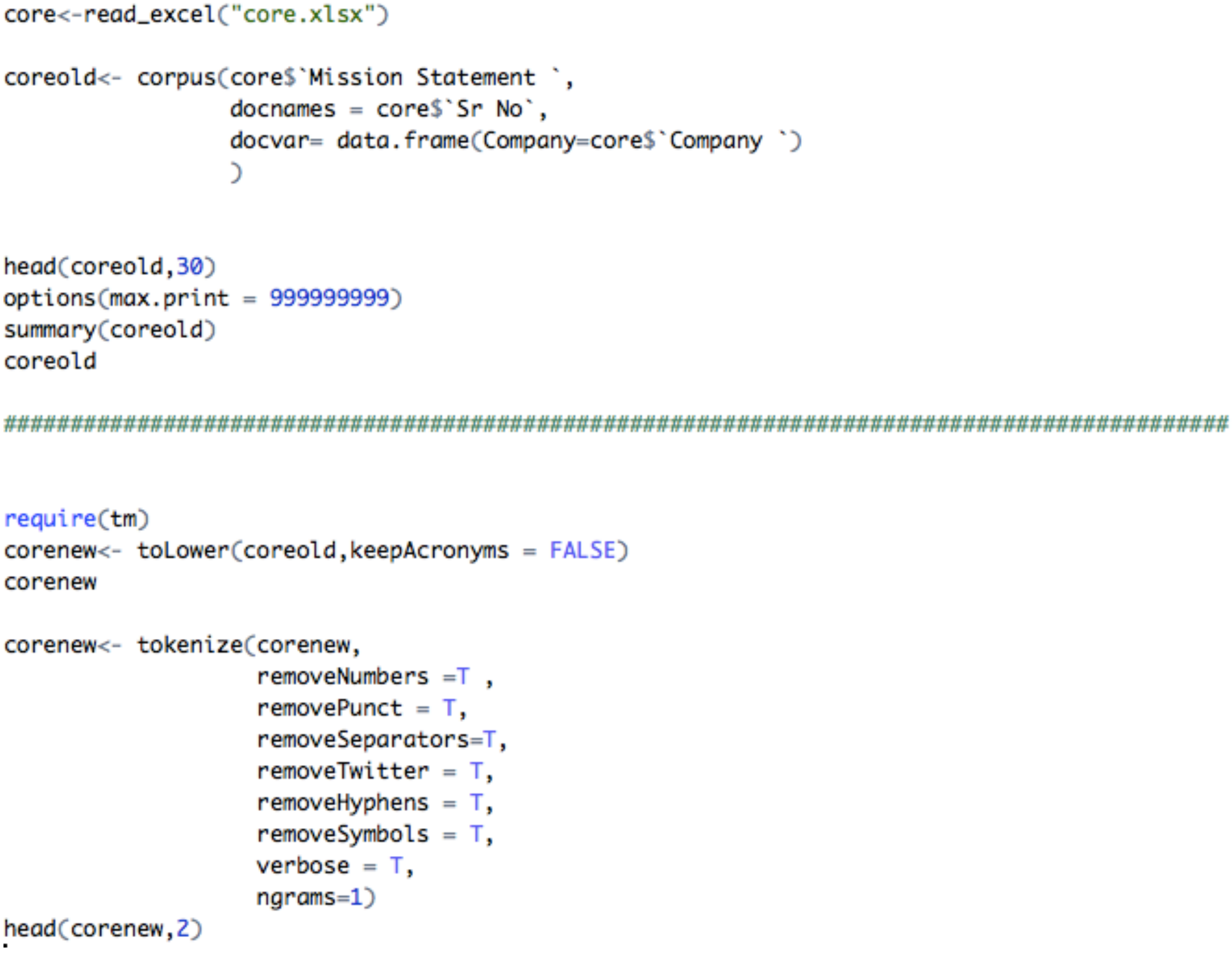
Assignment 7

Arnav Somani
Fall 2016

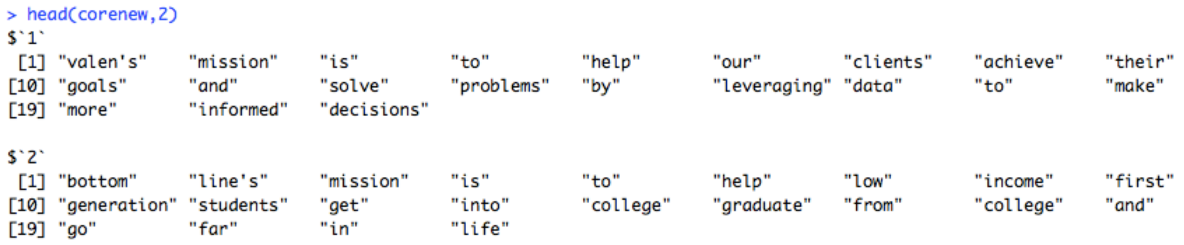
Part 1

Q.1) Find a list of successful data-driven companies (See this Forbes article for inspiration) and select 30 organizations.
A. The list in excel form which I have attached with this assignment.

Q.2) Create a corpus of their mission statements
A.

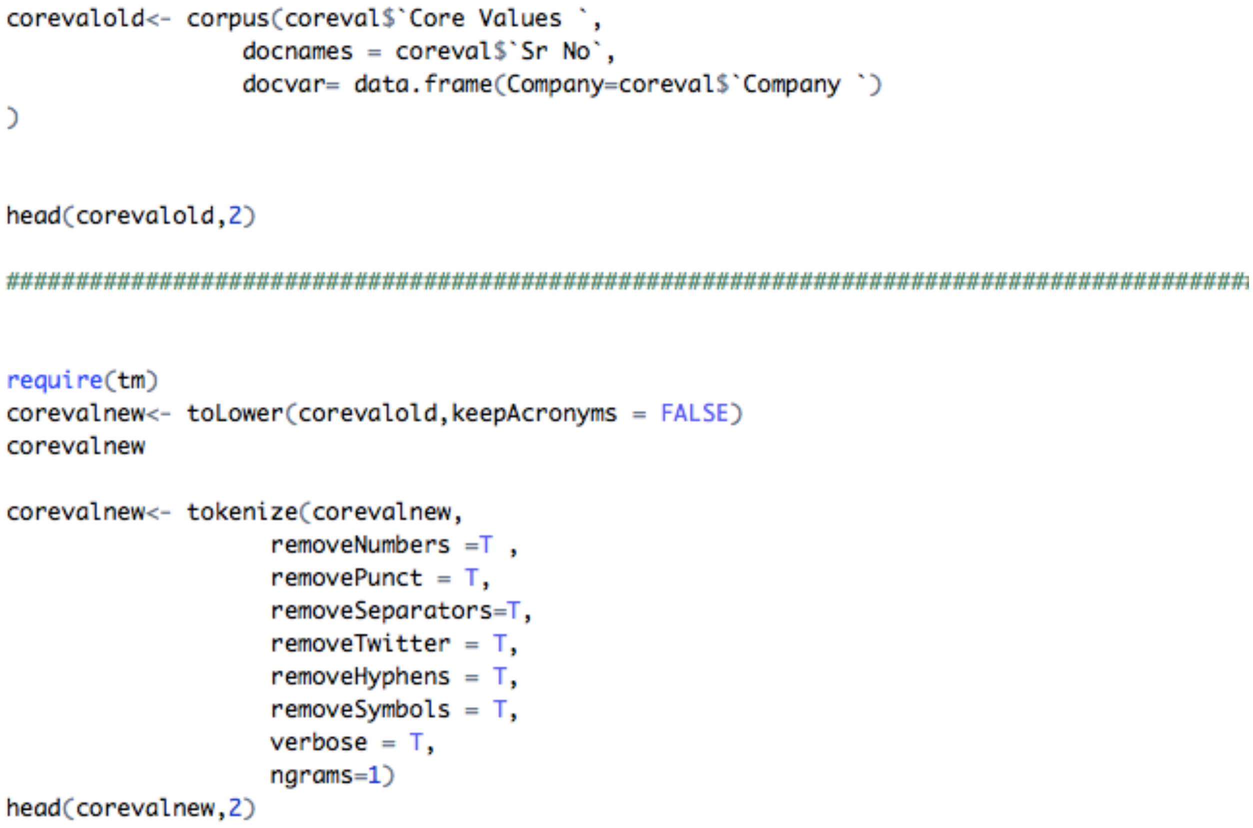


```
core<-read_excel("core.xlsx")

coreold<- corpus(core$`Mission Statement `,
                 docnames = core$`Sr No`,
                 docvar= data.frame(Company=core$`Company `)
                 )


head(coreold,30)
options(max.print = 999999999)
summary(coreold)
coreold

############################################################################################


require(tm)
corenew<- toLower(coreold,keepAcronyms = FALSE)
corenew

corenew<- tokenize(corenew,
                   removeNumbers =T ,
                   removePunct = T,
                   removeSeparators=T,
                   removeTwitter = T,
                   removeHyphens = T,
                   removeSymbols = T,
                   verbose = T,
                   ngrams=1)
head(corenew,2)
```

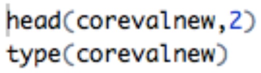


```
> head(corenew,2)
$`1`
 [1] "valen's"    "mission"    "is"         "to"         "help"       "our"        "clients"    "achieve"    "their"
[10] "goals"      "and"        "solve"      "problems"   "by"         "leveraging" "data"       "to"         "make"
[19] "more"       "informed"   "decisions"

$`2`
 [1] "bottom"     "line's"     "mission"    "is"         "to"         "help"       "low"        "income"     "first"
[10] "generation" "students"   "get"        "into"       "college"    "graduate"   "from"       "college"    "and"
[19] "go"         "far"        "in"         "life"
```

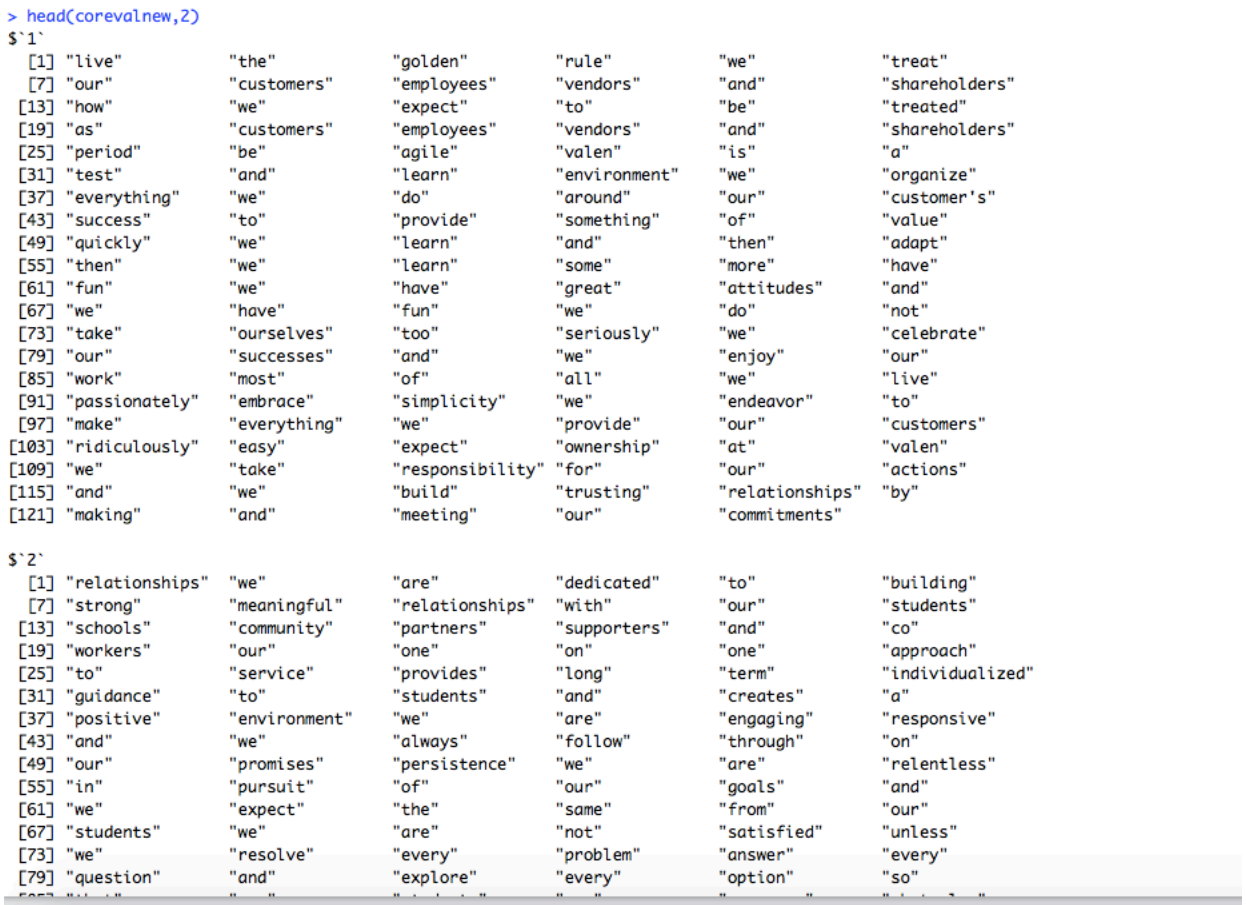
Q.3) Create a corpus of their core values
A.



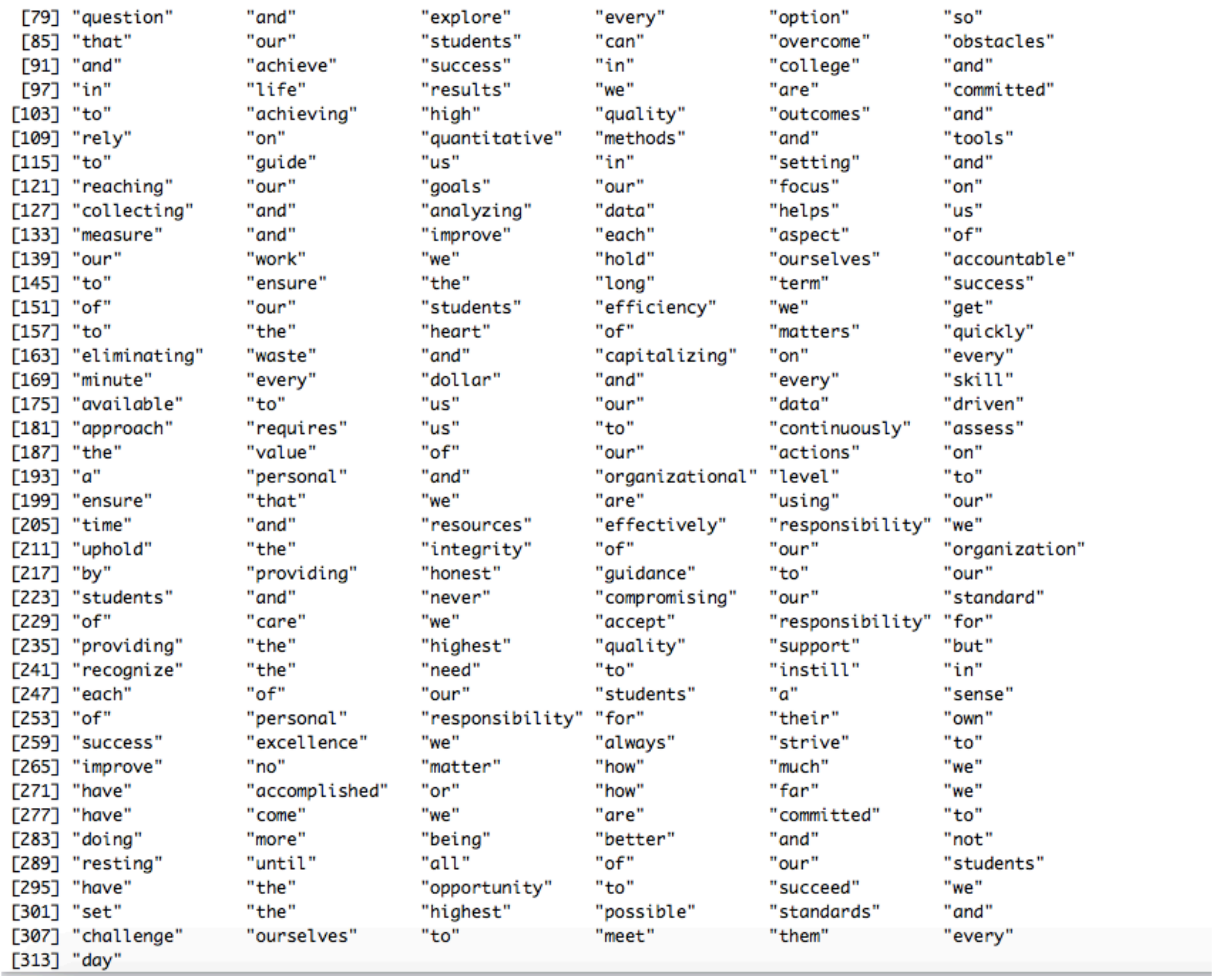
```
corevalold<- corpus(coreval$`Core Values `,
                    docnames = coreval$`Sr No`,
                    docvar= data.frame(Company=coreval$`Company `)
)


head(corevalold,2)

#########################################################################################


require(tm)
corevalnew<- toLower(corevalold,keepAcronyms = FALSE)
corevalnew

corevalnew<- tokenize(corevalnew,
                      removeNumbers =T ,
                      removePunct = T,
                      removeSeparators=T,
                      removeTwitter = T,
                      removeHyphens = T,
                      removeSymbols = T,
                      verbose = T,
                      ngrams=1)
head(corevalnew,2)
```



```
head(corevalnew,2)
type(corevalnew)
```



```
> head(corevalnew,2)
$`1`
  [1] "live"          "the"           "golden"         "rule"          "we"              "treat"
  [7] "our"           "customers"     "employees"      "vendors"       "and"             "shareholders"
 [13] "how"           "we"            "expect"         "to"            "be"              "treated"
 [19] "as"            "customers"     "employees"      "vendors"       "and"             "shareholders"
 [25] "period"        "be"            "agile"          "valen"         "is"              "a"
 [31] "test"          "and"           "learn"          "environment"   "we"              "organize"
 [37] "everything"    "we"            "do"             "around"        "our"             "customer's"
 [43] "success"       "to"            "provide"        "something"     "of"              "value"
 [49] "quickly"       "we"            "learn"          "and"           "then"            "adapt"
 [55] "then"          "we"            "learn"          "some"          "more"            "have"
 [61] "fun"           "we"            "have"           "great"         "attitudes"       "and"
 [67] "we"            "have"          "fun"            "we"            "do"              "not"
 [73] "take"          "ourselves"     "too"            "seriously"     "we"              "celebrate"
 [79] "our"           "successes"     "and"            "we"            "enjoy"           "our"
 [85] "work"          "most"          "of"             "all"           "we"              "live"
 [91] "passionately"  "embrace"       "simplicity"     "we"            "endeavor"        "to"
 [97] "make"          "everything"    "we"             "provide"       "our"             "customers"
[103] "ridiculously"  "easy"          "expect"         "ownership"     "at"              "valen"
[109] "we"            "take"          "responsibility" "for"           "our"             "actions"
[115] "and"           "we"            "build"          "trusting"      "relationships"   "by"
[121] "making"        "and"           "meeting"        "our"           "commitments"

$`2`
  [1] "relationships" "we"            "are"            "dedicated"     "to"              "building"
  [7] "strong"        "meaningful"    "relationships"  "with"          "our"             "students"
 [13] "schools"       "community"     "partners"       "supporters"    "and"             "co"
 [19] "workers"       "our"           "one"            "on"            "one"             "approach"
 [25] "to"            "service"       "provides"       "long"          "term"            "individualized"
 [31] "guidance"      "to"            "students"       "and"           "creates"         "a"
 [37] "positive"      "environment"   "we"             "are"           "engaging"        "responsive"
 [43] "and"           "we"            "always"         "follow"        "through"         "on"
 [49] "our"           "promises"      "persistence"    "we"            "are"             "relentless"
 [55] "in"            "pursuit"       "of"             "our"           "goals"           "and"
 [61] "we"            "expect"        "the"            "same"          "from"            "our"
 [67] "students"      "we"            "are"            "not"           "satisfied"       "unless"
 [73] "we"            "resolve"       "every"          "problem"       "answer"          "every"
 [79] "question"      "and"           "explore"        "every"         "option"          "so"
```

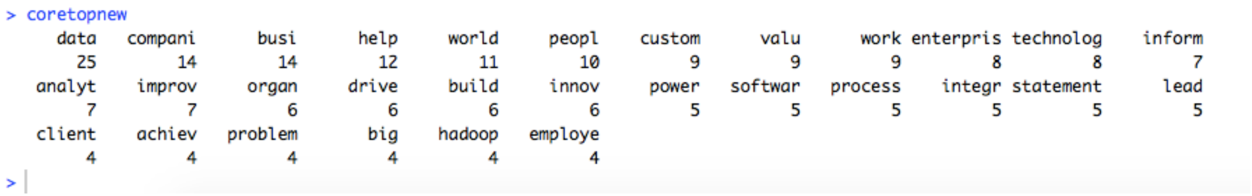


```
 [79] "question"       "and"            "explore"        "every"          "option"         "so"
 [85] "that"           "our"            "students"       "can"            "overcome"       "obstacles"
 [91] "and"            "achieve"        "success"        "in"             "college"        "and"
 [97] "in"             "life"           "results"        "we"             "are"            "committed"
[103] "to"             "achieving"      "high"           "quality"        "outcomes"       "and"
[109] "rely"           "on"             "quantitative"   "methods"        "and"            "tools"
[115] "to"             "guide"          "us"             "in"             "setting"        "and"
[121] "reaching"       "our"            "goals"          "our"            "focus"          "on"
[127] "collecting"     "and"            "analyzing"      "data"           "helps"          "us"
[133] "measure"        "and"            "improve"        "each"           "aspect"         "of"
[139] "our"            "work"           "we"             "hold"           "ourselves"      "accountable"
[145] "to"             "ensure"         "the"            "long"           "term"           "success"
[151] "of"             "our"            "students"       "efficiency"     "we"             "get"
[157] "to"             "the"            "heart"          "of"             "matters"        "quickly"
[163] "eliminating"    "waste"          "and"            "capitalizing"   "on"             "every"
[169] "minute"         "every"          "dollar"         "and"            "every"          "skill"
[175] "available"      "to"             "us"             "our"            "data"           "driven"
[181] "approach"       "requires"       "us"             "to"             "continuously"   "assess"
[187] "the"            "value"          "of"             "our"            "actions"        "on"
[193] "a"              "personal"       "and"            "organizational" "level"          "to"
[199] "ensure"         "that"           "we"             "are"            "using"          "our"
[205] "time"           "and"            "resources"      "effectively"    "responsibility" "we"
[211] "uphold"         "the"            "integrity"      "of"             "our"            "organization"
[217] "by"             "providing"      "honest"         "guidance"       "to"             "our"
[223] "students"       "and"            "never"          "compromising"   "our"            "standard"
[229] "of"             "care"           "we"             "accept"         "responsibility" "for"
[235] "providing"      "the"            "highest"        "quality"        "support"        "but"
[241] "recognize"      "the"            "need"           "to"             "instill"        "in"
[247] "each"           "of"             "our"            "students"       "a"              "sense"
[253] "of"             "personal"       "responsibility" "for"            "their"          "own"
[259] "success"        "excellence"     "we"             "always"         "strive"         "to"
[265] "improve"        "no"             "matter"         "how"            "much"           "we"
[271] "have"           "accomplished"   "or"             "how"            "far"            "we"
[277] "have"           "come"           "we"             "are"            "committed"      "to"
[283] "doing"          "more"           "being"          "better"         "and"            "not"
[289] "resting"        "until"          "all"            "of"             "our"            "students"
[295] "have"           "the"            "opportunity"    "to"             "succeed"        "we"
[301] "set"            "the"            "highest"        "possible"       "standards"      "and"
[307] "challenge"      "ourselves"      "to"             "meet"           "them"           "every"
[313] "day"
```

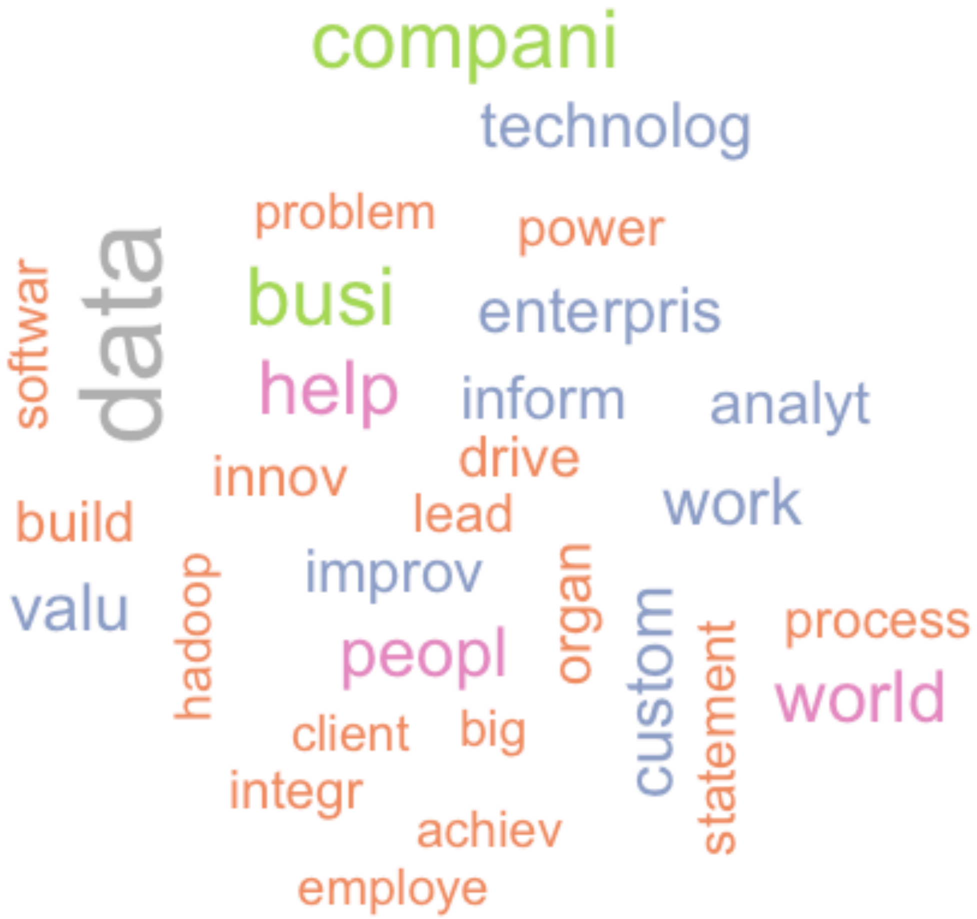
Q.4) Analyze the corpus and provide insight on how to structure a firm for data-analysis readiness

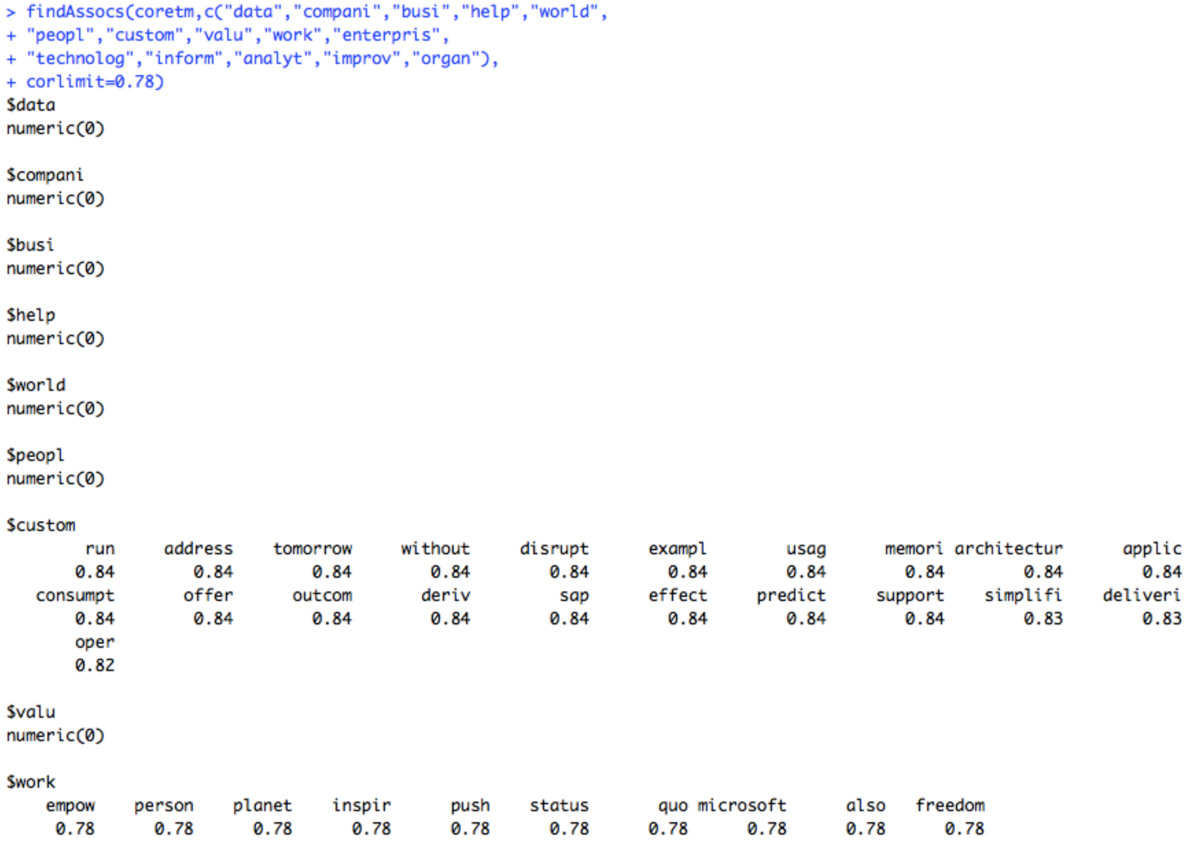
A.

For Mission Statement

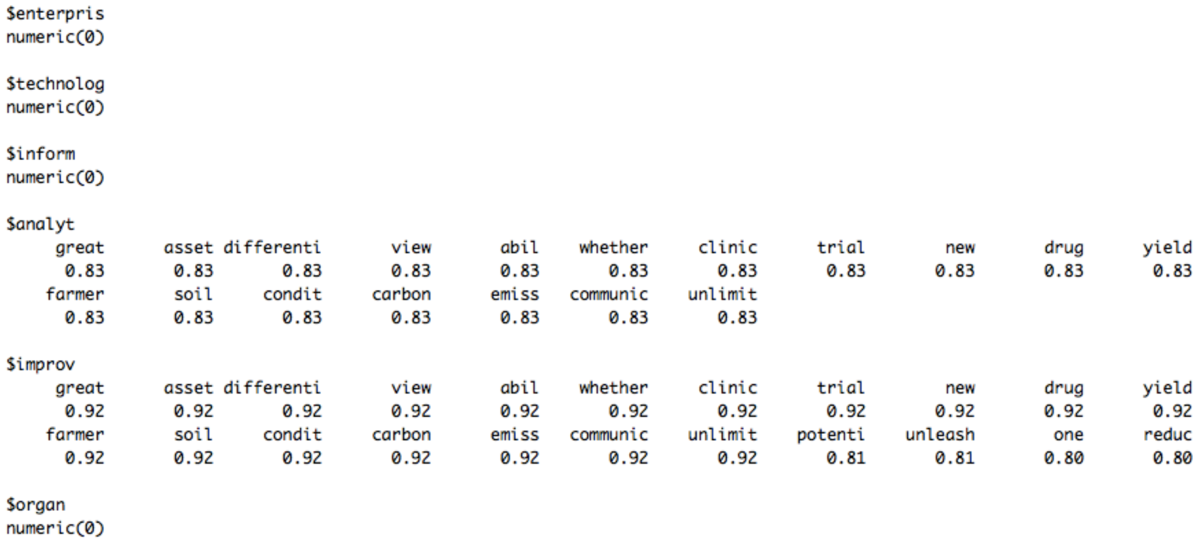


```
> coretopnew
     data    compani       busi       help      world      peopl     custom       valu       work enterpris technolog     inform
       25         14         14         12         11         10          9          9          9         8         8          7
   analyt     improv      organ      drive      build      innov      power    softwar    process    integr statement       lead
        7          7          6          6          6          6          5          5          5         5         5          5
   client     achiev    problem        big     hadoop    employe
        4          4          4          4          4          4
> |
```



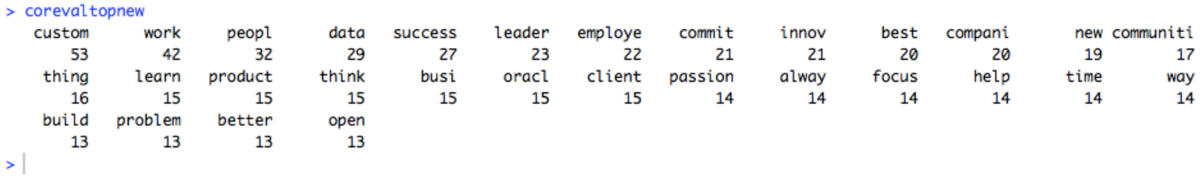


```
> findAssocs(coretm,c("data","compani","busi","help","world",
+ "peopl","custom","valu","work","enterpris",
+ "technolog","inform","analyt","improv","organ"),
+ corlimit=0.78)
$data
numeric(0)

$compani
numeric(0)

$busi
numeric(0)

$help
numeric(0)

$world
numeric(0)

$peopl
numeric(0)

$custom
        run     address    tomorrow     without     disrupt      exampl        usag      memori architectur      applic 
       0.84        0.84        0.84        0.84        0.84        0.84        0.84        0.84        0.84        0.84 
   consumpt       offer     outcom       deriv         sap      effect     predict     support    simplifi    deliveri 
       0.84        0.84        0.84        0.84        0.84        0.84        0.84        0.84        0.83        0.83 
       oper 
       0.82 

$valu
numeric(0)

$work
    empow    person    planet    inspir      push    status       quo microsoft      also   freedom 
     0.78      0.78      0.78      0.78      0.78      0.78      0.78      0.78      0.78      0.78 
```

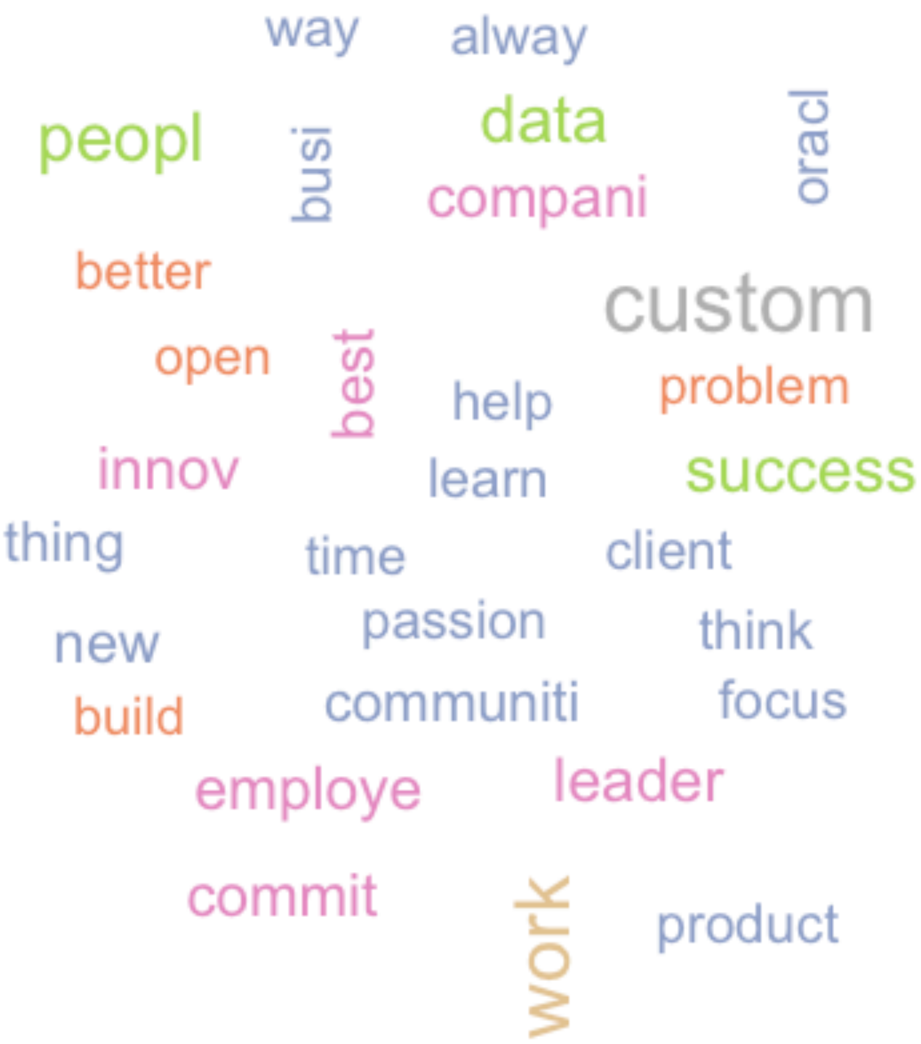


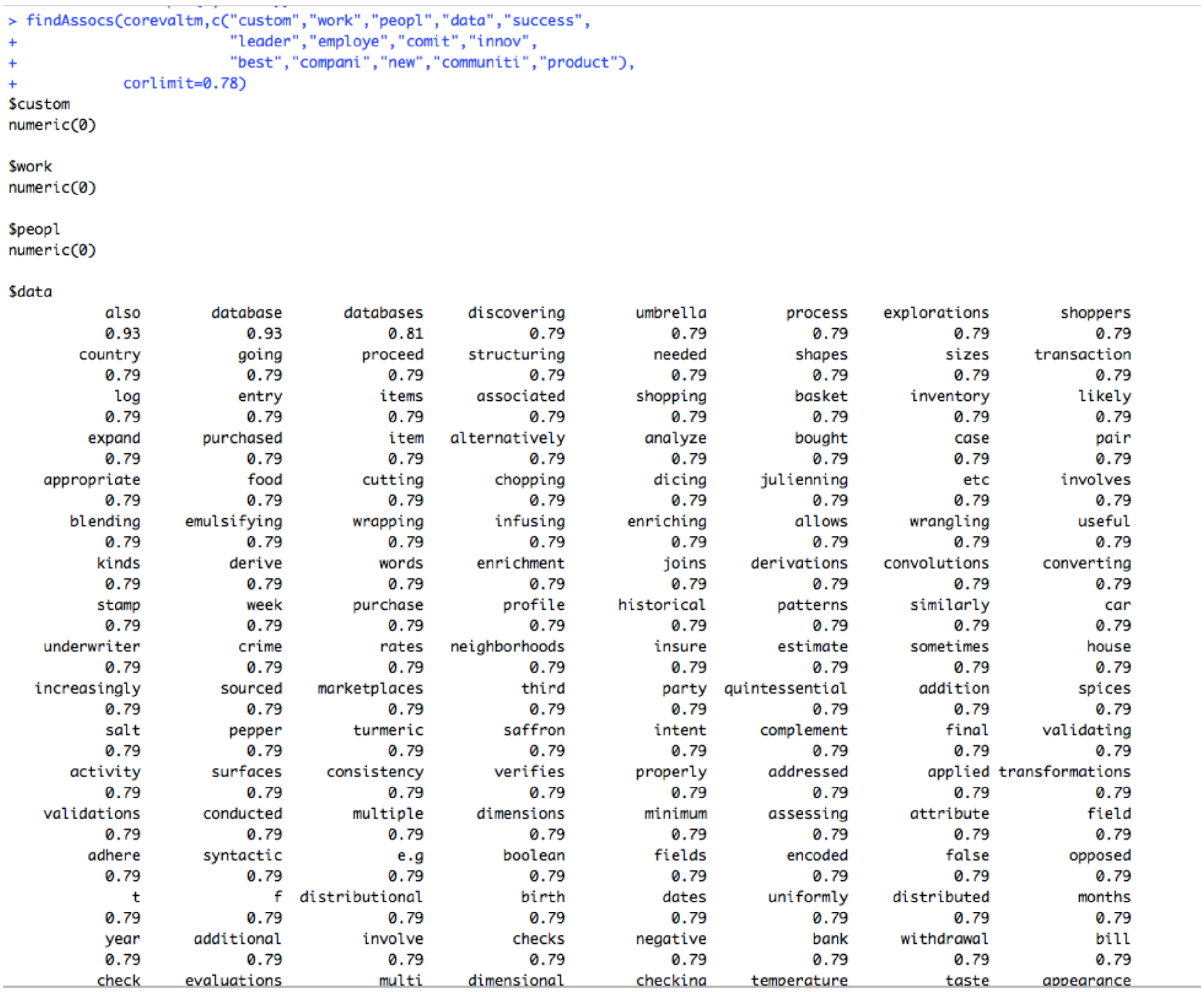
```
$enterpris
numeric(0)

$technolog
numeric(0)

$inform
numeric(0)

$analyt
     great      asset differenti       view       abil    whether     clinic      trial        new       drug      yield 
      0.83       0.83       0.83       0.83       0.83       0.83       0.83       0.83       0.83       0.83       0.83 
    farmer       soil     condit     carbon      emiss   communic    unlimit 
      0.83       0.83       0.83       0.83       0.83       0.83       0.83 

$improv
     great      asset differenti       view       abil    whether     clinic      trial        new       drug      yield 
      0.92       0.92       0.92       0.92       0.92       0.92       0.92       0.92       0.92       0.92       0.92 
    farmer       soil     condit     carbon      emiss   communic    unlimit    potenti    unleash        one      reduc 
      0.92       0.92       0.92       0.92       0.92       0.92       0.92       0.81       0.81       0.80       0.80 

$organ
numeric(0)
```

For core values

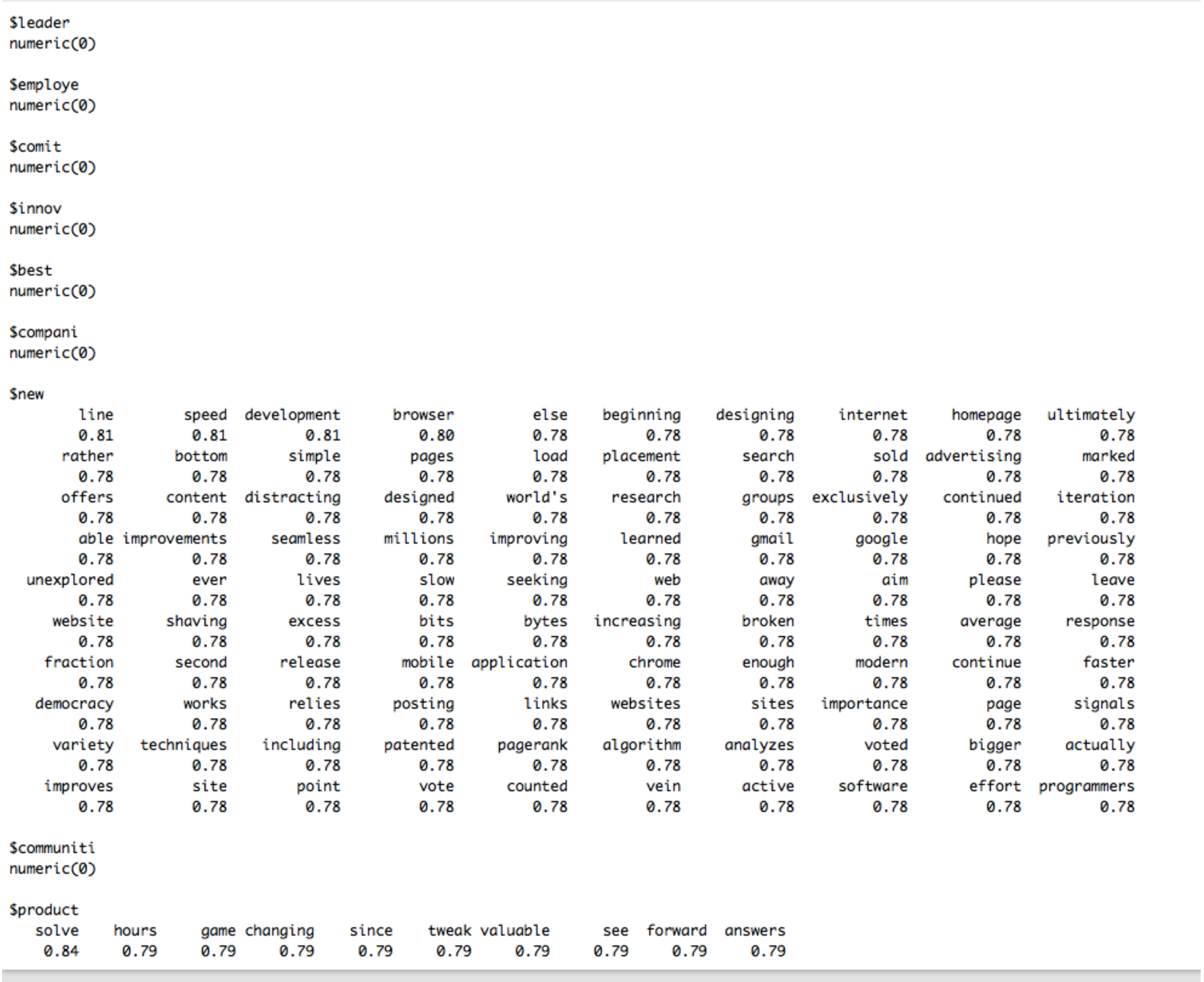


```
> corevaltopnew
    custom      work     peopl      data   success    leader   employe    commit     innov      best   compani       new communiti
        53        42        32        29        27        23        22        21        21        20        20        19        17
     thing     learn   product     think      busi     oracl    client   passion     alway     focus      help      time       way
        16        15        15        15        15        15        15        14        14        14        14        14        14
     build   problem    better      open
        13        13        13        13
> |
```





```
> findAssocs(corevaltm,c("custom","work","peopl","data","success",
+                        "leader","employe","comit","innov",
+                        "best","compani","new","communiti","product"),
+            corlimit=0.78)
$custom
numeric(0)

$work
numeric(0)

$peopl
numeric(0)

$data
           also        database       databases     discovering        umbrella         process    explorations        shoppers
           0.93            0.93            0.81            0.79            0.79            0.79            0.79            0.79
        country           going         proceed     structuring          needed          shapes           sizes     transaction
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
            log           entry           items      associated        shopping          basket       inventory          likely
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
         expand       purchased            item   alternatively         analyze          bought            case            pair
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
    appropriate            food         cutting        chopping          dicing      julienning             etc        involves
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
       blending     emulsifying        wrapping        infusing       enriching          allows       wrangling          useful
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
          kinds          derive           words      enrichment           joins     derivations    convolutions      converting
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
          stamp            week        purchase         profile      historical        patterns       similarly             car
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
    underwriter           crime           rates   neighborhoods          insure        estimate       sometimes           house
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
   increasingly         sourced    marketplaces           third           party  quintessential        addition          spices
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
           salt          pepper        turmeric         saffron          intent      complement           final      validating
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
       activity        surfaces     consistency        verifies        properly       addressed         applied transformations
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
    validations       conducted        multiple      dimensions         minimum       assessing       attribute           field
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
         adhere       syntactic             e.g         boolean          fields         encoded           false         opposed
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
              t               f  distributional           birth           dates       uniformly     distributed          months
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
           year      additional         involve          checks        negative            bank      withdrawal            bill
           0.79            0.79            0.79            0.79            0.79            0.79            0.79            0.79
          check     evaluations           multi     dimensional        checking     temperature           taste      appearance
```



```
$leader
numeric(0)

$employe
numeric(0)

$comit
numeric(0)

$innov
numeric(0)

$best
numeric(0)

$compani
numeric(0)

$new
        line        speed  development      browser         else    beginning    designing     internet     homepage   ultimately
        0.81         0.81         0.81         0.80         0.78         0.78         0.78         0.78         0.78         0.78
      rather       bottom       simple        pages         load    placement       search         sold  advertising       marked
        0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78
      offers      content  distracting     designed      world's     research       groups  exclusively    continued    iteration
        0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78
        able improvements     seamless     millions    improving      learned        gmail       google         hope   previously
        0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78
  unexplored         ever        lives         slow      seeking          web         away          aim       please        leave
        0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78
     website      shaving       excess         bits        bytes   increasing       broken        times      average     response
        0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78
    fraction       second      release       mobile  application       chrome       enough       modern     continue       faster
        0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78
   democracy        works       relies      posting        links     websites        sites   importance         page      signals
        0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78
     variety   techniques    including     patented     pagerank    algorithm     analyzes        voted       bigger     actually
        0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78
    improves         site        point         vote      counted         vein       active     software       effort  programmers
        0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78         0.78

$communiti
numeric(0)

$product
   solve    hours     game changing    since    tweak valuable      see  forward  answers
    0.84     0.79     0.79     0.79     0.79     0.79     0.79     0.79     0.79     0.79
```

- Here, doing analysis on the mission statements, we can say that after stemming the words, root words used in mission statement like “custom”,” work”,”analyt”,”imporv” and root words used in core values like “data”, ”new”, ”product” should be used as these words have highest correlation with the associated paragraphs.
- Hence we can say that, these words are form the characteristics features of a firm that adapts well to business analystics.

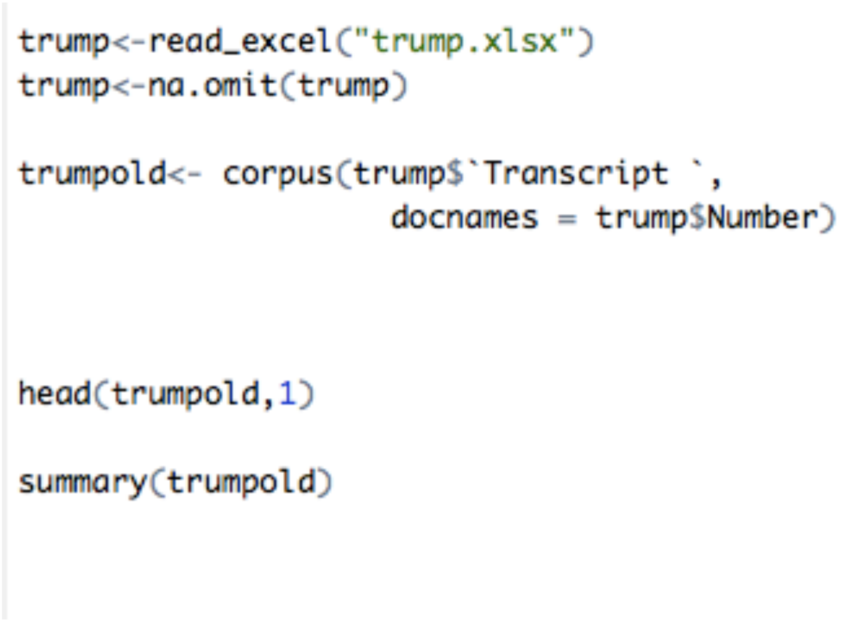
Q.5) Are there any other data-driven approaches you would recommend the CEO to implement?

A. Based on the review of the customers or the employee itself from the glassdoor or google, we can get an idea of what are the pros and cons of a business analytics firm which can drive our decision to form the characteristic features.

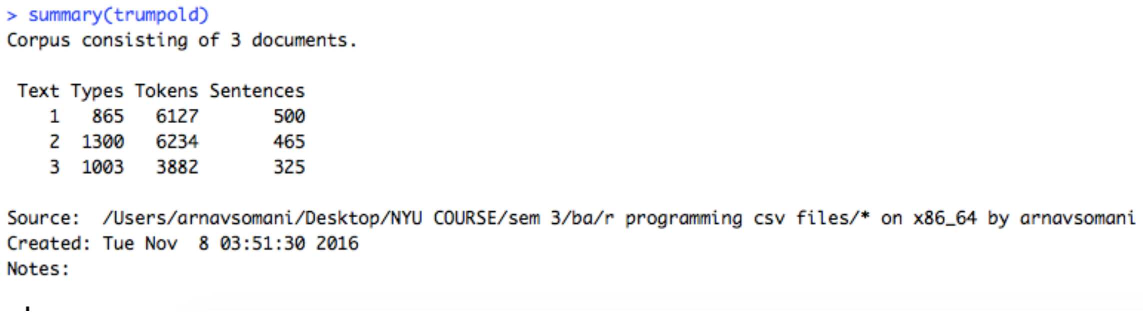
PART II

Q.1) Search for “Donald Trump speech transcript” and select 3 speeches of your choice
A. The transcript of Donald Trump speech has been copied in a excel file which I have attached with the assignment.

Q.2) Create a corpus for the speeches
A.

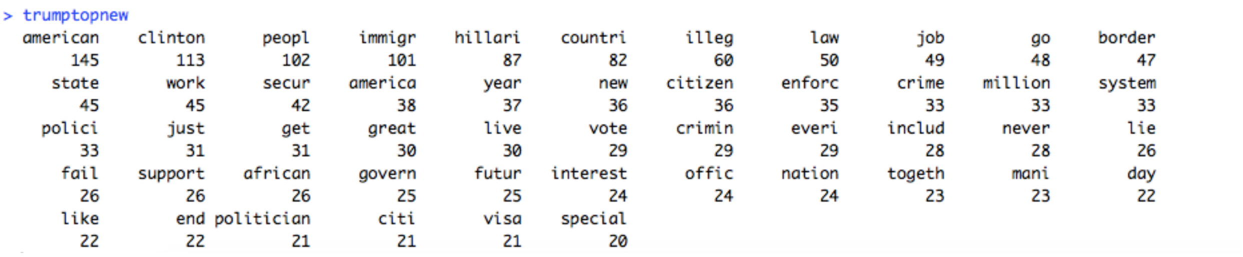


```
trump<-read_excel("trump.xlsx")
trump<-na.omit(trump)

trumpold<- corpus(trump$`Transcript `,
                  docnames = trump$Number)



head(trumpold,1)

summary(trumpold)
```

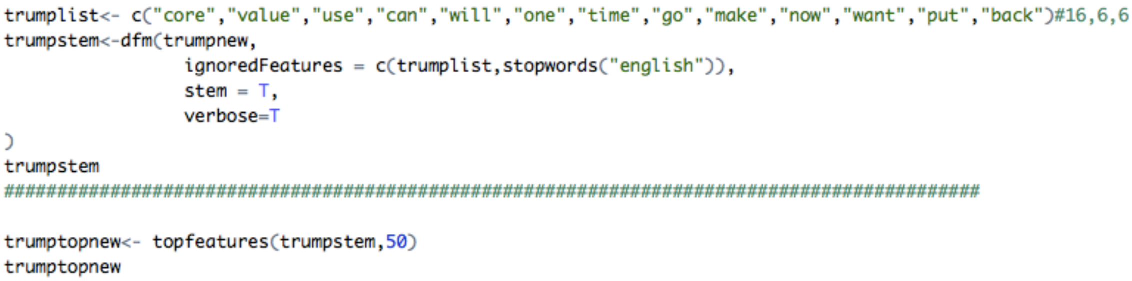


```
> summary(trumpold)
Corpus consisting of 3 documents.

 Text Types Tokens Sentences
    1   865   6127       500
    2  1300   6234       465
    3  1003   3882       325

Source:  /Users/arnavsomani/Desktop/NYU COURSE/sem 3/ba/r programming csv files/* on x86_64 by arnavsomani
Created: Tue Nov  8 03:51:30 2016
Notes:
  .
```

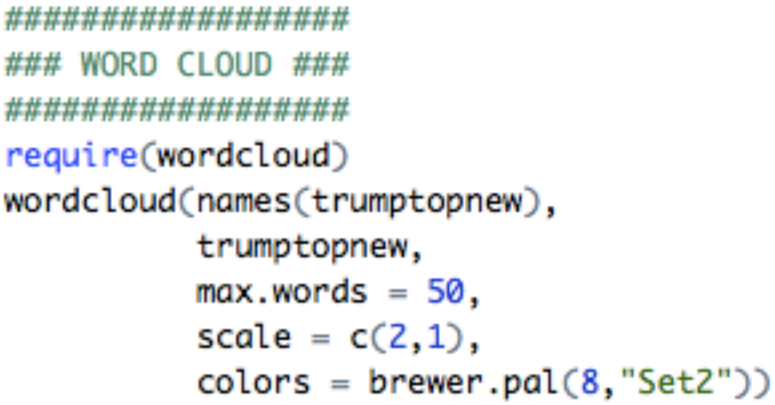
Q.3) Complete a frequency analysis of word usage
A.



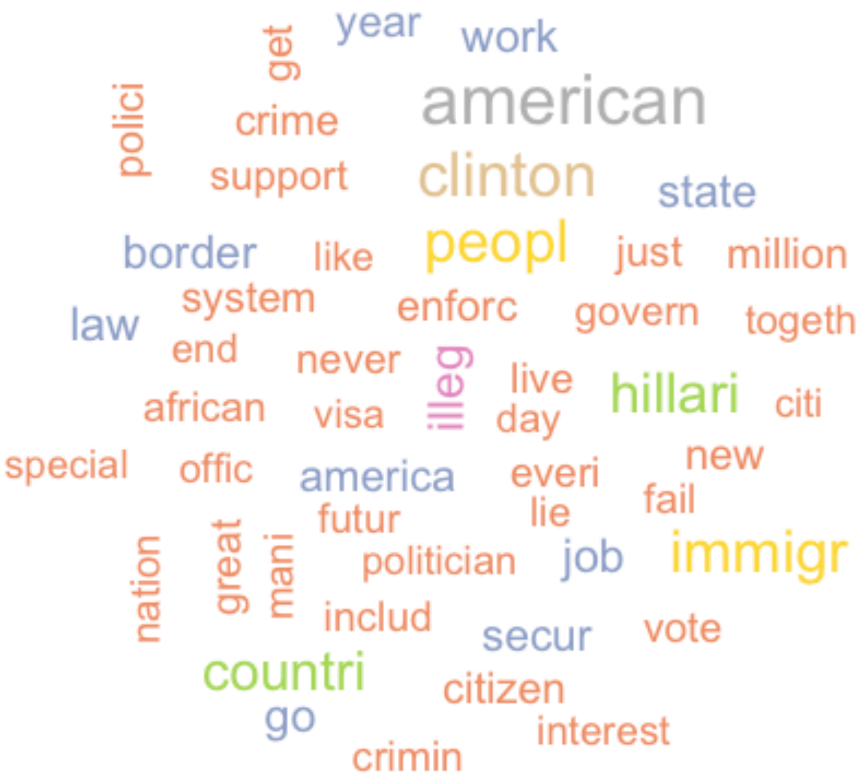
```
> trumptopnew
  american    clinton      peopl     immigr    hillari    countri      illeg        law        job         go     border
       145        113        102        101         87         82         60         50         49         48         47
     state       work      secur    america       year        new    citizen     enforc      crime    million     system
        45         45         42         38         37         36         36         35         33         33         33
    polici       just        get      great       live       vote     crimin      everi     includ      never        lie
        33         31         31         30         30         29         29         29         28         28         26
      fail    support    african     govern      futur   interest      offic     nation     togeth       mani        day
        26         26         26         25         25         24         24         24         23         23         22
      like        end politician       citi       visa    special
        22         22         21         21         21         20
```



```
trumplist<- c("core","value","use","can","will","one","time","go","make","now","want","put","back")#16,6,6
trumpstem<-dfm(trumpnew,
                 ignoredFeatures = c(trumplist,stopwords("english")),
                 stem = T,
                 verbose=T
)
trumpstem
############################################################################################

trumptopnew<- topfeatures(trumpstem,50)
trumptopnew
```

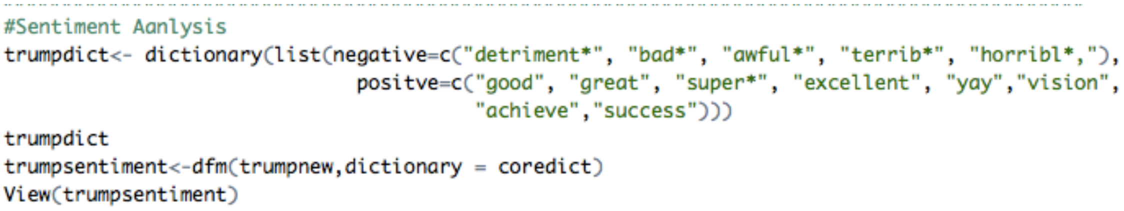


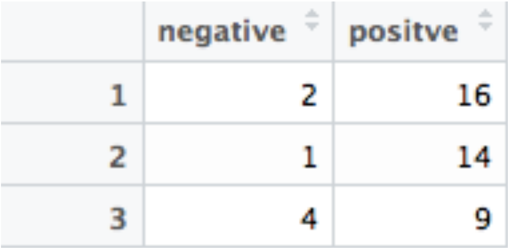
```
##################
### WORD CLOUD ###
##################
require(wordcloud)
wordcloud(names(trumptopnew),
          trumptopnew,
          max.words = 50,
          scale = c(2,1),
          colors = brewer.pal(8,"Set2"))
```



Q.4) Complete a sentiment analysis

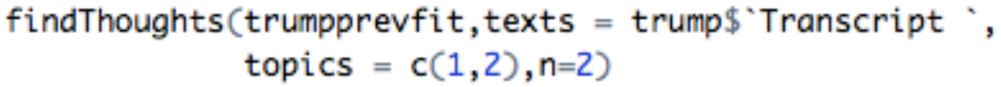
A.


```
#Sentiment Aanlysis
trumpdict<- dictionary(list(negative=c("detriment*", "bad*", "awful*", "terrib*", "horribl*,"),
                            positve=c("good", "great", "super*", "excellent", "yay","vision",
                                      "achieve","success")))
trumpdict
trumpsentiment<-dfm(trumpnew,dictionary = coredict)
View(trumpsentiment)
```

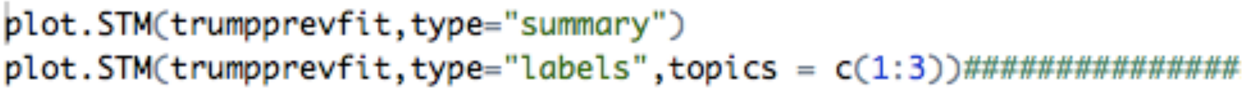


	negative	positve
1	2	16
2	1	14
3	4	9

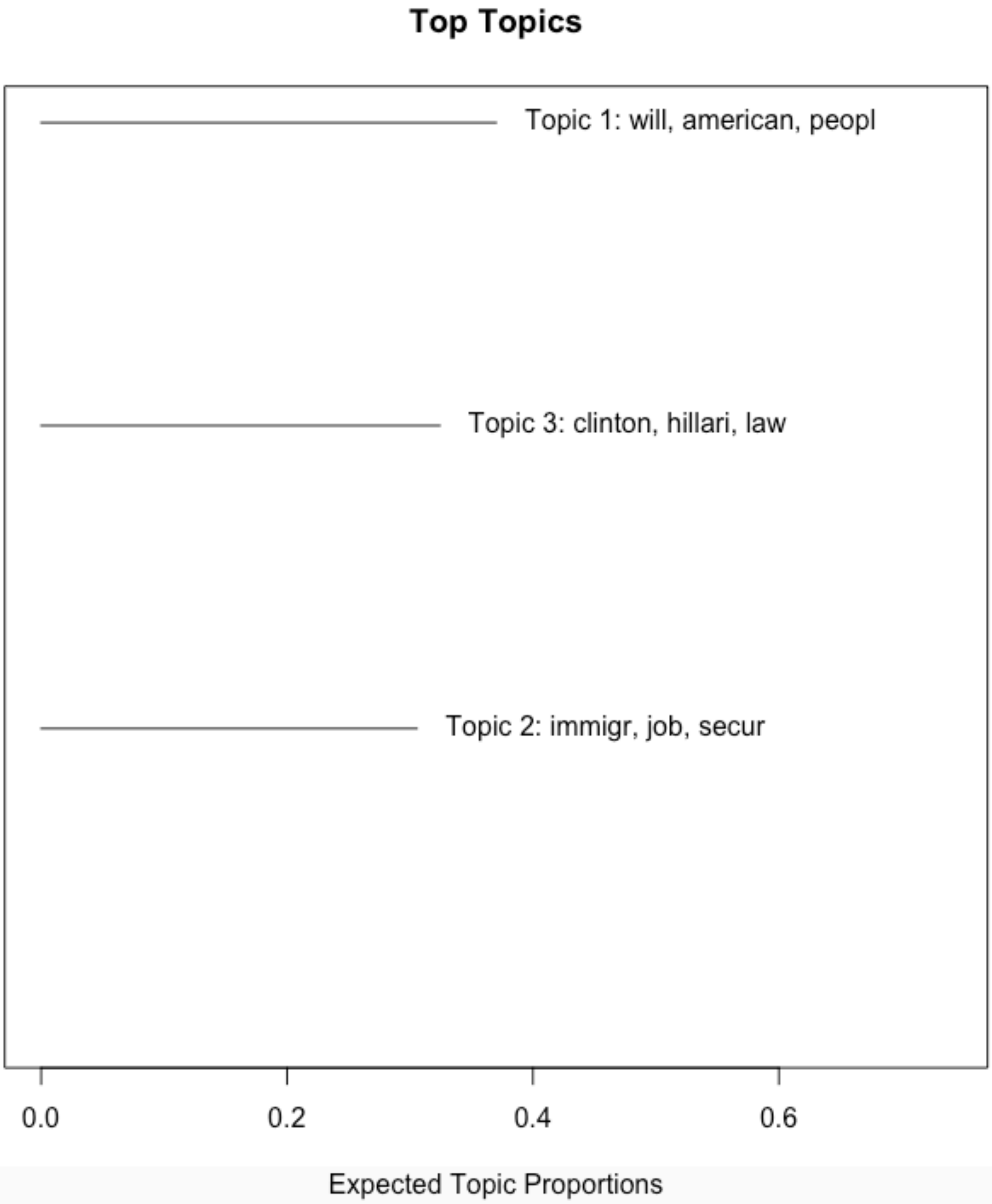
Q.5) What are the common topics in the corpus
A.

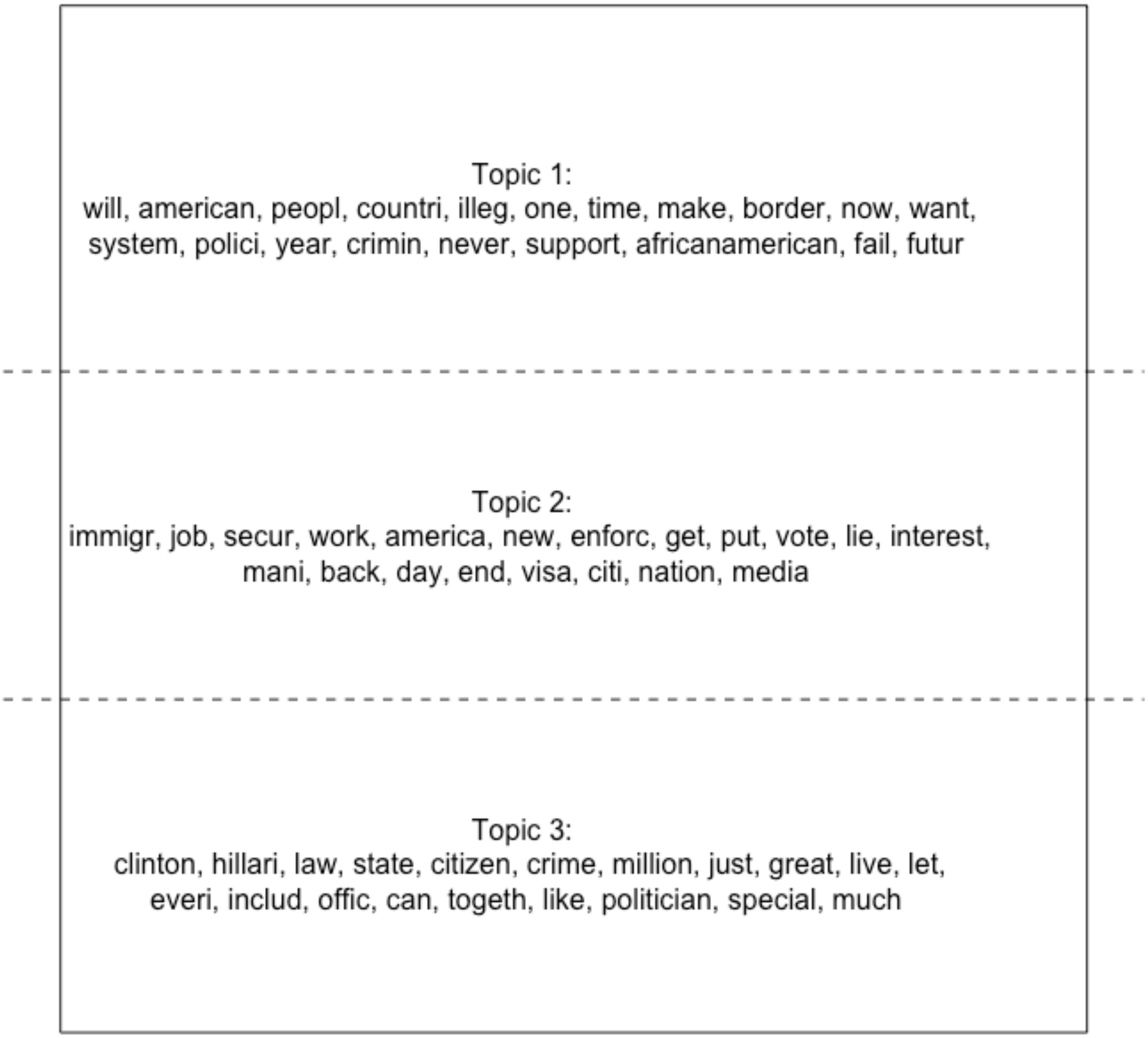


```
findThoughts(trumpprevfit,texts = trump$`Transcript `,
             topics = c(1,2),n=2)
```



```
plot.STM(trumpprevfit,type="summary")
plot.STM(trumpprevfit,type="labels",topics = c(1:3))###############
```


Top Topics
Topic 1: will, american, peopl
Topic 3: clinton, hillari, law
Topic 2: immigr, job, secur
0.0
0.2
0.4
0.6
Expected Topic Proportions



Topic 1:
will, american, peopl, countri, illeg, one, time, make, border, now, want, system, polici, year, crimin, never, support, africanamerican, fail, futur

Topic 2:
immigr, job, secur, work, america, new, enforc, get, put, vote, lie, interest, mani, back, day, end, visa, citi, nation, media

Topic 3:
clinton, hillari, law, state, citizen, crime, million, just, great, live, let, everi, includ, offic, can, togeth, like, politician, special, much

Q.6) Write a memo style report summarizing Trump's linguistic effectiveness.

A. There were some solid substantial issues this election, but somehow Trump always managed to find some smaller details to it. Similar to countless aspects of Trump's candidacy, voters also hold divided opinion of his candidacy. Here, I took three transcript of trump's speeches and analyzed it. What I found out through my analysis is, Trump chooses any one topic in any particular speech and talks about that in whole speech. Like in Topic one he chooses to speak on making wall on the border of America, in Topic 2 he chooses immigration and in topic 3 he choose to speak on his competitor Hillary Clinton. In spite of his negative actions over the years, trump has a vivid eyes on using more positive words in his speech. For every 1 negative word he uses, he balances it with using 3 positive words.