SOTL-first pass

Ebner

May 27, 2016

first try to do topic modeling with abstracts from several journals. 24 journals, 5625 non empty abstracts. there are many parameters here to change the analysis.

Topic modeling requires that 'chunks' of text are input. in this ananlysis, each abstract is a 'chunk'. in the Text Analysis book, Jockers uses between 500 and 1000 words per chunk.

Number of topics is another factor.

List of stop words another.

part of speech another.

```
library(mallet)
```

```
## Warning: package 'mallet' was built under R version 3.1.3

## Loading required package: rJava
```

```
library(wordcloud)
```

```
## Warning: package 'wordcloud' was built under R version 3.1.3

## Loading required package: RColorBrewer

## Warning: package 'RColorBrewer' was built under R version 3.1.3
```

```
library(reshape2)
library("Matrix")
#library(SnowballC) stemmer was kind of shitty
#library(hunspell) stemmer better and worse, ergo, still shitty
library(tm)
```

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

```
NUMTOPICS <- 10
```

The input file here is "Combined clean 5-12-16.csv"

```
# read in the whole csv file for processing
c <- read.csv("Combined clean  5-12-16.csv", stringsAsFactors = FALSE)
#dim(c)
# [1] 6775   27

# remove rows where abstract is NA
c <- c[which(c$abstract != ""),]
```

```
#dim(c)
# [1] 5625   27
s <- c
#s <- c[1:5,]
# split each abstract into words
ids.v <- paste(s$journal,s$year,s$Volume.number,s$Issue,s$page,sep="_")
ids.v <- gsub(" ","",ids.v)
words <- s$abstract
words.lower <- tolower(words)
words.nopunct <- gsub("[^[:alnum:][:space:]]", " ", words.lower)
words.nopunct <- words.nopunct[words.nopunct != ""]

# Split first to stem (this is done here becuase it also takes 's out)

# Make corpus dictionary from all words in all abstracts (after stop removal)
#using tm
words.stemmed <- NULL
# create the total corpus dictionary
dict <- unlist(strsplit(words.nopunct, "\\s+"))
dict <- removeWords(dict, stopwords("english"))
dict <- dict[dict != ""]
for (i in 1:length(words.nopunct)) {
    cat(".")
    corp <- unlist(strsplit(words.nopunct[[i]],"\\s+"))
    corp <- removeWords(corp, stopwords("english"))
    corp <- corp[corp != ""]

    stemmedCorpus <- stemDocument(corp)
    #stemCompletedCorpus <- stemCompletion(stemmedCorpus, dictionary = dict) #too slow
    stemCompletedCorpus <- stemmedCorpus
    stemCompletedCorpus <- stripWhitespace(stemCompletedCorpus)
    # collapse into single string
    stemCompletedCorpus <- paste(stemCompletedCorpus,sep="",collapse=" ")
    words.stemmed <- c(words.stemmed,stemCompletedCorpus)
}
```

```
## ...................................................................................................
```

```
# get rid of punctuation
words.l <- strsplit(words.stemmed, "\\s+")
# make into list of chr's, was list of array of chr's, where each was one word
chunks.v <- unlist(lapply(words.l, paste, collapse = " "))
#str(ids.v)
#str(chunks.v)
#summary(nchar(chunks.v))
```

summary info for the abstract lengths are:

```
summary(nchar(chunks.v))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.0   403.0   514.0   537.2   650.0  2776.0
```

Create the model for training and topic modeling

```r
mallet.instances <- mallet.import(ids.v, chunks.v,
                                  "stoplist.csv",
                                  FALSE,
                                  token.regexp = "[\\p{L}]+")
topic.model <- MalletLDA(num.topics = NUMTOPICS)
topic.model$loadDocuments(mallet.instances)

vocabulary <- topic.model$getVocabulary()
length(vocabulary)
```

```
## [1] 9812
```

```r
word.freqs <- mallet.word.freqs(topic.model = topic.model)
# top words in entire set
word.freqs <- word.freqs[ order(-word.freqs[,3], -word.freqs[,2]), ]
head(word.freqs,100)
```

```
##           words term.freq doc.freq
## 5       student     10890     3769
## 76         educ      4383     2112
## 121       learn      4410     1855
## 7         studi      3217     1850
## 21        cours      3538     1770
## 150      provid      2278     1711
## 238       manag      3425     1701
## 56       result      1828     1434
## 409       teach      2434     1427
## 84       articl      1859     1417
## 6          busi      2862     1405
## 25     research      2448     1397
## 142      author      1886     1299
## 16      account      4293     1287
## 103     discuss      1614     1249
## 36         base      1668     1228
## 495      effect      1636     1204
## 248     present      1504     1202
## 67        paper      1678     1182
## 358      experi      1793     1172
## 195     develop      1730     1149
## 335      examin      1229     1023
## 132      differ      1461     1013
## 11      univers      1394      986
## 210     suggest      1131      978
## 26       import      1237      974
## 257       relat      1237      955
## 57         find      1188      943
## 97         case      2225      941
## 585     program      2058      927
## 89      describ      1092      925
## 382        work      1316      904
## 20   undergradu      1119      900
## 254      design      1195      885
## 245  understand      1081      879
## 190        issu      1152      873
## 50     approach      1405      870
## 327      includ      1026      866
## 59        model      1625      859
## 336     practic      1115      835
## 369       skill      1483      826
## 365     process      1221      824
## 152     increas      1056      816
## 35        level      1099      815
## 1255     econom      1788      789
## 351    signific      1013      785
## 321       class      1218      784
## 270   classroom      1055      769
## 63       assess      1264      768
## 156      requir      1041      761
## 2          year       994      742
## 114       focus       844      721
## 53      analysi       958      720
## 203        data       995      718
## 13      perform      1359      698
## 353       offer       842      698
## 375       group      1191      685
## 40       inform      1017      676
## 713        make       831      663
## 540      school      1073      662
## 352    identifi       794      659
## 310  instructor      1022      657
## 418       indic       737      652
## 88      concept       936      647
## 3897     hospit      1522      642
## 180      improv       807      640
## 532      academ      1000      636
## 275       activ       958      634
## 115      specif       702      628
## 73       report       917      625
## 295         set       788      625
## 487      method       920      623
## 433       evalu       852      615
## 306        time       805      615
## 324        mani       696      613
## 470       howev       640      603
## 635       chang       873      594
## 290    knowledg       853      592
## 289    abstract       666      592
## 189      survey       751      588
## 200      integr       792      579
## 601      explor       653      579
## 798       decis       906      570
## 127     success       693      570
## 340     respons       703      567
## 431    opportun       664      558
## 405     problem       927      557
## 393      implic       592      557
## 314     graduat       866      551
## 432     current       659      550
## 104      higher       721      548
## 125      theori       790      544
## 730    challeng       685      535
## 240    interest       645      534
## 81        addit       566      524
## 236        role       632      521
## 176     financi      1044      520
## 80       consid       593      520
## 149       posit       657      519
## 388    particip       755      516
```

Now Train the Model

```
topic.model$train(400)

topic.words.m <- mallet.topic.words(topic.model,
                                    smoothed = TRUE,
                                    normalized = TRUE)
topic.words.notNormed.m <- mallet.topic.words(topic.model,
                                    smoothed = TRUE,
                                    normalized = FALSE)
topic.words.notNormed.m <- round(topic.words.notNormed.m)

dim(topic.words.m)
```

```
## [1]   10 9812
```

```
colnames(topic.words.m) <- vocabulary
colnames(topic.words.notNormed.m) <- vocabulary
#topic.words.m[,1:6]
# topic.words.notNormed.df is a data frame of
# 10 rows (topics) by 15963 columns (word 'types)
# the entries of row,col are the frequency of word in a given topic
topic.words.notNormed.df <- as.data.frame(topic.words.notNormed.m)

# top 3 words in each topic
mallet.topic.labels(topic.model,topic.words.m)
```

```
##  [1] "student learn exercis"       "busi manag educ"
##  [3] "educ entrepreneurship learn" "case financi account"
##  [5] "student studi learn"         "hospit program studi"
##  [7] "account ethic educ"          "student econom cours"
##  [9] "model game author"           "teach student learn"
```

Show an array of plots that roughly show the words that are in each topic

```
par( mfcol=c(2,5) )
#Sort rows by topic value.
topic.words.notNormed.sorted.df <-
```

```
    topic.words.notNormed.df[, order(topic.words.notNormed.df[1,], topic.words.notNormed.df[2,], topic.w
topic.words.notNormed.sorted.m <- as.matrix(topic.words.notNormed.sorted.df)
for (i in 1:NUMTOPICS) {
    nonzeros <- sum(topic.words.notNormed.sorted.df[i,]!=0)
    print(paste(nonzeros,"Unique words in topic",i))
    title <- paste("Topic",i, "word dist")
    barplot(topic.words.notNormed.sorted.m[i,],horiz = TRUE)
    title(main=title)
}
```

```
## [1] "1479 Unique words in topic 1"
```

```
## [1] "1652 Unique words in topic 2"
```

```
## [1] "1488 Unique words in topic 3"
```

```
## [1] "2057 Unique words in topic 4"
```

```
## [1] "1489 Unique words in topic 5"
```

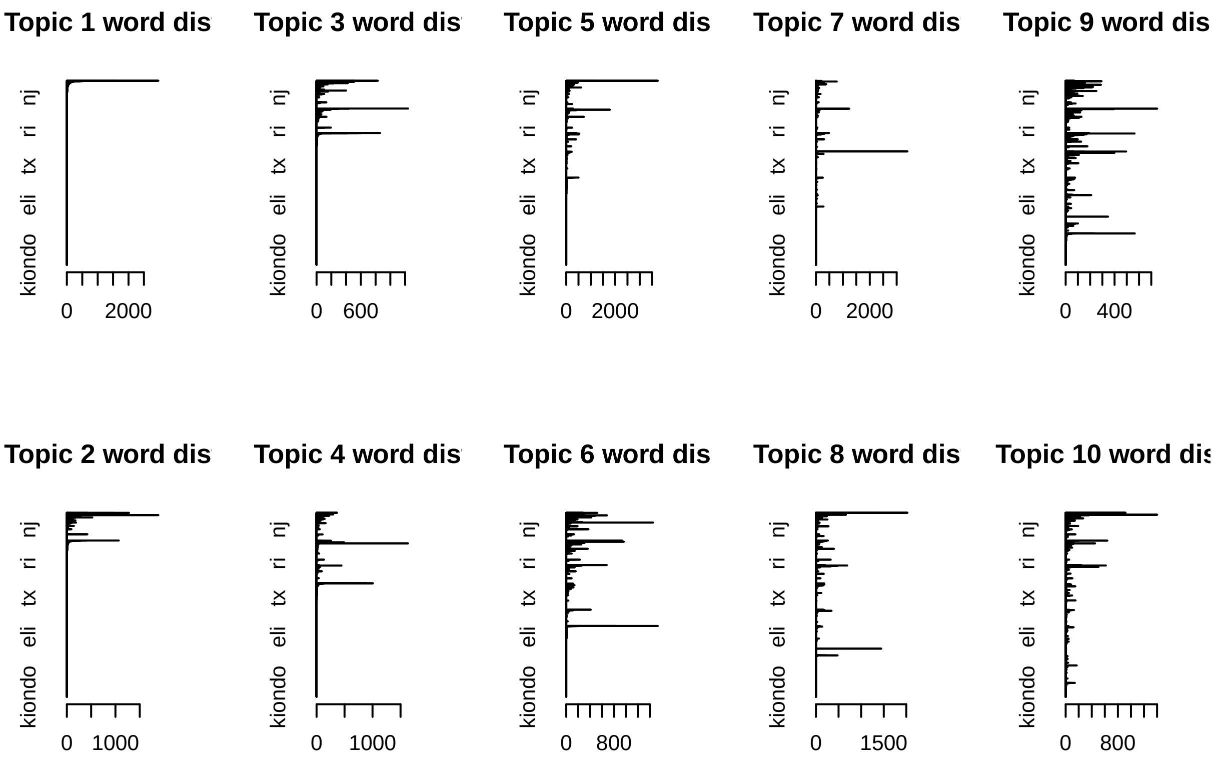
```
## [1] "1464 Unique words in topic 6"
```

```
## [1] "1729 Unique words in topic 7"
```

```
## [1] "1362 Unique words in topic 8"
```

```
## [1] "2054 Unique words in topic 9"
```

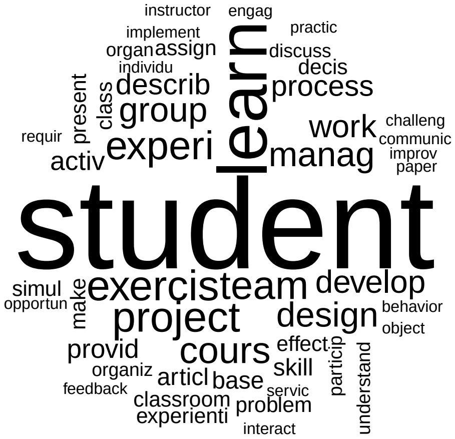
```
## [1] "1808 Unique words in topic 10"
```



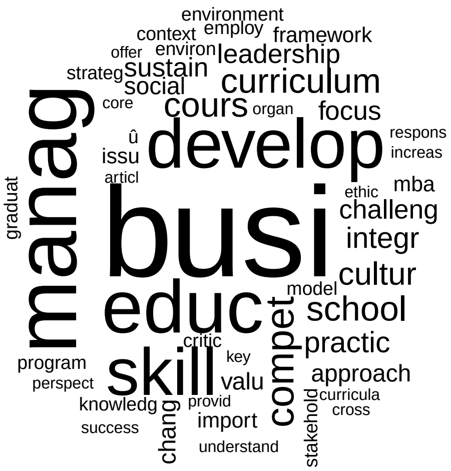
Create a word cloud for each topic found in the text. also print out word frequencies

```
for (i in 1:NUMTOPICS) {
    topic.top.words <- mallet.top.words(topic.model,
                                        topic.words.m[i,], 50)
    topic.top.words.notNormed <- mallet.top.words(topic.model,
                                        topic.words.notNormed.m[i,], 50)
    topic.top.words.notNormed$weights <-
        as.integer(topic.top.words.notNormed$weights)
    print(paste("TOPIC ",i))
    print(topic.top.words.notNormed)
    wordcloud(topic.top.words$words,
          topic.top.words$weights,random.order = F,
          c(4,0.2))
}
```

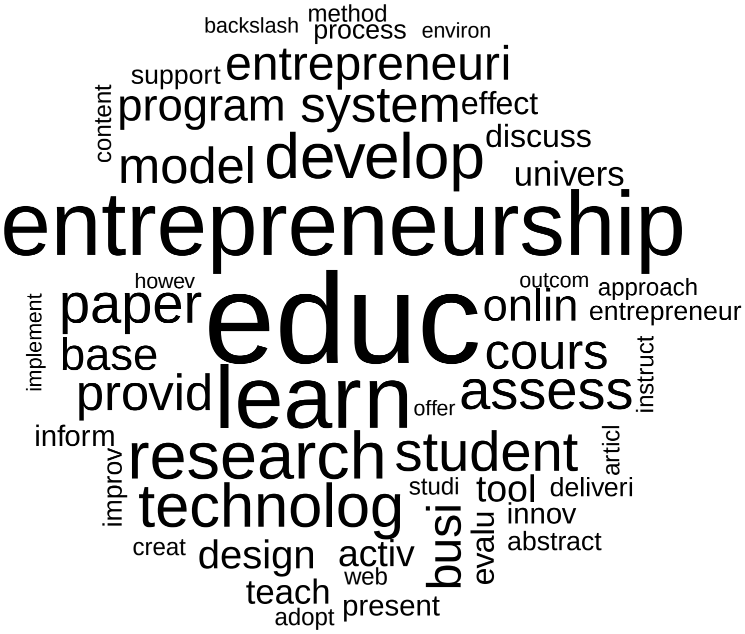
```
## [1] "TOPIC  1"
##                  words weights
## student        student    2971
## learn            learn    1552
## exercis        exercis     899
## project        project     865
## team              team     847
## experi          experi     807
## cours            cours     737
## manag            manag     712
## group            group     681
## design         design     673
## develop       develop     658
## work             work     593
## process       process     571
## describ       describ     559
## activ           activ     523
## provid         provid     484
## articl         articl     454
## base             base     441
## skill           skill     423
## class           class     394
## classroom   classroom     391
## make             make     387
## assign         assign     383
## effect         effect     374
## decis           decis     370
## present       present     343
## simul           simul     342
## problem       problem     338
## experienti experienti     334
## particip     particip     329
## organiz       organiz     320
## discuss       discuss     302
## organ           organ     285
## understand understand     279
## improv         improv     267
## servic         servic     263
## individu     individu     260
## behavior     behavior     254
## paper           paper     245
## challeng     challeng     245
## practic       practic     243
## implement   implement     243
## engag           engag     239
## opportun     opportun     231
## communic     communic     224
## feedback     feedback     222
## interact     interact     218
## requir         requir     215
## instructor instructor     213
## object         object     203
```



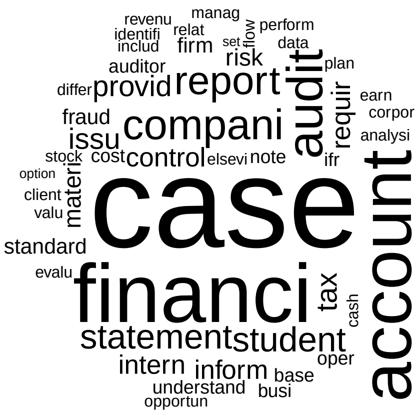
```
## [1] "TOPIC  2"
##                   words weights
## busi               busi    1885
## manag             manag    1281
## educ               educ    1070
## develop         develop     970
## skill             skill     937
## compet           compet     525
## cours             cours     435
## curriculum   curriculum     424
## school           school     423
## cultur           cultur     404
## practic         practic     349
## integr           integr     343
## leadership   leadership     324
## challeng       challeng     323
## focus             focus     306
## sustain         sustain     296
## social           social     285
## valu               valu     278
## chang             chang     268
## approach       approach     265
## issu               issu     247
## import           import     215
## mba                 mba     212
## framework     framework     211
## employ           employ     207
## environ         environ     205
## critic           critic     205
## model             model     199
## strateg         strateg     194
## program         program     190
## knowledg       knowledg     188
## context         context     185
## graduat         graduat     181
## stakehold     stakehold     179
## environment environment     177
## û                     û     174
## articl           articl     169
## core               core     165
## understand   understand     159
## organ             organ     158
## ethic             ethic     157
## increas         increas     156
## respons         respons     155
## provid           provid     153
## offer             offer     151
## perspect       perspect     146
## key                 key     145
## curricula     curricula     145
## success         success     144
## cross             cross     141
```



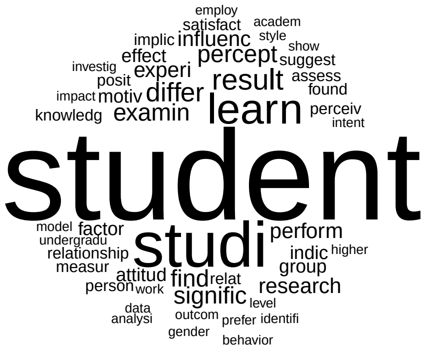
```
## [1] "TOPIC  3"
##                             words weights
## educ                         educ    1241
## entrepreneurship entrepreneurship     862
## learn                       learn     831
## research                 research     610
## develop                   develop     585
## technolog               technolog     559
## student                   student     516
## assess                     assess     512
## paper                       paper     503
## system                     system     465
## cours                       cours     463
## model                       model     442
## provid                     provid     439
## entrepreneuri       entrepreneuri     432
## busi                         busi     426
## onlin                       onlin     408
## program                   program     401
## base                         base     394
## design                     design     347
## activ                       activ     310
## tool                         tool     305
## univers                   univers     297
## teach                       teach     280
## discuss                   discuss     274
## effect                     effect     251
## evalu                       evalu     250
## present                   present     241
## inform                     inform     232
## innov                       innov     229
## process                   process     218
## support                   support     211
## deliveri                 deliveri     208
## improv                     improv     205
## abstract                 abstract     203
## entrepreneur         entrepreneur     195
## studi                       studi     193
## web                           web     193
## method                     method     183
## content                   content     179
## approach                 approach     178
## creat                       creat     178
## articl                     articl     177
## adopt                       adopt     171
## instruct                 instruct     166
## howev                       howev     157
## backslash               backslash     156
## implement               implement     155
## offer                       offer     151
## environ                   environ     150
## outcom                     outcom     149
```



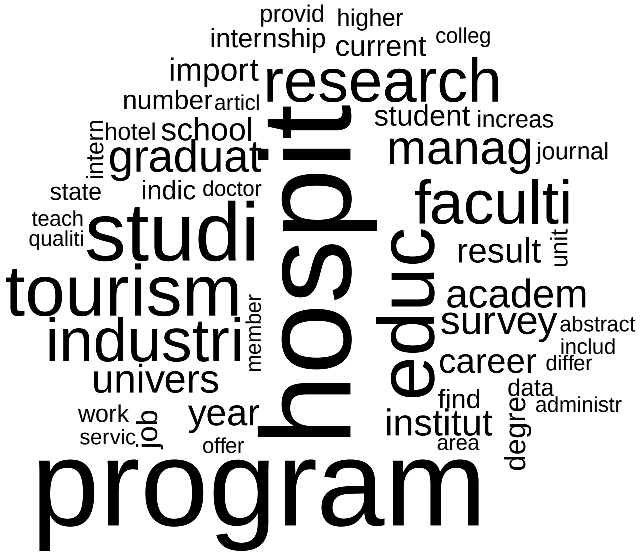
```
## [1] "TOPIC  4"
##                 words weights
## case             case    1636
## financi       financi    1009
## account       account     888
## audit           audit     586
## compani       compani     495
## report         report     446
## statement   statement     373
## student       student     368
## provid         provid     311
## issu             issu     299
## tax               tax     293
## control       control     269
## requir         requir     268
## intern         intern     259
## inform         inform     256
## risk             risk     228
## materi         materi     228
## firm             firm     216
## fraud           fraud     210
## standard     standard     200
## auditor       auditor     178
## ifr               ifr     173
## understand understand     169
## cost             cost     165
## note             note     164
## busi             busi     163
## base             base     152
## oper             oper     151
## analysi       analysi     147
## corpor         corpor     144
## flow             flow     143
## cash             cash     141
## elsevi         elsevi     139
## plan             plan     136
## data             data     135
## earn             earn     133
## manag           manag     130
## opportun     opportun     128
## relat           relat     127
## includ         includ     123
## valu             valu     123
## evalu           evalu     121
## stock           stock     120
## differ         differ     119
## client         client     119
## perform       perform     116
## revenu         revenu     114
## identifi     identifi     114
## option         option     110
## set               set     108
```



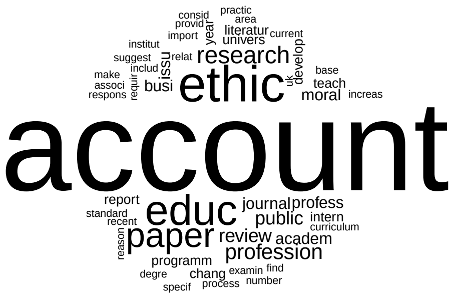
```
## [1] "TOPIC  5"
##                      words weights
## student           student    3740
## studi               studi    1784
## learn               learn    1121
## result             result     727
## differ             differ     613
## examin             examin     546
## find                 find     537
## signific         signific     525
## percept           percept     504
## perform           perform     472
## experi             experi     463
## research         research     431
## influenc         influenc     422
## factor             factor     406
## group               group     394
## attitud           attitud     387
## indic               indic     370
## effect             effect     362
## motiv               motiv     352
## relat               relat     312
## perceiv           perceiv     307
## assess             assess     303
## posit               posit     301
## relationship relationship     290
## knowledg         knowledg     287
## found               found     284
## person             person     273
## suggest           suggest     266
## satisfact       satisfact     260
## measur             measur     259
## implic             implic     259
## work                 work     254
## undergradu     undergradu     249
## style               style     246
## data                 data     245
## level               level     242
## impact             impact     242
## analysi           analysi     237
## investig         investig     227
## outcom             outcom     224
## gender             gender     224
## prefer             prefer     221
## behavior         behavior     220
## intent             intent     217
## show                 show     210
## higher             higher     209
## academ             academ     208
## identifi         identifi     199
## employ             employ     199
## model               model     198
```



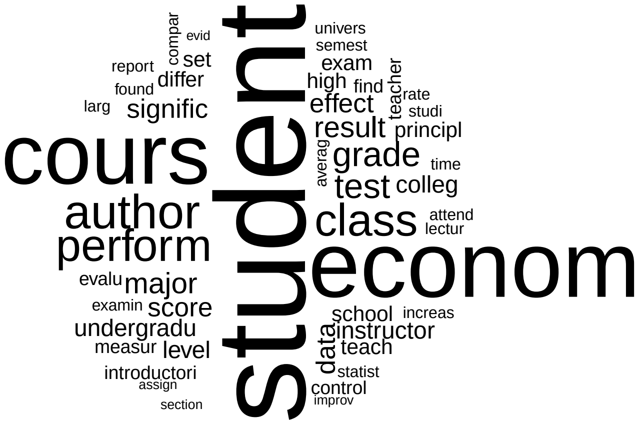
```
## [1] "TOPIC  6"
##                   words weights
## hospit         hospit    1534
## program       program    1454
## studi           studi     965
## educ             educ     935
## tourism       tourism     830
## industri     industri     697
## faculti       faculti     683
## research     research     679
## manag           manag     522
## graduat       graduat     487
## univers       univers     435
## survey         survey     421
## academ         academ     411
## year             year     393
## institut     institut     374
## career         career     373
## result         result     362
## school         school     312
## import         import     311
## degre           degre     306
## student       student     282
## job               job     279
## current       current     278
## internship internship     264
## find              find     257
## number          number     254
## indic            indic     243
## intern          intern     237
## data              data     229
## increas        increas     220
## hotel            hotel     219
## work              work     216
## higher          higher     215
## provid          provid     209
## unit              unit     208
## journal        journal     208
## state            state     207
## administr   administr     203
## servic          servic     200
## doctor          doctor     194
## abstract      abstract     192
## articl          articl     191
## differ          differ     190
## colleg          colleg     190
## qualiti        qualiti     190
## includ          includ     189
## area              area     188
## member          member     186
## offer            offer     181
## teach            teach     178
```



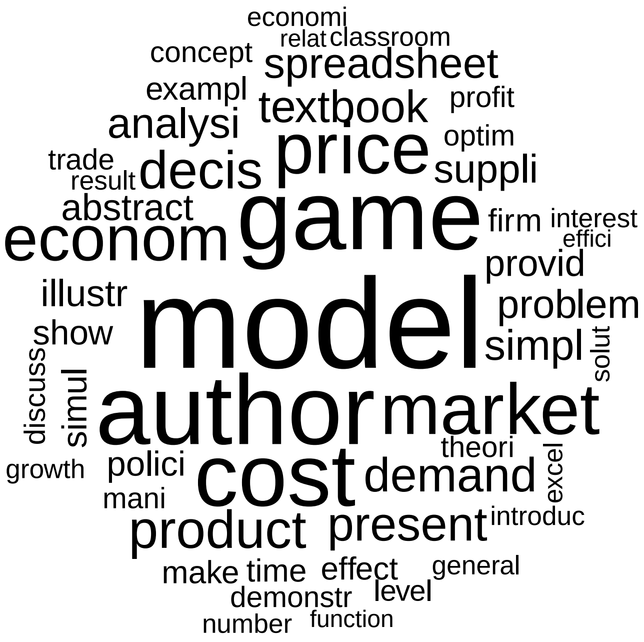
```
## [1] "TOPIC  7"
##                  words weights
## account       account    3405
## ethic           ethic    1242
## educ             educ    1038
## paper           paper     775
## research     research     490
## profession profession     387
## review         review     364
## public         public     308
## journal       journal     306
## issu             issu     302
## profess       profess     287
## moral           moral     284
## busi             busi     262
## academ         academ     260
## develop       develop     216
## univers       univers     205
## programm     programm     188
## report         report     181
## literatur   literatur     174
## teach           teach     164
## year             year     159
## chang           chang     158
## intern         intern     157
## import         import     152
## provid         provid     146
## examin         examin     146
## suggest       suggest     142
## requir         requir     135
## standard     standard     135
## associ         associ     134
## base             base     133
## respons       respons     131
## reason         reason     128
## process       process     124
## relat           relat     121
## uk                 uk     118
## institut     institut     117
## current       current     112
## area             area     110
## number         number     110
## make             make     110
## find             find     109
## recent         recent     107
## includ         includ     107
## practic       practic     107
## curriculum curriculum     106
## degre           degre     102
## consid         consid     102
## increas       increas      97
## specif         specif      96
```



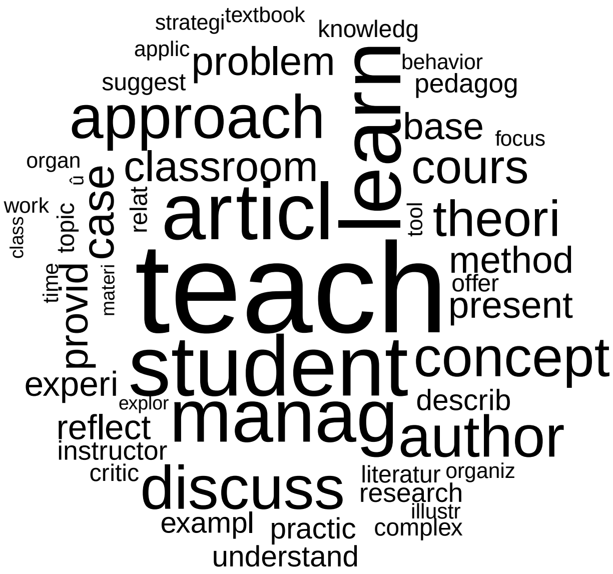
```
## [1] "TOPIC  8"
##                        words weights
## student              student    2027
## econom                econom    1446
## cours                  cours    1336
## author                author     694
## class                  class     666
## perform              perform     665
## test                    test     481
## grade                  grade     480
## result                result     401
## major                  major     387
## effect                effect     350
## score                  score     346
## signific            signific     334
## data                    data     327
## undergradu        undergradu     311
## level                  level     307
## instructor        instructor     294
## colleg                colleg     280
## school                school     265
## differ                differ     257
## teach                  teach     254
## high                    high     245
## exam                    exam     244
## principl            principl     233
## set                          set     231
## evalu                      evalu     220
## measur                    measur     215
## find                        find     215
## teacher                  teacher     190
## introductori introductori     188
## control                  control     186
## examin                    examin     179
## statist                  statist     174
## attend                    attend     172
## semest                    semest     170
## compar                    compar     168
## report                    report     167
## found                      found     163
## increas                  increas     162
## time                        time     162
## studi                      studi     153
## lectur                    lectur     146
## univers                  univers     141
## rate                        rate     141
## larg                        larg     141
## averag                    averag     140
## assign                    assign     135
## evid                        evid     134
## section                  section     132
## improv                    improv     131
```



```
## [1] "TOPIC  9"
##                   words weights
## model             model     749
## game               game     566
## author           author     564
## cost               cost     495
## price             price     399
## market           market     396
## econom           econom     347
## decis             decis     294
## product         product     289
## present         present     258
## demand           demand     252
## textbook       textbook     242
## spreadsheet spreadsheet     229
## analysi         analysi     225
## simpl             simpl     217
## problem         problem     215
## suppli           suppli     210
## illustr         illustr     198
## abstract       abstract     195
## provid           provid     193
## show               show     179
## polici           polici     173
## simul             simul     170
## time               time     158
## firm               firm     158
## effect           effect     155
## exampl           exampl     154
## make               make     153
## theori           theori     149
## optim             optim     148
## profit           profit     147
## mani               mani     146
## discuss         discuss     142
## concept         concept     141
## trade             trade     141
## demonstr       demonstr     139
## level             level     138
## result           result     130
## introduc       introduc     130
## general         general     130
## classroom     classroom     128
## excel             excel     127
## growth           growth     124
## solut             solut     123
## number           number     119
## interest       interest     118
## economi         economi     117
## function       function     113
## relat             relat     112
## effici           effici     112
```



```
## [1] "TOPIC  10"
##                 words weights
## teach           teach    1403
## student       student     918
## learn           learn     906
## articl         articl     840
## manag           manag     780
## approach     approach     640
## discuss       discuss     638
## author         author     617
## concept       concept     562
## theori         theori     505
## cours           cours     473
## case             case     451
## classroom   classroom     418
## provid         provid     381
## problem       problem     373
## present       present     359
## method         method     358
## base             base     348
## experi         experi     316
## reflect       reflect     311
## describ       describ     278
## exampl         exampl     270
## practic       practic     264
## understand understand     255
## research     research     242
## pedagog       pedagog     222
## instructor instructor     221
## suggest       suggest     218
## topic           topic     208
## knowledg     knowledg     207
## critic         critic     205
## relat           relat     200
## complex       complex     195
## literatur   literatur     193
## offer           offer     189
## focus           focus     182
## illustr       illustr     179
## behavior     behavior     179
## textbook     textbook     171
## applic         applic     168
## tool             tool     167
## organ           organ     167
## work             work     165
## organiz       organiz     160
## time             time     158
## strategi     strategi     157
## explor         explor     155
## class           class     154
## û                   û     154
## materi         materi     153
```

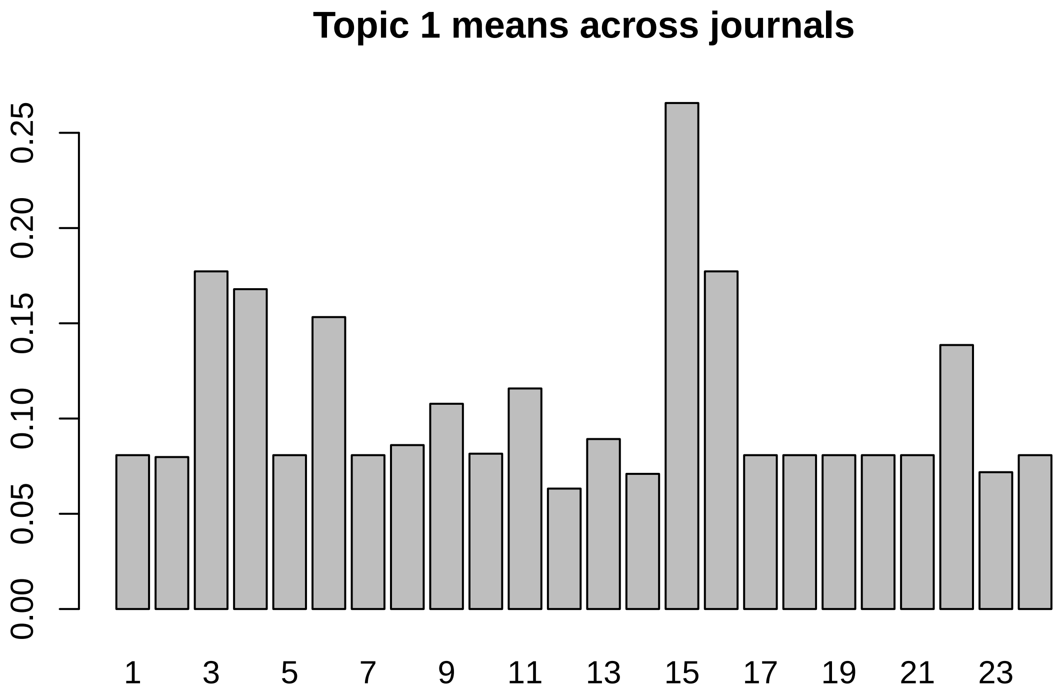


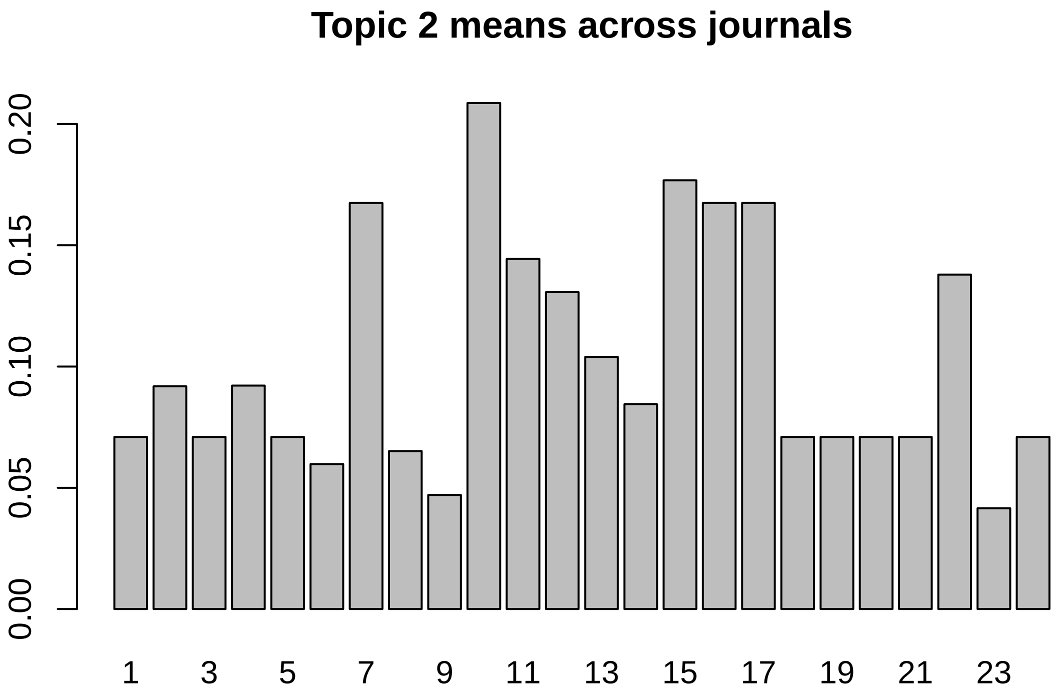
Topic Coherence shows the probability of a given 'topic' in a given abstract Rows are abstracts, and columns are topics. Since there are > 5000 abstracts we want to show this in a more digestible way by showing average topic probability based on journal name.(could be other, e.g. discipline)

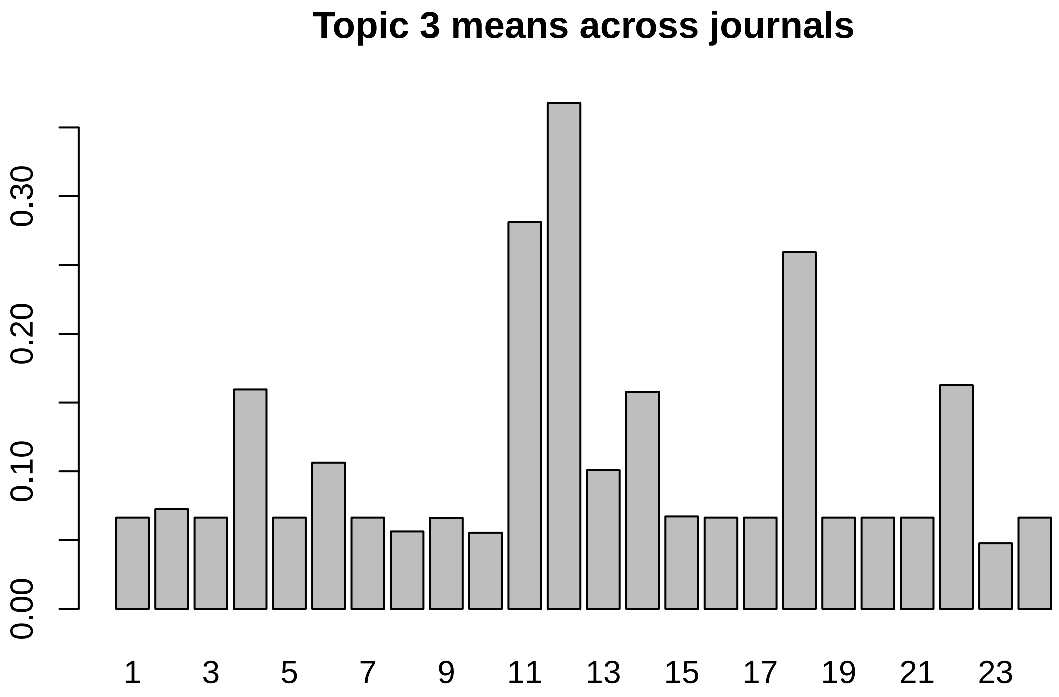
```r
#list journals:
u <- unique(s$journal)
print(u)
```

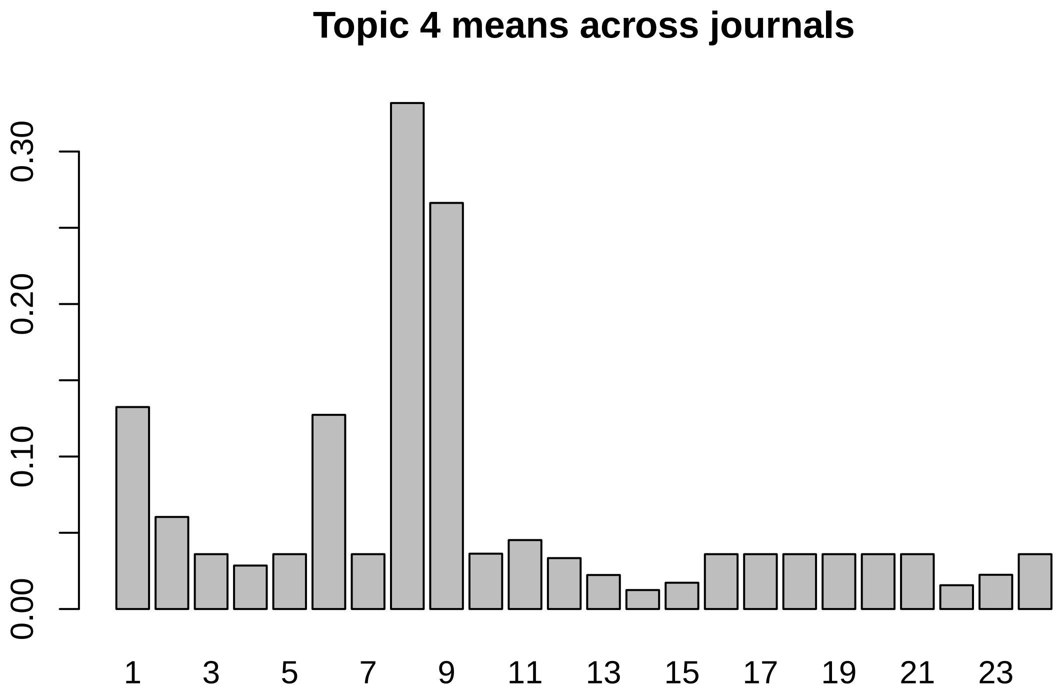
```
##  [1] "Accounting Education"
##  [2] "Issues in Accounting Education"
##  [3] "INFORMS: Transactions on Education"
##  [4] "Decision Sciences Journal of Innovative Education"
##  [5] "The Journal of Economic Education"
##  [6] "The BMW Model: A New Framework for Teaching Monetary Economics"
##  [7] "Teaching New Keynesian Open Economy Macroeconomics at the Intermediate Level"
##  [8] "Teaching Real Business Cycles to Undergraduates"
##  [9] "Intermediate Macroeconomics Tutorials and Applets"
## [10] "Teaching Economics to Undergraduates in Europe: Volume, Structure, and Contents"
## [11] "Exploring Duopoly Markets with Conjectural Variations"
## [12] "Cubic Cost Functions and Major Market Structures"
## [13] "The International Journal of Management Education"
## [14] "Journal of Management Education"
## [15] "Management Education for Practicing Managers: Combining Academic Rigor With Personal Change and
## [16] "More Than a Game: Learning About Climate Change Through Role-Play "
## [17] " Finland\""
## [18] "Journal of Hospitality, Leisure, Sport {\\&} Tourism Education"
## [19] "Journal of Hospitality {\\&} Tourism Education"
## [20] "The Portfolio‰ÛÓAn Alternative Assessment Method in Hospitality and Tourism Management Educati
## [21] "Journal of Entrepreneurship Education"
## [22] "journal of entrepreneurship education"
## [23] "Journal of Business Ethics Education"
## [24] "Journal of Accounting Education"
```

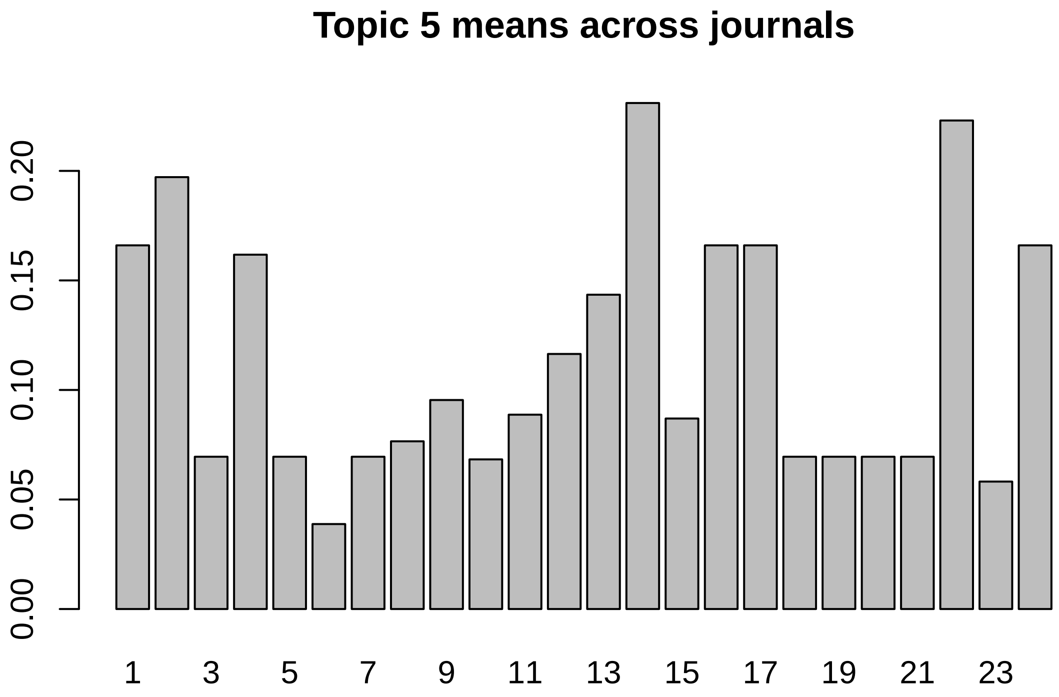
```r
doc.topics.m <- mallet.doc.topics(topic.model,
                                  smoothed = T,
                                  normalized = T)
doc.topics.df <- as.data.frame(doc.topics.m)
#5625 abstracts x 10 topics
journals <- s$journal
doc.topics.df <- cbind(journals,doc.topics.df)
doc.topics.mean.df <- aggregate(doc.topics.df[, 2:ncol(doc.topics.df)],
                              list(doc.topics.df[,1]),
                                   mean)
for (i in 1:NUMTOPICS) {
    title <- paste("Topic",i, "means across journals")
    colVal <- paste("V",i,sep = "")
    barplot(doc.topics.mean.df[,colVal],names.arg = c(1:nrow(doc.topics.mean.df)))
    title(main = title)
}
```

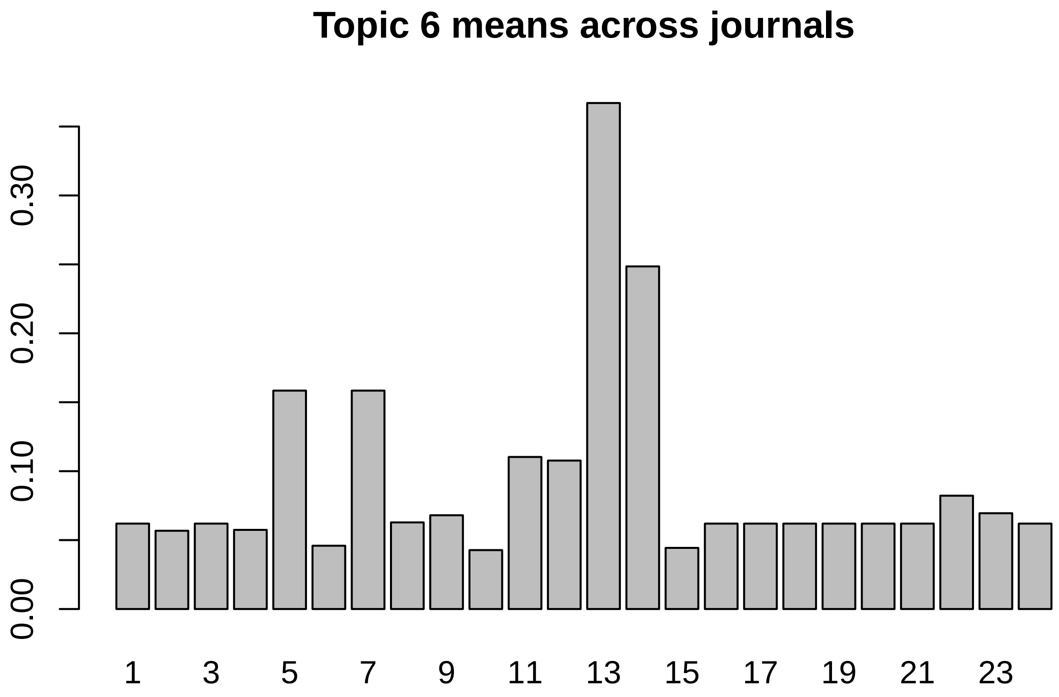

Topic 1 means across journals

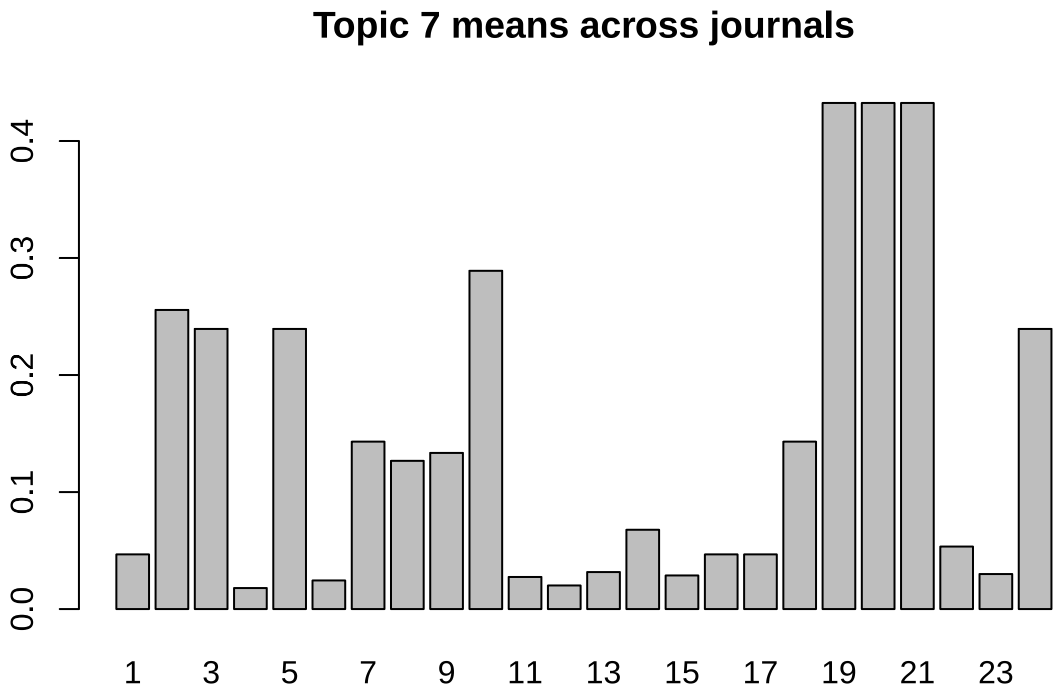

Topic 2 means across journals
0.00
0.05
0.10
0.15
0.20
1
3
5
7
9
11
13
15
17
19
21
23

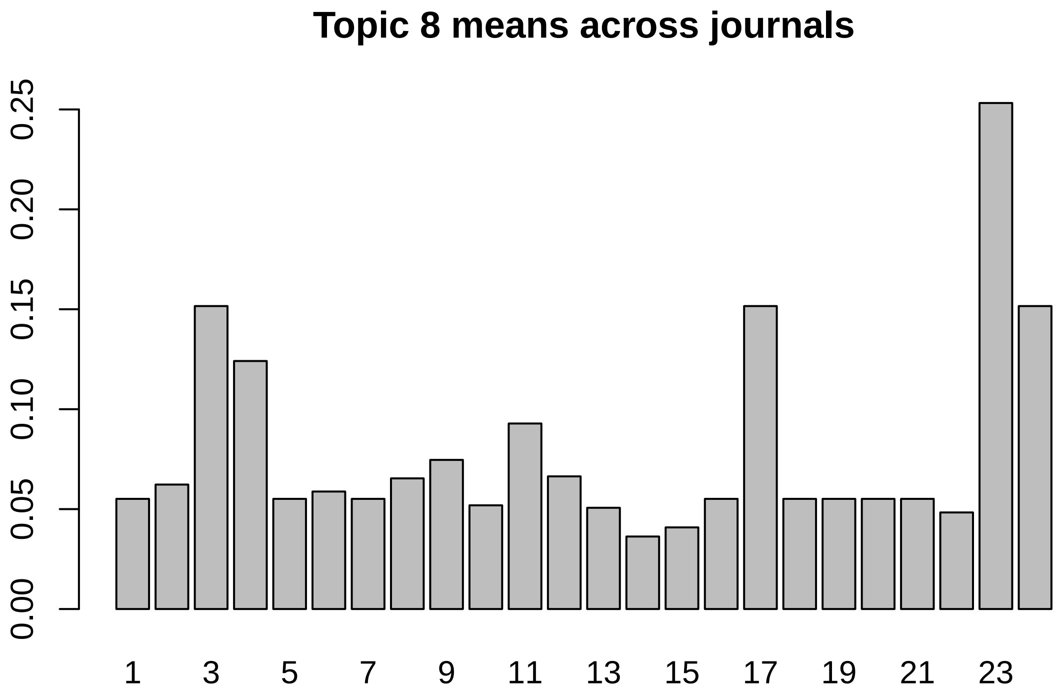

Topic 3 means across journals
0.00
0.10
0.20
0.30
1
3
5
7
9
11
13
15
17
19
21
23

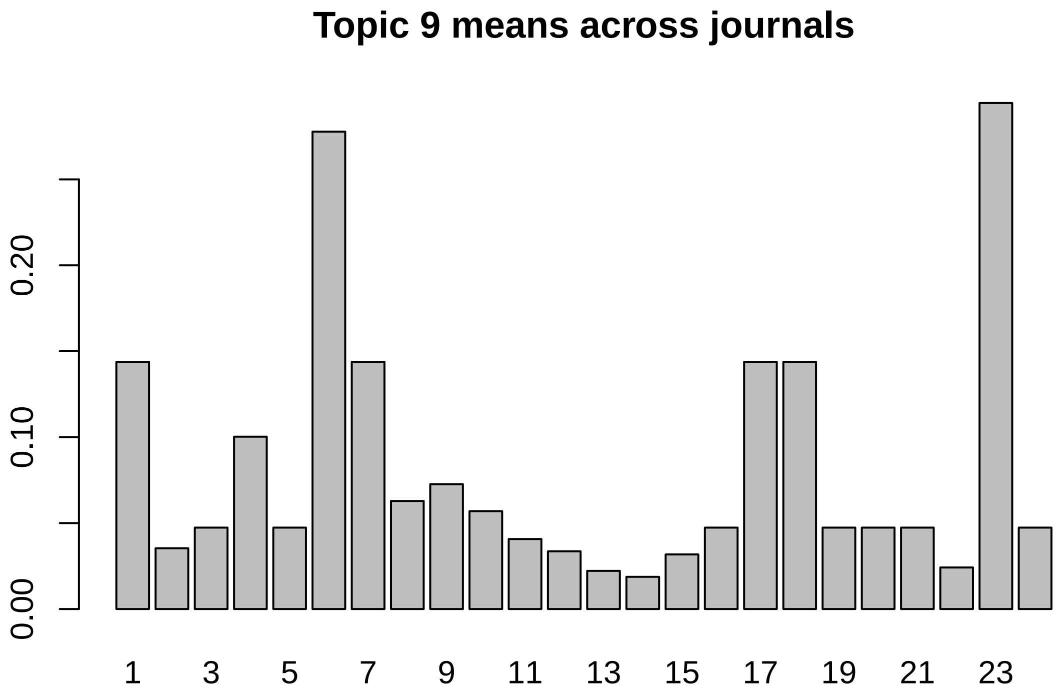

Topic 4 means across journals
0.00
0.10
0.20
0.30
1
3
5
7
9
11
13
15
17
19
21
23

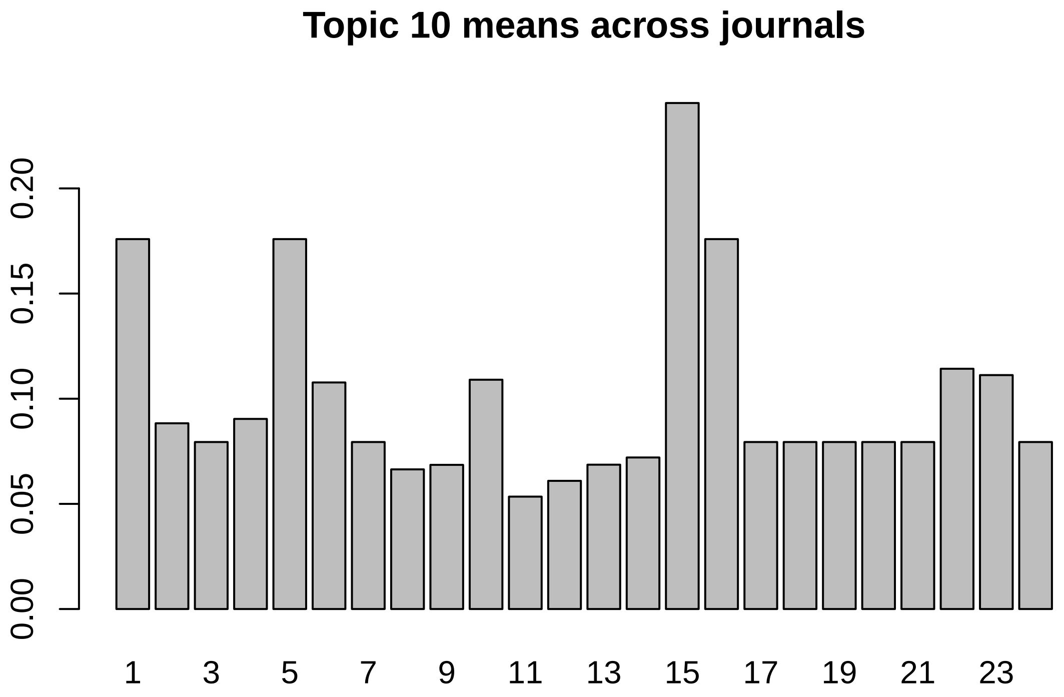

Topic 5 means across journals
0.00
0.05
0.10
0.15
0.20
1
3
5
7
9
11
13
15
17
19
21
23


Topic 6 means across journals
0.00
0.10
0.20
0.30
1
3
5
7
9
11
13
15
17
19
21
23


Topic 7 means across journals
0.0
0.1
0.2
0.3
0.4
1
3
5
7
9
11
13
15
17
19
21
23


Topic 8 means across journals
0.00
0.05
0.10
0.15
0.20
0.25
1
3
5
7
9
11
13
15
17
19
21
23


Topic 9 means across journals
0.00
0.10
0.20
1
3
5
7
9
11
13
15
17
19
21
23



Co-occurence matrix creation.

```
# convert the wide style to long style
idx <- data.frame(1:NUMTOPICS)
colnames(idx) <- "topic"
topic.words.notNormed.df <- cbind(idx,topic.words.notNormed.df)
topic.words.long.df <- melt(topic.words.notNormed.df, id=c("topic"))
dim(topic.words.long.df)
```

```
## [1] 98110     3
```

```
dim(topic.words.long.df)
```

```
## [1] 98110     3
```

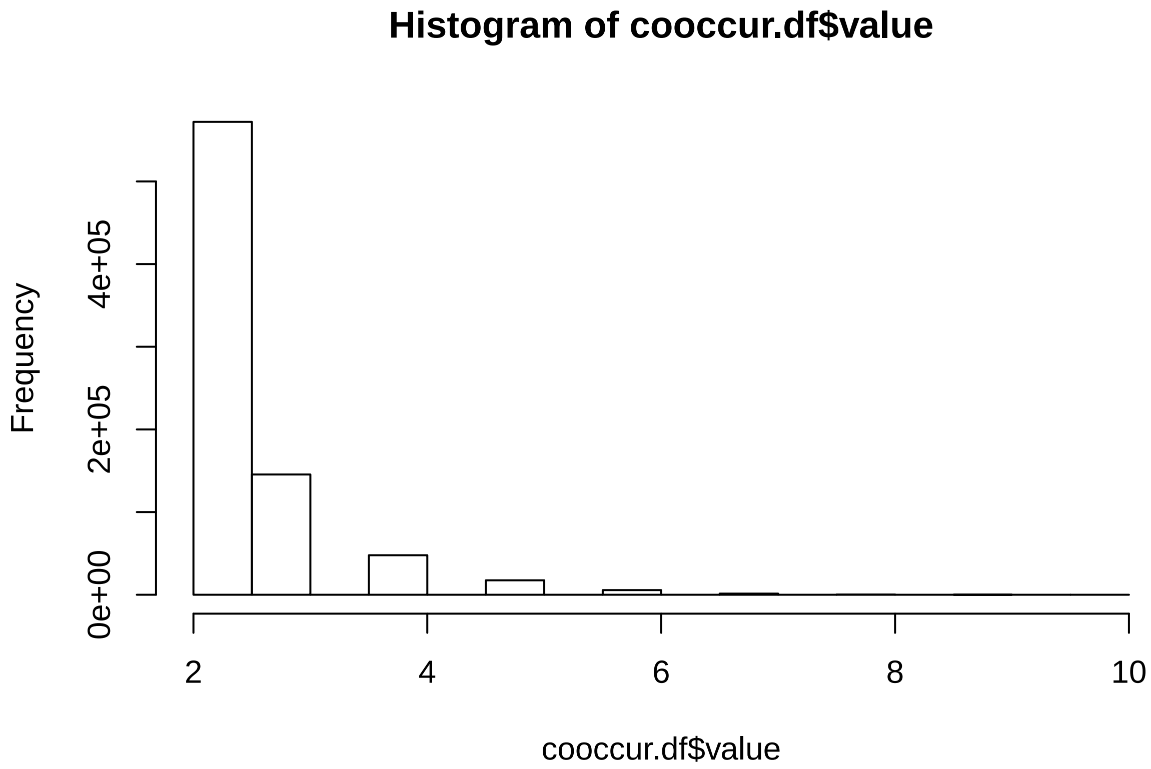
```
# sort topics x words matrix to show what words belong to which topics.
topic.words.sorted <- topic.words.long.df[ order(topic.words.long.df[,1]), ]
```

Calculate cooccurance matrix as t(m) * m, where m is 15k words wide, and 10 topics high. the resultant matrix is 15kx15k and holds number of cooccurances.

```
# topic.words.long.df is a long form matrix with topic, word, value
# First Remove all 0s and 1s from topic.words.long.df
# because if there is only one occurance of a word, it cannot co-occur!
#topic.words.long.df <- topic.words.long.df[topic.words.long.df$value > 1,]
```

```
topic.words.long.df <- topic.words.long.df[topic.words.long.df$value != 0,]
topic.words.long.df <- topic.words.long.df[topic.words.long.df$value != 1,]
topic.fac <- topic.words.long.df[,1]
words.fac <- topic.words.long.df[,2]
sparseM <- sparseMatrix(
        as.numeric(topic.fac),
        as.numeric(words.fac),
        dimnames = list(
                as.character(levels(topic.fac)),
                as.character(levels(words.fac))),
        x = 1)
# calculating co-occurrences
v <- t(sparseM) %*% sparseM
# setting diagonal to zero
diag(v) <- 0

# Now lets take a look at it
summ <- summary(v)
summLU <- summ[summ$i > summ$j,]
cooccur.df <- data.frame(word1 = rownames(v)[summLU$i],
                        word2  = colnames(v)[summLU$j],
                        value  = summLU$x)
cooccur.df <- cooccur.df[cooccur.df$value != 0,]
cooccur.df <- cooccur.df[cooccur.df$value != 1,]
# now sort big to small
cooccur.df <- cooccur.df[ order(-cooccur.df[,3], cooccur.df[,1], cooccur.df[,2]), ]
hist(cooccur.df$value)
```


Histogram of cooccur.df$value
Frequency
0e+00
2e+05
4e+05
2
4
6
8
10
cooccur.df$value