Thoracic Outlet Syndrome Articles and Abstracts

Janis Corona

12/12/2019

## This script takes articles or abstracts on Thoracic Outlet Syndrome-TOS from PubMed or Google Scholar

This creates a directory to stem the abstracts and preprocess from the csv file into a corpus of files in a folder called TOS.

Auto <- read.csv('ThoracicOutletSyndrome.csv', sep=',',  
 header=TRUE, na.strings=c('',' ','NA'))

auto <- Auto[complete.cases(Auto$article),]  
  
  
dir.create('./TOS')  
  
ea <- as.character(auto$article)  
setwd('./TOS')  
  
for (j in 1:length(ea)){  
 write(ea[j], paste(paste('EA',j, sep='.'), '.txt', sep=''))  
}  
setwd('../')

This code preprocesses and stems the corpus

library(tm)  
library(SnowballC)  
library(wordcloud)  
library(ggplot2)  
  
TOS <- Corpus(DirSource("TOS"))  
  
  
TOS

## <<SimpleCorpus>>  
## Metadata: corpus specific: 1, document level (indexed): 0  
## Content: documents: 47

TOS <- tm\_map(TOS, removePunctuation)  
#TOS <- tm\_map(TOS, removeNumbers)  
TOS <- tm\_map(TOS, tolower)  
TOS <- tm\_map(TOS, removeWords, stopwords("english"))  
TOS <- tm\_map(TOS, stripWhitespace)  
TOS <- tm\_map(TOS, stemDocument)  
  
dtmTOS <- DocumentTermMatrix(TOS)  
  
freq <- colSums(as.matrix(dtmTOS))

This code orders words stemmed by frequency and finds input correlations

FREQ <- data.frame(freq)  
ord <- order(freq, decreasing=TRUE)  
  
freq[head(ord, 25)]

## patient tos thorac syndrom outlet rib   
## 241 167 148 142 128 89   
## symptom arteri first surgeri compress studi   
## 84 80 68 68 67 66   
## vascular nerv neurogen treatment can subclavian   
## 62 61 60 57 56 56   
## posit brachial clinic plexus venous physic   
## 52 50 50 50 50 49   
## test   
## 49

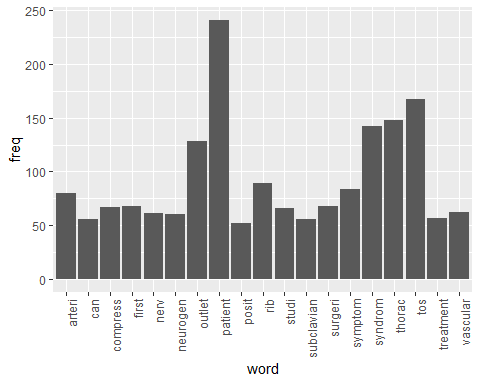
findAssocs(dtmTOS, "patient", corlimit=0.7)

## $patient  
## user neurogen appropri activ venous work   
## 0.82 0.78 0.72 0.71 0.70 0.70

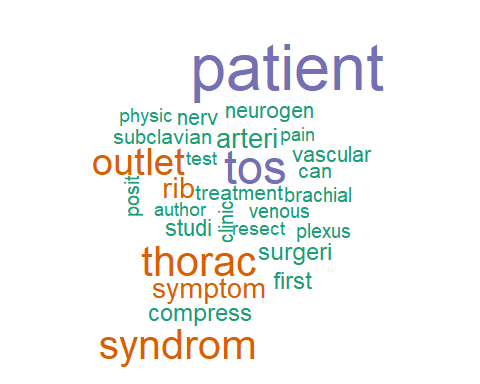
findAssocs(dtmTOS, "pain", corlimit=0.9)

## $pain  
## mani asymptomat volunt call chang   
## 0.95 0.94 0.94 0.92 0.90

wf <- data.frame(word=names(freq), freq=freq)  
p <- ggplot(subset(wf, freq>50), aes(word, freq))  
p <- p + geom\_bar(stat= 'identity')   
p <- p + theme(axis.text.x=element\_text(angle=90, hjust=1))   
p

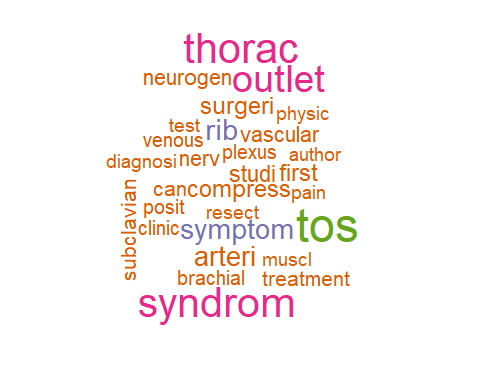


wordcloud(names(freq), freq, min.freq=48,colors=brewer.pal(3,'Dark2'))



wordcloud(names(freq), freq, max.words=30,colors=brewer.pal(6,'Dark2'))

## Warning in wordcloud(names(freq), freq, max.words = 30, colors =  
## brewer.pal(6, : patient could not be fit on page. It will not be plotted.



### The above stemmed the corpus, this will lemmatize the original csv file

and add the field to the table and write out to csv, followed by plot the word count frequencies that were lemmatized and the word clouds

library(textstem)  
  
lemma <- lemmatize\_strings(auto$article, dictionary=lexicon::hash\_lemmas)  
  
Lemma <- as.data.frame(lemma)  
Lemma <- cbind(Lemma, auto)  
  
colnames(Lemma) <- c('lemmatizedAbstract','article', 'source')  
  
write.csv(Lemma, 'LemmatizedTOS.csv', row.names=FALSE)

dir.create('./TOS-Lemma')  
  
ea <- as.character(Lemma$lemmatizedAbstract)  
setwd('./TOS-Lemma')  
  
for (j in 1:length(ea)){  
 write(ea[j], paste(paste('EAL',j, sep='.'), '.txt', sep=''))  
}  
setwd('../')

library(tm)  
library(SnowballC)  
library(wordcloud)  
library(ggplot2)

TOS <- Corpus(DirSource("TOS-Lemma"))  
  
TOS

## <<SimpleCorpus>>  
## Metadata: corpus specific: 1, document level (indexed): 0  
## Content: documents: 46

TOS <- tm\_map(TOS, removePunctuation)  
#TOS <- tm\_map(TOS, removeNumbers)  
TOS <- tm\_map(TOS, tolower)  
TOS <- tm\_map(TOS, removeWords, stopwords("english"))  
TOS <- tm\_map(TOS, stripWhitespace)  
  
dtmTOS <- DocumentTermMatrix(TOS)  
dtmTOS

## <<DocumentTermMatrix (documents: 46, terms: 2429)>>  
## Non-/sparse entries: 6132/105602  
## Sparsity : 95%  
## Maximal term length: 26  
## Weighting : term frequency (tf)

freq <- colSums(as.matrix(dtmTOS))  
  
FREQ <- data.frame(freq)  
ord <- order(freq, decreasing=TRUE)  
  
freq[head(ord, 25)]

## patient tos thoracic syndrome outlet rib   
## 239 176 151 144 131 92   
## symptom first surgery much study nerve   
## 84 69 68 67 66 65   
## neurogenic vascular can subclavian treatment compression   
## 63 63 61 59 57 54   
## plexus brachial venous test muscle physical   
## 53 52 52 51 49 49   
## arterial   
## 48

pain <- as.data.frame(findAssocs(dtmTOS, "pain", corlimit=0.9))  
  
symptom <- as.data.frame(findAssocs(dtmTOS, "symptom", corlimit=0.95))  
  
  
treatment <- as.data.frame(findAssocs(dtmTOS, "treatment", corlimit=0.61))  
  
pain

## pain  
## many 0.95  
## asymptomatic 0.94  
## volunteer 0.94  
## change 0.90

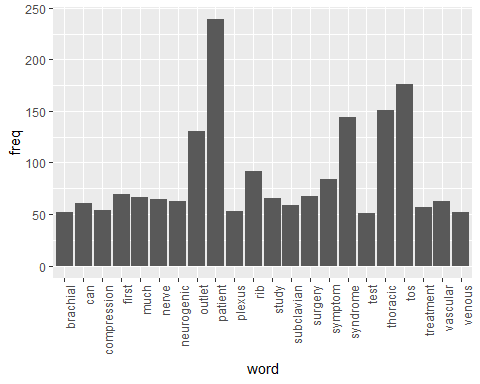
symptom

## symptom  
## pulse 0.97  
## 04050 0.96  
## 175kb 0.96  
## 180ø 0.96  
## 1927 0.96  
## 1945 0.96  
## 1947 0.96  
## 1963 0.96  
## 1965 0.96  
## 196615 0.96  
## 1980 0.96  
## 1998 0.96  
## 1st 0.96  
## 2007 0.96  
## 2500 0.96  
## 3rd 0.96  
## 4th 0.96  
## 5th 0.96  
## 601 0.96  
## 604 0.96  
## 90ø 0.96  
## abandon 0.96  
## abduct 0.96  
## abduction 0.96  
## absence 0.96  
## accident 0.96  
## account 0.96  
## activate 0.96  
## actively 0.96  
## adson 0.96  
## adson1 0.96  
## aer 0.96  
## agent 0.96  
## ago 0.96  
## alfred 0.96  
## alteration 0.96  
## antecubital 0.96  
## arteriography 0.96  
## asm 0.96  
## atos 0.96  
## auto 0.96  
## behind 0.96  
## bring 0.96  
## broad 0.96  
## chin 0.96  
## circumference 0.96  
## classical 0.96  
## claudication 0.96  
## coldness 0.96  
## color 0.96  
## comparable 0.96  
## contralateral 0.96  
## cyanosis 0.96  
## digital 0.96  
## diminish 0.96  
## distend 0.96  
## dorsiflex 0.96  
## duplication 0.96  
## dye 0.96  
## ear 0.96  
## east 0.96  
## easy 0.96  
## elevate 0.96  
## elevation 0.96  
## elicit 0.96  
## eliminate 0.96  
## elvey 0.96  
## elvey5 0.96  
## erroneous 0.96  
## erroneously 0.96  
## essentially 0.96  
## every 0.96  
## evident 0.96  
## exam 0.96  
## examiner 0.96  
## excessive 0.96  
## execute 0.96  
## expense 0.96  
## express 0.96  
## fairly 0.96  
## famous 0.96  
## favor 0.96  
## finger 0.96  
## flex 0.96  
## formation 0.96  
## frequently 0.96  
## gilroy 0.96  
## hammondmdabneal 0.96  
## headache 0.96  
## historically 0.96  
## imply 0.96  
## inclusive 0.96  
## infallible 0.96  
## investigate 0.96  
## just 0.96  
## lab 0.96  
## label 0.96  
## large 0.96  
## leg 0.96  
## light 0.96  
## limit 0.96  
## lodge 0.96  
## maneuver 0.96  
## meyer14 0.96  
## mislead 0.96  
## misnomer 0.96  
## modifier 0.96  
## move 0.96  
## mural 0.96  
## ncv 0.96  
## necessarily 0.96  
## nervous 0.96  
## noninvasive 0.96  
## nonthrombotic 0.96  
## now 0.96  
## obliterate 0.96  
## obliteration 0.96  
## occasional 0.96  
## occipital 0.96  
## opposite 0.96  
## original 0.96  
## originally 0.96  
## overactive 0.96  
## pallor 0.96  
## panelrichard 0.96  
## passively 0.96  
## past 0.96  
## pathognomonic 0.96  
## permit 0.96  
## popularize 0.96  
## precede 0.96  
## press 0.96  
## progressively 0.96  
## provocative 0.96  
## raise 0.96  
## raobab 0.96  
## really 0.96  
## recorder 0.96  
## recording 0.96  
## relabel 0.96  
## rename 0.96  
## reproduction 0.96  
## res 0.96  
## ron 0.96  
## rotation 0.96  
## run 0.96  
## sandersmdab 0.96  
## scar 0.96  
## schrotter 0.96  
## seat 0.96  
## second 0.96  
## serve 0.96  
## silent 0.96  
## simple 0.96  
## simply 0.96  
## since 0.96  
## sit 0.96  
## situation 0.96  
## sometimes 0.96  
## specify 0.96  
## spontaneously 0.96  
## statistic 0.96  
## stick 0.96  
## stoney 0.96  
## straight 0.96  
## stress 0.96  
## strong 0.96  
## succession 0.96  
## suggestive 0.96  
## superficial 0.96  
## tension 0.96  
## theoretically 0.96  
## thing 0.96  
## thrombotic 0.96  
## tight 0.96  
## today 0.96  
## touch 0.96  
## transfemoral 0.96  
## trapezius 0.96  
## turn 0.96  
## typical 0.96  
## ultt 0.96  
## uncertainty 0.96  
## unnecessary 0.96  
## unrelated 0.96  
## unreliable 0.96  
## unwarranted 0.96  
## upright 0.96  
## usually 0.96  
## utilize 0.96  
## wall 0.96  
## way 0.96  
## weak 0.96  
## whereas 0.96  
## will 0.96  
## woods8 0.96  
## wright7 0.96  
## wrist 0.96  
## position 0.95  
## side 0.95  
## test 0.95

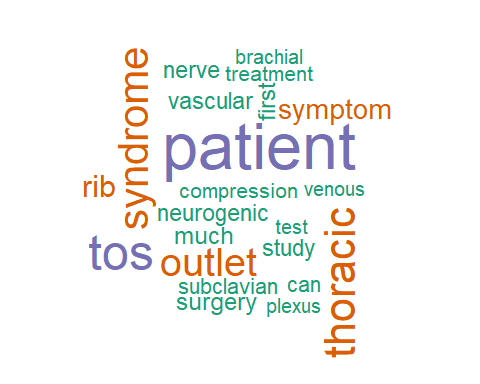
treatment

## treatment  
## placebo 0.70  
## 1002 0.68  
## 14651858 0.68  
## 2010 0.68  
## agree 0.68  
## altmetric 0.68  
## amed 0.68  
## analog 0.68  
## authors 0.68  
## beneficial 0.68  
## bias 0.68  
## cd007218 0.68  
## cochrane 0.68  
## collection 0.68  
## dare 0.68  
## declaration 0.68  
## despite 0.68  
## disputed 0.68  
## espa¤ol 0.68  
## exclusively 0.68  
## extract 0.68  
## great 0.68  
## guideline 0.68  
## handbook 0.68  
## hansson 0.68  
## interest 0.68  
## language 0.68  
## minimum 0.68  
## months 0.68  
## natural 0.68  
## neuroplasty 0.68  
## paresthesias 0.68  
## people 0.68  
## povlsen 0.68  
## progression 0.68  
## pub3 0.68  
## publication 0.68  
## quasi 0.68  
## rcts 0.68  
## register 0.68  
## rigorous 0.68  
## saline 0.68  
## sebastian 0.68  
## snbp 0.68  
## specialize 0.68  
## tfrr 0.68  
## thomas 0.68  
## tingle 0.68  
## version 0.68  
## view 0.68  
## visual 0.68  
## whats 0.68  
## accept 0.67  
## randomize 0.67  
## strength 0.66  
## six 0.66  
## trial 0.65  
## lack 0.64  
## search 0.64  
## generally 0.62

wf <- data.frame(word=names(freq), freq=freq)  
p <- ggplot(subset(wf, freq>50), aes(word, freq))  
p <- p + geom\_bar(stat= 'identity')   
p <- p + theme(axis.text.x=element\_text(angle=90, hjust=1))   
p



wordcloud(names(freq), freq, min.freq=50,colors=brewer.pal(3,'Dark2'))



wordcloud(names(freq), freq, max.words=30,colors=brewer.pal(6,'Dark2'))

