Low Back Pain PubMed

Janis Corona

12/9/2019

This script takes ten articles from the abstracts on low back pain articles from NCBI's PubMed

This creates a directory to stem the abstracts and preprocess from the csv file into a corpus of 20 files in a folder called LowBackPain.

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

auto <- Auto[complete.cases(Auto$abstract),]


dir.create('./LowBackPain')

ea <- as.character(auto$abstract)
setwd('./LowBackPain')

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)

LowBackPain <- Corpus(DirSource("LowBackPain"))


LowBackPain
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 20
```

```
#LowBackPain <- tm_map(LowBackPain, removePunctuation)
#LowBackPain <- tm_map(LowBackPain, removeNumbers)
LowBackPain <- tm_map(LowBackPain, tolower)
LowBackPain <- tm_map(LowBackPain, removeWords, stopwords("english"))
LowBackPain <- tm_map(LowBackPain, stripWhitespace)
LowBackPain <- tm_map(LowBackPain, stemDocument)

dtmLowBackPain <- DocumentTermMatrix(LowBackPain)
```

```
freq <- colSums(as.matrix(dtmLowBackPain))
```

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       pain       back      studi        low      group
##         68         62         50         46         37         32
##    chronic      joint     compar     improv     differ      pain.
##         28         27         24         23         22         20
##   signific     effect     includ        lbp  treatment   function
##         20         19         18         18         18         18
##     outcom     report        one        use     observ sacroiliac
##         17         17         16         16         16         15
##     disabl
##         13
```

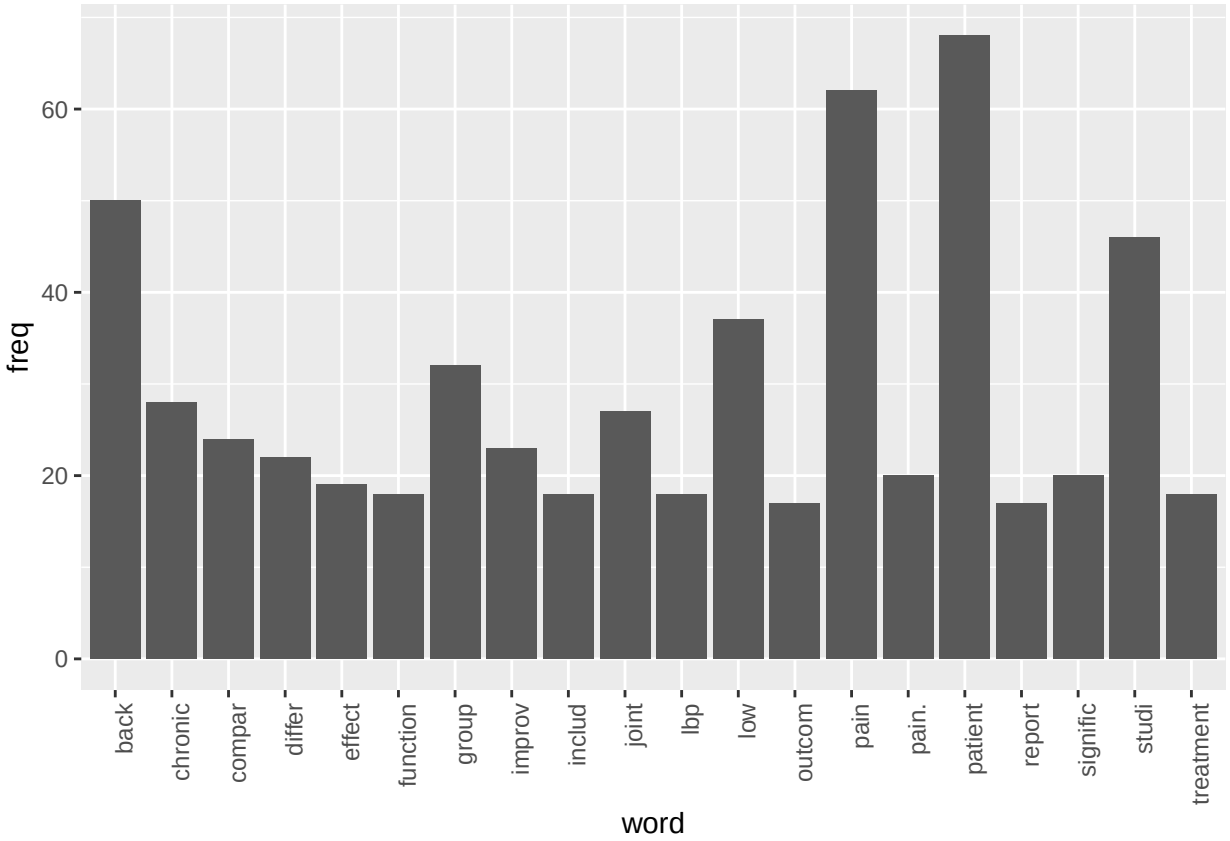
```
findAssocs(dtmLowBackPain, "patient", corlimit=0.7)
```

```
## $patient
##        cours characterist questionnair        score       outcom
##         0.83         0.78         0.78         0.77         0.74
##       compar       improv      complet
##         0.73         0.72         0.70
```

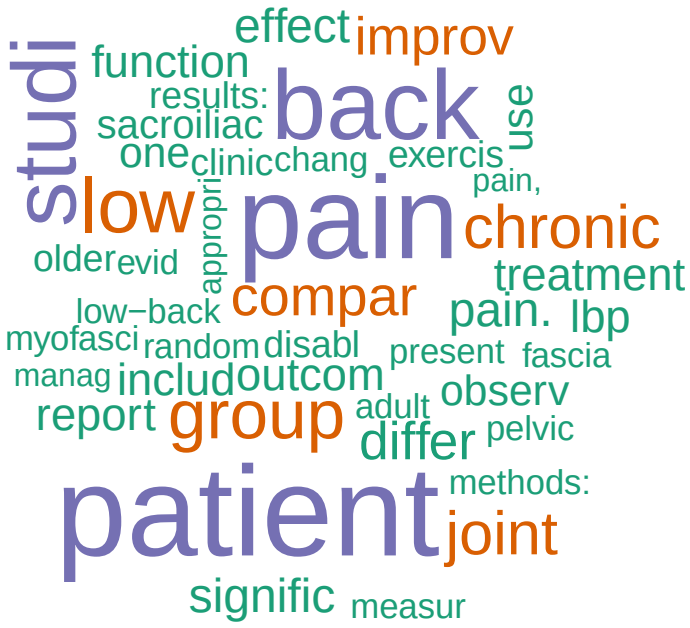
```
findAssocs(dtmLowBackPain, "pain", corlimit=0.62)
```

```
## $pain
##         also     although     advantag      convent       especi
##         0.69         0.68         0.64         0.64         0.64
##      general   non-specif      possibl single-blind        appli
##         0.64         0.64         0.64         0.64         0.63
##     myofasci
##         0.63
```

```
wf <- data.frame(word=names(freq), freq=freq)
p <- ggplot(subset(wf, freq>16), 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=10,colors=brewer.pal(3,'Dark2'))
```




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

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$abstract, dictionary=lexicon::hash_lemmas)

Lemma <- as.data.frame(lemma)
Lemma <- cbind(Lemma, auto)

colnames(Lemma) <- c('lemmatizedAbstract','abstract', 'source')

write.csv(Lemma, 'LemmatizedLowBackPain.csv', row.names=FALSE)
```

```
dir.create('./LowBackPain-Lemma')

ea <- as.character(Lemma$lemmatizedAbstract)
setwd('./LowBackPain-Lemma')

for (j in 1:length(ea)){
  write(ea[j], paste(paste('EAL',j, sep='.'), '.txt', sep=''))
}
setwd('../')
```

```r
library(tm)
library(SnowballC)
library(wordcloud)
library(ggplot2)
```

```r
LowBackPain <- Corpus(DirSource("LowBackPain-Lemma"))

LowBackPain
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 20
```

```r
#LowBackPain <- tm_map(LowBackPain, removePunctuation)
#LowBackPain <- tm_map(LowBackPain, removeNumbers)
LowBackPain <- tm_map(LowBackPain, tolower)
LowBackPain <- tm_map(LowBackPain, removeWords, stopwords("english"))
LowBackPain <- tm_map(LowBackPain, stripWhitespace)

dtmLowBackPain <- DocumentTermMatrix(LowBackPain)
dtmLowBackPain
```

```
## <<DocumentTermMatrix (documents: 20, terms: 1575)>>
## Non-/sparse entries: 2556/28944
## Sparsity           : 92%
## Maximal term length: 19
## Weighting          : term frequency (tf)
```

```r
freq <- colSums(as.matrix(dtmLowBackPain))

FREQ <- data.frame(freq)
ord <- order(freq, decreasing=TRUE)

freq[head(ord, 25)]
```

```
##     patient        pain        back         low       study       group
##          67          63          61          53          48          36
##     chronic       joint         lbp     compare       pain.      report
##          28          27          26          23          21          20
##   treatment     include         one     outcome         use        much
##          19          18          18          17          17          16
## improvement  functional  sacroiliac significant conclusion:    exercise
##          15          15          15          14          13          13
##     result:
##          13
```

```r
patient <- as.data.frame(findAssocs(dtmLowBackPain, "patient", corlimit=0.62))

result <- as.data.frame(findAssocs(dtmLowBackPain, "result", corlimit=0.56))
```

```r
treatment <- as.data.frame(findAssocs(dtmLowBackPain, "treatment", corlimit=0.65))

patient
```

```
##                patient
## course            0.85
## characteristic    0.80
## questionnaire     0.80
## score             0.79
## improvement       0.77
## compare           0.71
## 001               0.70
## 01.               0.70
## 15.               0.70
## 214               0.70
## 34.               0.70
## background        0.70
## behavioral        0.70
## clbp              0.70
## clbp.             0.70
## clearly           0.70
## clinically        0.70
## cognitive         0.70
## completion        0.70
## control.          0.70
## costly            0.70
## datum:            0.70
## disturbance,      0.70
## disturbance.      0.70
## empirical         0.70
## fatigue,          0.70
## function,         0.70
## health,           0.70
## historical        0.70
## include:          0.70
## interference,     0.70
## ipp               0.70
## ipps              0.70
## lack              0.70
## match             0.70
## mdq               0.70
## meaningful        0.70
## pair              0.70
## participation     0.70
## post              0.70
## potentially       0.70
## pre               0.70
## program           0.70
## promis            0.70
## propensity        0.70
## pt.               0.70
## questionnaire.    0.70
## satisfaction,     0.70
```

```
## seventeen         0.70
## summary           0.70
## upon              0.70
## ò10               0.70
## ò3                0.70
## outcome           0.69
## complete          0.68
## measure           0.66
## objective:        0.66
## similar           0.64
## 60.               0.63
## measure.          0.63
```

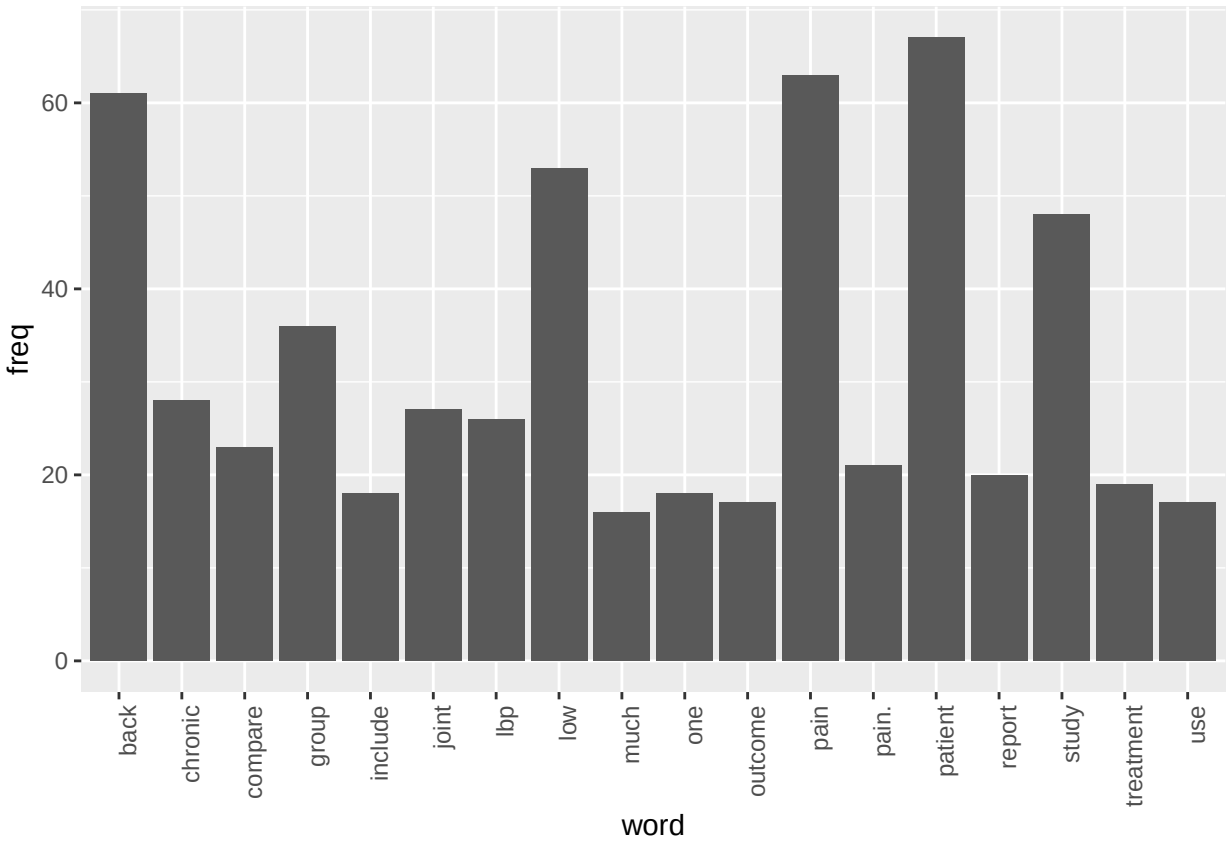
```
result
```

```
##              result
## furthermore,   0.76
## network        0.76
## trial          0.60
```

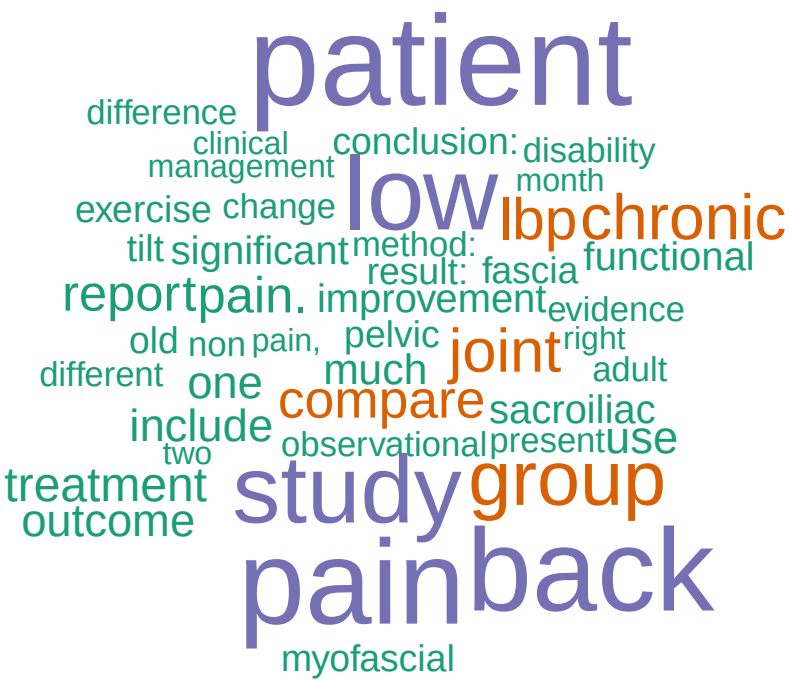
```
treatment
```

```
##            treatment
## reflect         0.83
## future          0.74
## individual      0.74
## relate          0.74
## core            0.73
## group.          0.73
## receive         0.71
## little          0.70
## condition.      0.70
## apply           0.70
## program.        0.70
```

```
wf <- data.frame(word=names(freq), freq=freq)
p <- ggplot(subset(wf, freq>15), 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=10,colors=brewer.pal(3,'Dark2'))
```




```
wordcloud(names(freq), freq, max.words=40,colors=brewer.pal(6,'Dark2'))
```


back
significant
treatment
result:
study
joint
pain.
pain
lbp
one
outcome
change
exercise
observational
group
sacroiliac
old
adult
present
disability
report
non
fascia
use
evidence
chronic
much
compare
different
tilt
myofascial
pelvic
difference
include
method:
conclusion:
functional
low
improvement
patient