Calf Cramps PubMed

Janis Corona

12/9/2019

This script takes ten articles from the abstracts on earache 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 Earache.

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

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


dir.create('./Calf_Cramps')

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

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)

Calf_Cramps <- Corpus(DirSource("Calf_Cramps"))


Calf_Cramps
```

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

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

```
dtmCalf_Cramps <- DocumentTermMatrix(Calf_Cramps)

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

This code orders words stemmed by frequency and finds input correlations

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

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

```
##     cramp     muscl      calf       leg      pain   patient   nocturn
##        61        47        43        41        32        31        24
##      week    effect     studi  results:  signific    compar     sleep
##        20        17        16        15        15        14        13
## treatment  pregnant    clinic   cramps.   qualiti    report    stimul
##        13        13        12        11        11        11        11
##    inject     group  methods:   syndrom
##        11        10        10        10
```

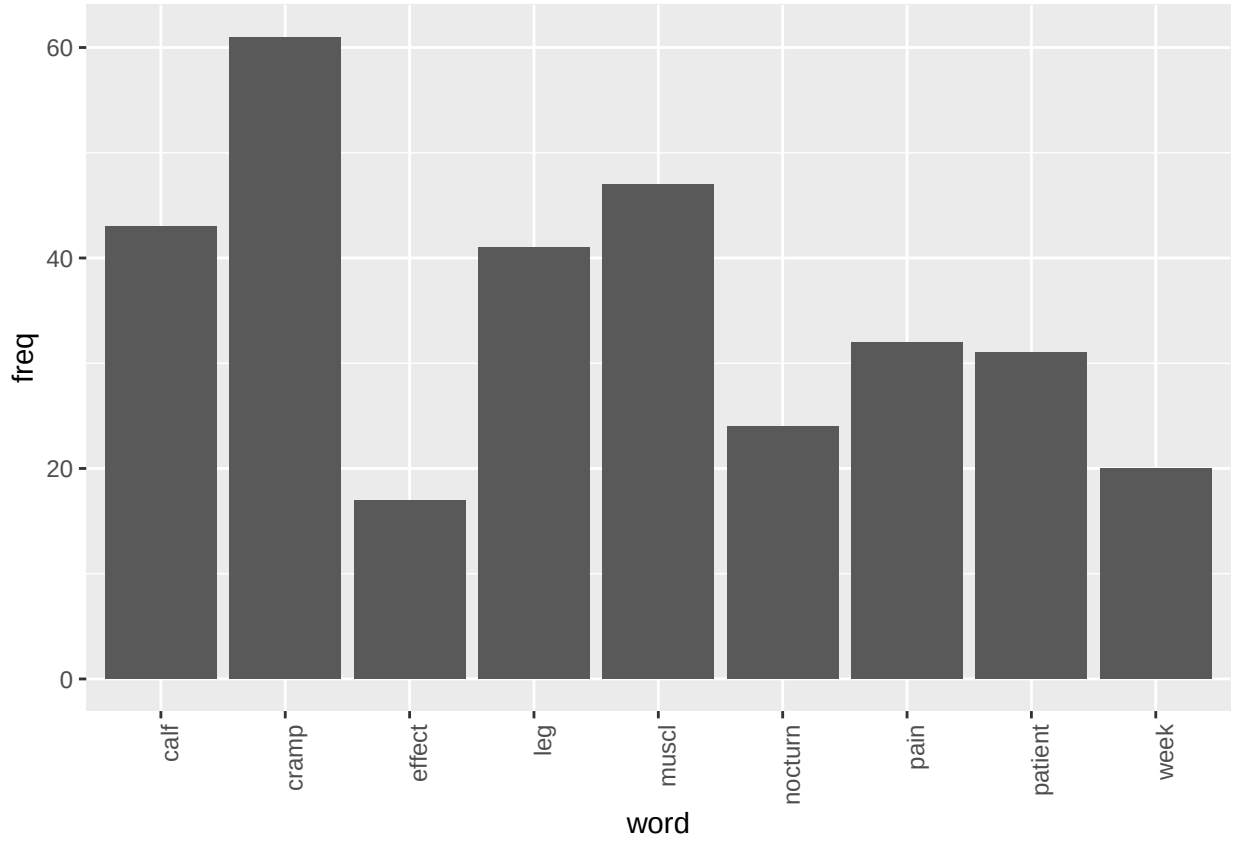
```
findAssocs(dtmCalf_Cramps, "sleep", corlimit=0.7)
```

```
## $sleep
##    qualiti  (mos-ss).    0.001),     0.003)    0.007),      0.02)
##       0.96       0.95       0.95       0.95       0.95       0.95
##     0.02).     0.03).      adequ       age)     aspect australia.
##       0.95       0.95       0.95       0.95       0.95       0.95
##     bodili    central      coast  controls.     domain     eighti
##       0.95       0.95       0.95       0.95       0.95       0.95
##    explain    greater health-rel     impact       larg       less
##       0.95       0.95       0.95       0.95       0.95       0.95
##     mental     mos-ss      negat      never   newcastl      peopl
##       0.95       0.95       0.95       0.95       0.95       0.95
##  primarili    problem   purpose:   quantiti     region       role
##       0.95       0.95       0.95       0.95       0.95       0.95
##  sex-match      sf-36    sf-36v2     sleep.      snore  substanti
##       0.95       0.95       0.95       0.95       0.95       0.95
##    summari     survey     health      life.    nocturn       life
##       0.95       0.95       0.89       0.83       0.80       0.76
##    disturb       age-      south     wales,       year      reduc
##       0.73       0.73       0.73       0.73       0.73       0.71
##  experienc
##       0.71
```

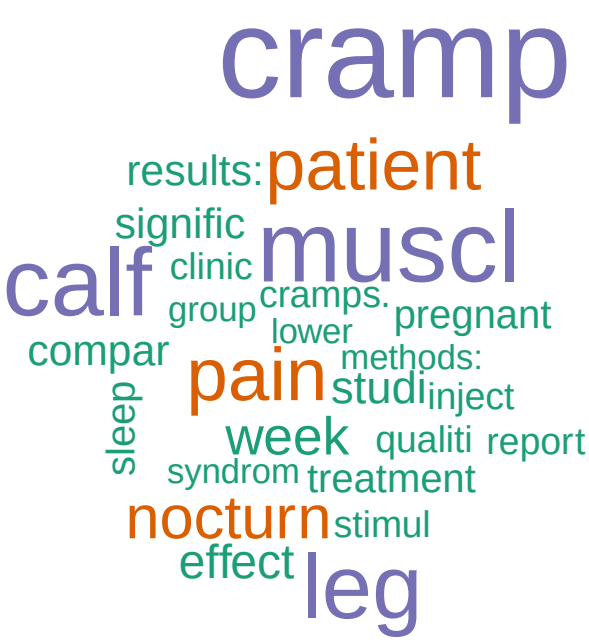
```
findAssocs(dtmCalf_Cramps, "pain", corlimit=0.5)
```

```
## $pain
##    patient  patients,       rate     study.     report      daili
##       0.60       0.58       0.58       0.58       0.57       0.56
##     cramp;   insomnia discomfort
##       0.52       0.51       0.51
```

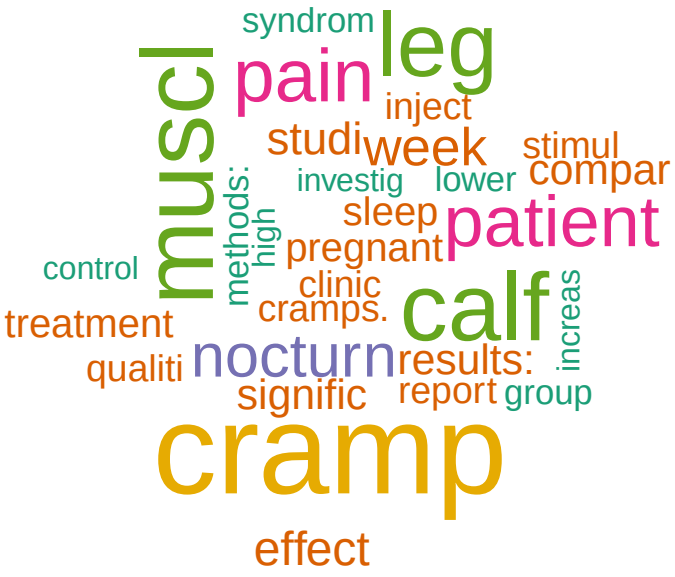
```
wf <- data.frame(word=names(freq), freq=freq)
p <- ggplot(subset(wf, freq>16), aes(word, freq))
p <- p + geom_bar(stat= 'identity')
p <- p + theme(axis.text.x=element_text(angle=90, hjust=1))
p
```




```
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, 'LemmatizedCalf_Cramps.csv', row.names=FALSE)
```

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

ea <- as.character(Lemma$lemmatizedAbstract)
setwd('./Calf_Cramps-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
Calf_Cramps <- Corpus(DirSource("Calf_Cramps-Lemma"))

Calf_Cramps
```

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

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

dtmCalf_Cramps <- DocumentTermMatrix(Calf_Cramps)
dtmCalf_Cramps
```

```
## <<DocumentTermMatrix (documents: 20, terms: 1209)>>
## Non-/sparse entries: 1989/22191
## Sparsity           : 92%
## Maximal term length: 20
## Weighting          : term frequency (tf)
```

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

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

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

```
##         cramp        muscle          calf           leg       patient
##            66            47            45            42            30
##          pain     nocturnal          week         study           btx
##            28            24            20            16            16
##       result:         sleep     treatment      clinical   conclusion:
##            15            15            14            13            13
##       method:         group       compare           low      pregnant
##            13            13            13            13            13
##        cramp.       quality        report       control significantly
##            11            11            11            10            10
```

```r
patient <- as.data.frame(findAssocs(dtmCalf_Cramps, "patient", corlimit=0.70))

Calf_Cramps <- as.data.frame(findAssocs(dtmCalf_Cramps, "calf", corlimit=0.75))
```

```
treatment <- as.data.frame(findAssocs(dtmCalf_Cramps, "treatment", corlimit=0.55))

patient
```

```
##                patient
## assessment        0.84
## global            0.84
## interventional    0.84
## prospective,      0.84
## improve           0.81
## frequent          0.77
## daily             0.75
## intensity         0.73
## patient,          0.72
```

```
Calf_Cramps
```

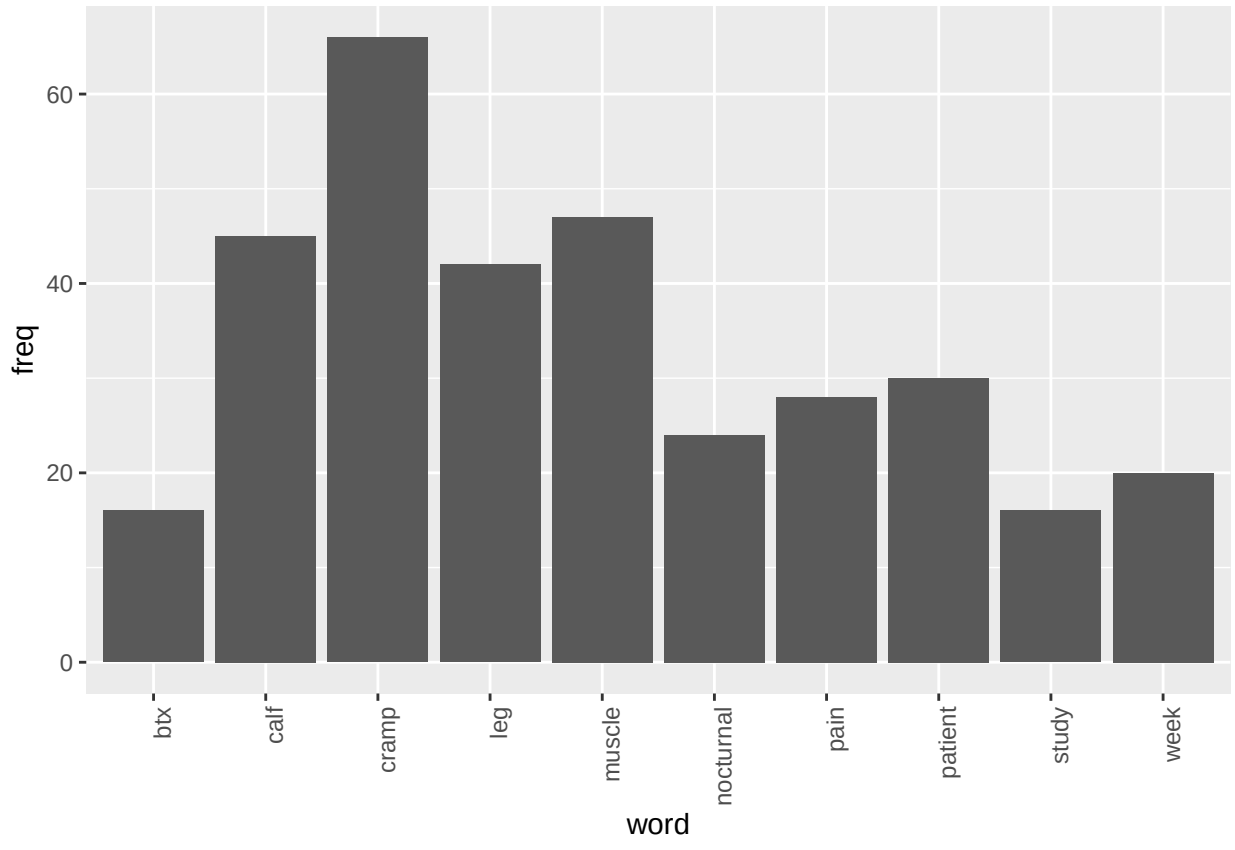
```
##                calf
## one            0.87
## 1.13           0.82
## 1.18           0.82
## 1.45           0.82
## 111            0.82
## 16,            0.82
## 180            0.82
## 186            0.82
## 2.96           0.82
## 2010           0.82
## 2012.          0.82
## 28.            0.82
## 32.            0.82
## 342            0.82
## 39,            0.82
## 420            0.82
## 492            0.82
## 5.76           0.82
## 50.            0.82
## 500            0.82
## 582            0.82
## 673            0.82
## 748            0.82
## 8.02           0.82
## 952            0.82
## 971            0.82
## =349.          0.82
## =6.            0.82
## area,          0.82
## balance        0.82
## call           0.82
## china          0.82
## china.         0.82
## chinese        0.82
```

```
## classify       0.82
## cluster        0.82
## county         0.82
## dairy          0.82
## demographic    0.82
## diagnose.      0.82
## dietary        0.82
## dynamic        0.82
## economic       0.82
## ffq            0.82
## first,         0.82
## fruit          0.82
## group,         0.82
## hypertension   0.82
## information,   0.82
## intake         0.82
## mainland.      0.82
## multi          0.82
## occupation,    0.82
## p=0.           0.82
## pattern        0.82
## pattern,       0.82
## pattern.       0.82
## period;        0.82
## peripartum     0.82
## probability    0.82
## proportional   0.82
## province       0.82
## quantitative   0.82
## questionnaire. 0.82
## randomization  0.82
## relatively     0.82
## residential    0.82
## respectively,  0.82
## sample         0.82
## semi           0.82
## size           0.82
## socio          0.82
## stage          0.82
## status,        0.82
## stratify       0.82
## study;         0.82
## take           0.82
## take.          0.82
## vegetable      0.82
## muscle         0.81
## different      0.81
## investigate    0.80
## use            0.80
## value          0.78
## trimester.     0.78
## prevalence     0.77
## factor         0.75
```

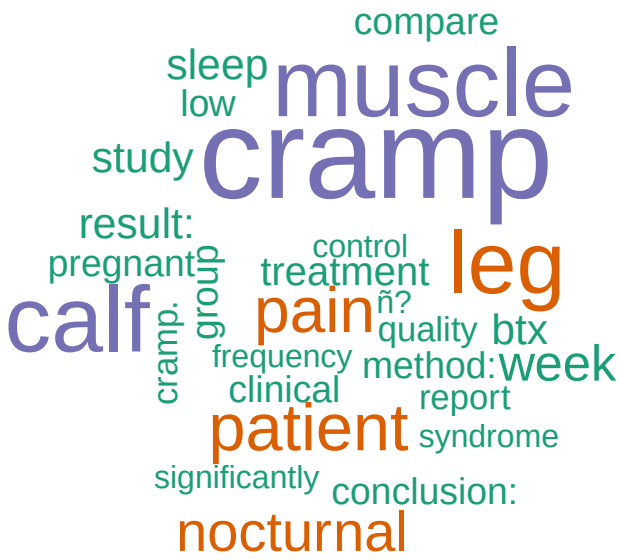
```
treatment
```

```
##               treatment
## effectiveness      0.81
## decrease           0.67
## baseline           0.66
## outcome            0.59
## follow             0.57
## inclusion          0.55
## intervention:      0.55
## stretch            0.55
## course             0.55
## main               0.55
## participant:       0.55
## btx                0.55
## safe               0.55
## toxin              0.55
```

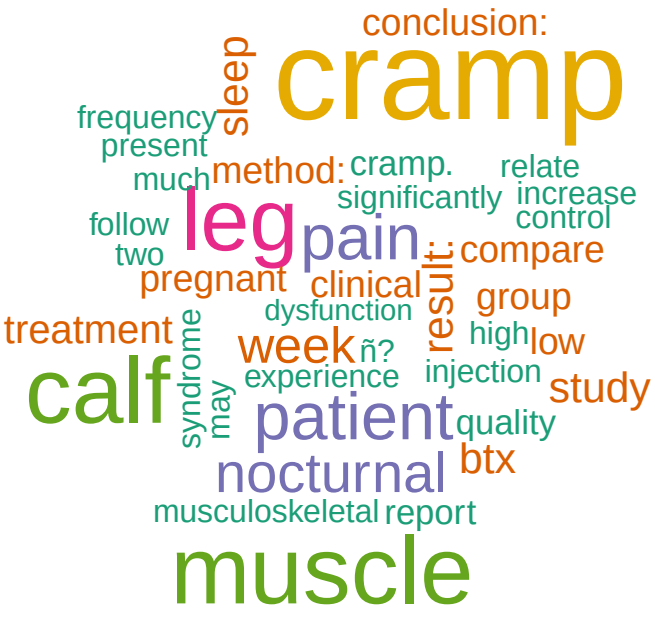
```
wf <- data.frame(word=names(freq), freq=freq)
p <- ggplot(subset(wf, freq>15), aes(word, freq))
p <- p + geom_bar(stat= 'identity')
p <- p + theme(axis.text.x=element_text(angle=90, hjust=1))
p
```



```
wordcloud(names(freq), freq, min.freq=10,colors=brewer.pal(3,'Dark2'))
```




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


conclusion:
cramp
sleep
frequency
present
much
method:
cramp.
relate
significantly
increase
control
follow
two
leg
pain
compare
pregnant
clinical
result:
group
dysfunction
treatment
week
ñ?
high
low
calf
syndrome
may
experience
injection
study
patient
quality
btx
nocturnal
musculoskeletal
report
muscle