Migraine PubMed

Janis Corona

12/9/2019

This script takes twenty articles from the abstracts on Migraine articles from NCBI's PubMed

This creates a directory to stem the abstracts and preprocess from the csv file into a corpus of 10 files in a folder called Migraines.

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

```
## [1] "abstract" "source"
```

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


dir.create('./Migraines')

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

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)

migraine <- Corpus(DirSource("Migraines"))


migraine
```

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

```
migraine <- tm_map(migraine, removePunctuation)
migraine <- tm_map(migraine, removeNumbers)
migraine <- tm_map(migraine, tolower)
migraine <- tm_map(migraine, removeWords, stopwords("english"))
```

```
migraine <- tm_map(migraine, stripWhitespace)
migraine <- tm_map(migraine, stemDocument)

dtmmigraine <- DocumentTermMatrix(migraine)

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

This code orders words stemmed by frequency and finds input correlations

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

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

```
##   migrain   headach     studi treatment    preval       use    disabl
##       247       159        83        62        61        52        46
##   consult      year   patient     medic  diagnosi     women    report
##        43        41        41        40        39        36        35
##  criteria  particip      acut    includ     among      need      rate
##        32        32        30        29        28        28        27
##     sever      care    health       men
##        26        24        24        22
```

```
findAssocs(dtmmigraine, "criteria", corlimit=0.75)
```

```
## $criteria
##                          onset                 acetaminophen
##                           0.87                          0.81
## acetaminophencaffeinebutalbit                 acetylsalicyl
##                           0.81                          0.81
##          acidcaffeinebutalbit                        codein
##                           0.81                          0.81
##                        consent                         cycl
##                           0.81                          0.81
##                       document                        earli
##                           0.81                          0.81
##                         enrol                        formal
##                           0.81                          0.81
##                         fulfil                         goal
##                           0.81                          0.81
##                           head               inclusionexclus
##                           0.81                          0.81
##                       instruct                     ketorolac
##                           0.81                          0.81
##                          local                         mens
##                           0.81                          0.81
##                      menstrual                          mrm
##                           0.81                          0.81
##                          newli                       newspap
##                           0.81                          0.81
##                        painfle                        phase
##                           0.81                          0.81
```

```
##                      plus                preliminari
##                      0.81                       0.81
##                   pretreat                   protocol
##                      0.81                       0.81
##                       rag                     relief
##                      0.81                       0.81
##                    satisfi                   seventyf
##                      0.81                       0.81
##                 sumatriptan                  thirtynin
##                      0.81                       0.81
##                      took                  undiagnos
##                      0.81                       0.81
##                       via                   withdrew
##                      0.81                       0.81
##                  satisfact                       hour
##                      0.80                       0.80
##                    respons                       pain
##                      0.80                       0.76
##                   advertis                    versus
##                      0.76                       0.76
```

```
findAssocs(dtmmigraine, "disabl", corlimit=0.63)
```

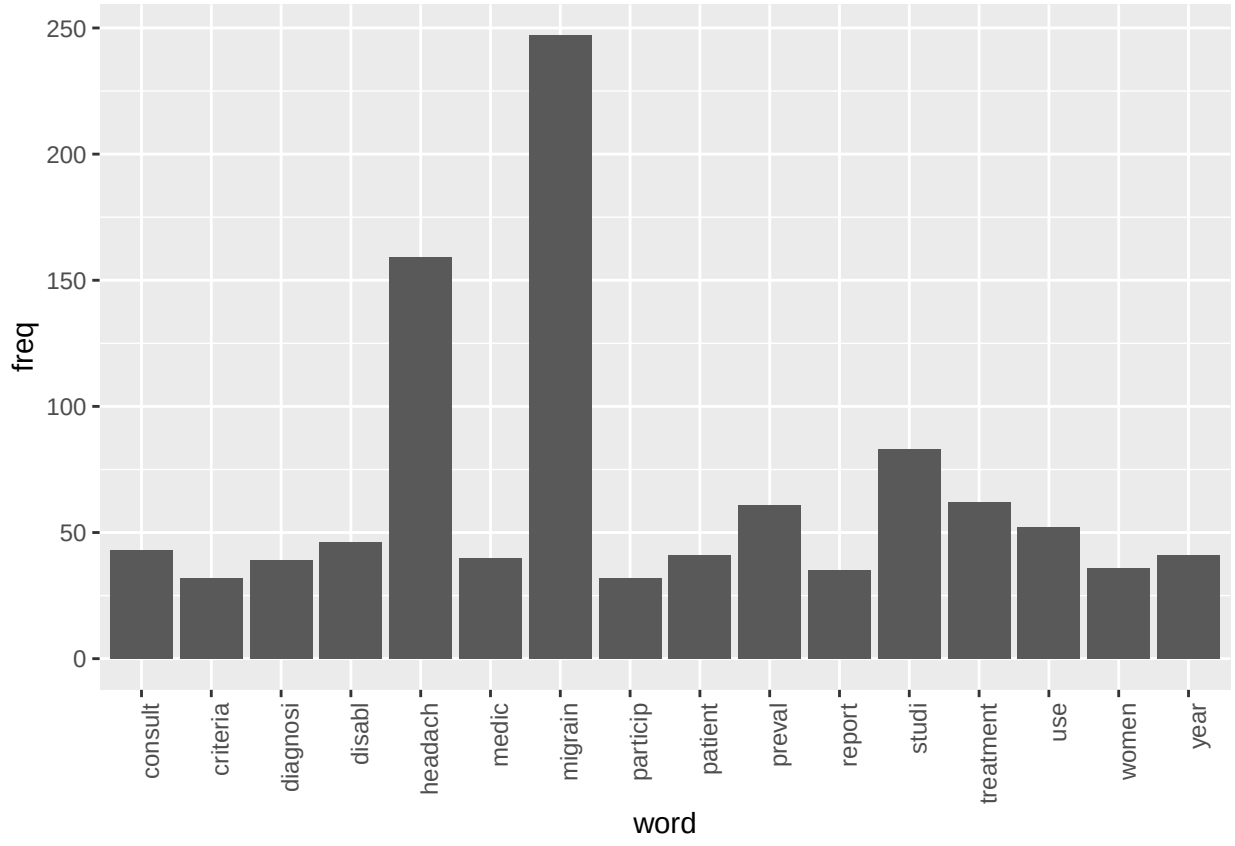
```
## $disabl
##            model          monitor             size            scale
##             0.76             0.74             0.74             0.66
##         function          measur           combin          amongst
##             0.65             0.65             0.65             0.64
##           accept              add          allianc            allow
##             0.63             0.63             0.63             0.63
##             alon           belief             dali           dimens
##             0.63             0.63             0.63             0.63
## disabilityadjust       diseaserel              ehf          environ
##             0.63             0.63             0.63             0.63
##         european           feder             full             gain
##             0.63             0.63             0.63             0.63
##          healthi        highlight            human              icf
##             0.63             0.63             0.63             0.63
##        incomplet           joint         knowledg           launch
##             0.63             0.63             0.63             0.63
##             lift          mortal           nonfat           organ
##             0.63             0.63             0.63             0.63
##         overview         paramet     publichealth           relev
##             0.63             0.63             0.63             0.63
##             role              top       undertaken              wha
##             0.63             0.63             0.63             0.63
##           whilst             whos             work             ylds
##             0.63             0.63             0.63             0.63
```

```
findAssocs(dtmmigraine, "women", corlimit=0.7)
```

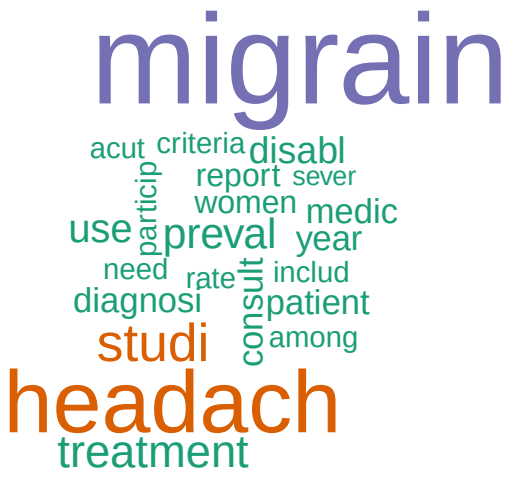
```
## $women
##                  slight                   calcul                canadian
```

```
##                0.93                0.71                0.71
##                less              nineti           nonspecif
##                0.71                0.71                0.71
##             perhap           psychosoci               seven
##                0.71                0.71                0.71
##           somewhat            standard              static
##                0.71                0.71                0.71
##          substanti triptansdihydroergotamin
##                0.71                0.71
```

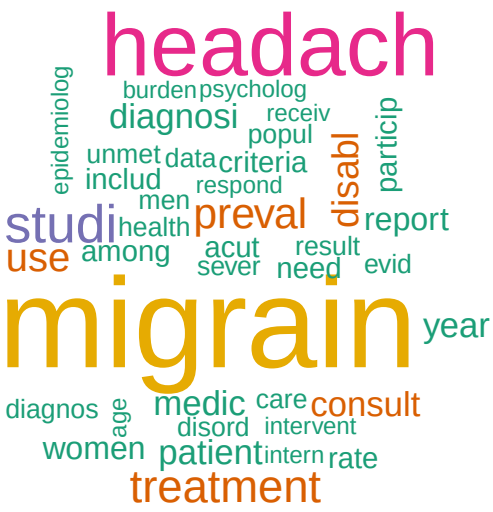
```
wf <- data.frame(word=names(freq), freq=freq)
p <- ggplot(subset(wf, freq>30), aes(word, freq))
p <- p + geom_bar(stat= 'identity')
p <- p + theme(axis.text.x=element_text(angle=90, hjust=1))
p
```




```
wordcloud(names(freq), freq, min.freq=25,colors=brewer.pal(3,'Dark2'))
```




```
wordcloud(names(freq), freq, max.words=40,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, 'Lemmatizedmigraine.csv', row.names=FALSE)
```

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

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

```
migraine <- Corpus(DirSource("migraine-Lemma"))

migraine
```

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

```
# this is an NCBI file so there are dashes and numbers in gene names at times
# that I would like to show, if they are frequent
# migraine <- tm_map(migraine, removePunctuation)
# migraine <- tm_map(migraine, removeNumbers)

migraine <- tm_map(migraine, tolower)
migraine <- tm_map(migraine, removeWords, stopwords("english"))
migraine <- tm_map(migraine, stripWhitespace)

dtmmigraine <- DocumentTermMatrix(migraine)
dtmmigraine
```

```
## <<DocumentTermMatrix (documents: 20, terms: 1507)>>
## Non-/sparse entries: 2971/27169
## Sparsity           : 90%
## Maximal term length: 17
## Weighting          : term frequency (tf)
```

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

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

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

```
##    migraine    headache       study  prevalence         use   treatment
##         209         153          63          56          49          44
##     patient      report        year        much       woman  disability
##          39          35          35          32          32          30
##       acute       among   diagnosis  population participant      health
##          30          28          28          27          27          27
##        high      relate        rate   criterion   migraine,        base
##          26          26          25          24          24          23
##     consult
##          23
```

```
health <- as.data.frame(findAssocs(dtmmigraine, "health", corlimit=0.6))

criteria <- as.data.frame(findAssocs(dtmmigraine, "criteria", corlimit=0.55))


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

health
```

```
##                health
## increasingly     0.78
## model            0.78
## ie,              0.75
## necessary        0.74
## public           0.73
## care             0.72
## non              0.70
## global           0.66
## barrier          0.65
## demographic,     0.65
## socioeconomic,   0.65
## step             0.65
## attributable     0.64
## quarter          0.64
## state            0.64
## publish          0.64
## 1.14             0.64
## care.            0.64
## consult,         0.64
## consult.         0.64
## insurance        0.64
## rate,            0.64
## successfully     0.64
## appropriate      0.63
## currently        0.63
## cause            0.61
```

```
criteria
```

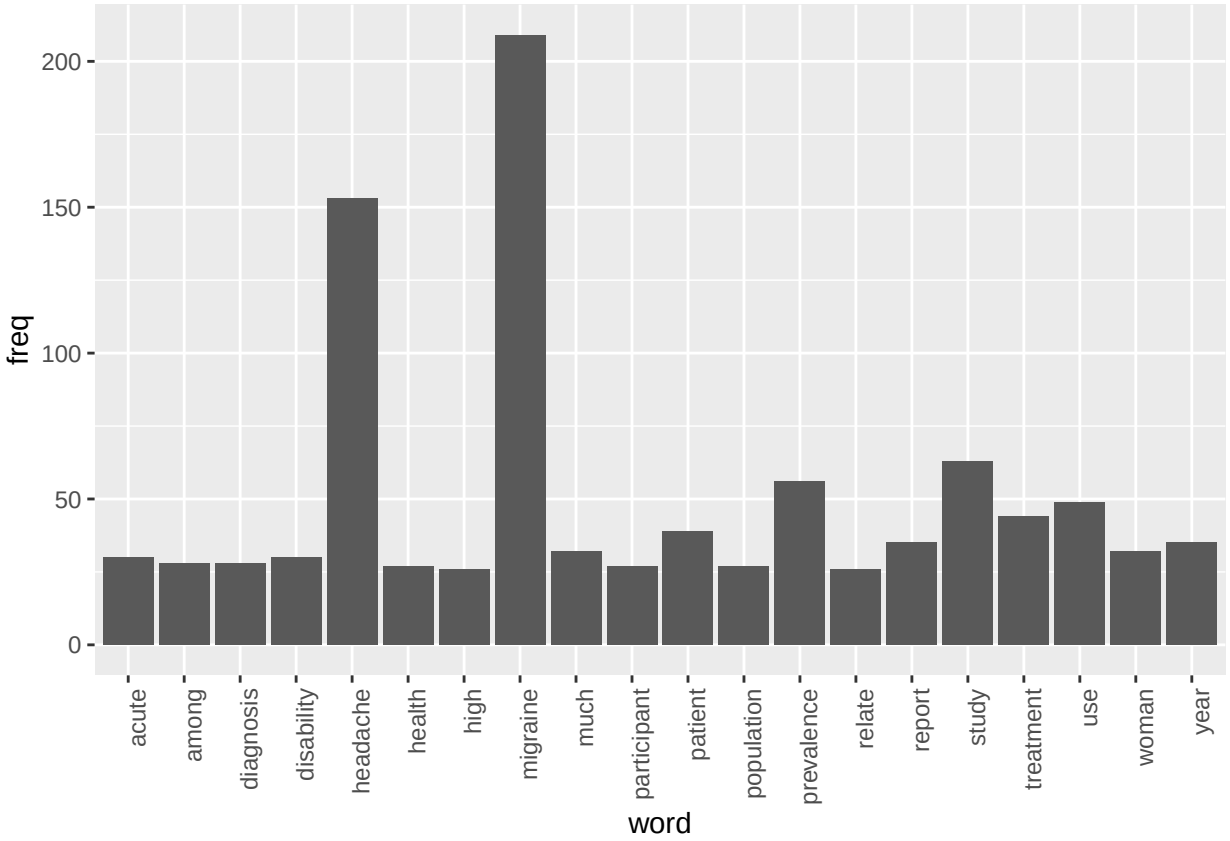
```
## [1] criteria
## <0 rows> (or 0-length row.names)
```

```
treatment
```

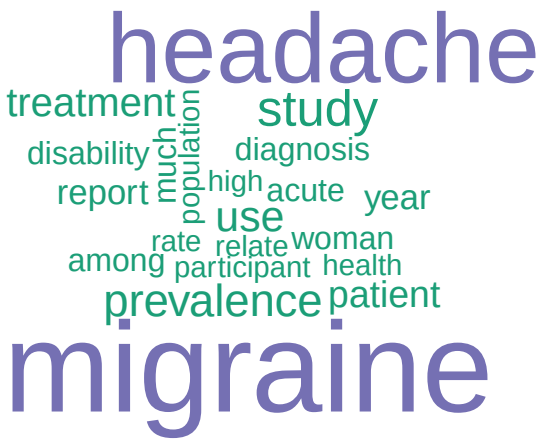
```
##               treatment
## acute              0.81
## relate             0.78
## assessment         0.77
## moderate           0.75
## background:        0.71
## identify           0.70
## less               0.70
```

```
## current            0.69
## american           0.68
## prevention         0.68
## use                0.67
## opioid             0.66
## 2009               0.62
## respondent         0.62
## treatment,         0.62
## medication         0.61
## per                0.61
## therapy.           0.60
## meet               0.59
## frequency          0.59
## sample             0.59
## conclusion:        0.58
## result:            0.58
## episodic           0.58
## objective:         0.58
## method:            0.57
## much               0.57
## category           0.57
## opportunity        0.57
## anti               0.57
## inflammatory       0.57
## recruit            0.57
## 38.                0.57
## beta               0.57
## contrast           0.57
## drug               0.57
## longitudinal,      0.57
## management,        0.57
```

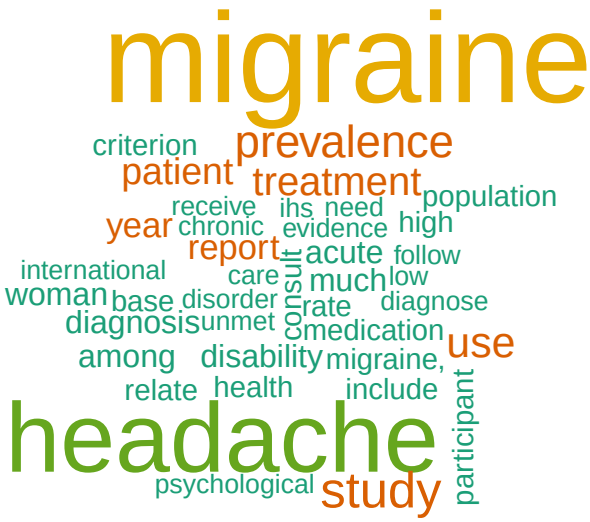
```r
wf <- data.frame(word=names(freq), freq=freq)
p <- ggplot(subset(wf, freq>25), aes(word, freq))
p <- p + geom_bar(stat= 'identity')
p <- p + theme(axis.text.x=element_text(angle=90, hjust=1))
p
```




```
wordcloud(names(freq), freq, min.freq=25,colors=brewer.pal(3,'Dark2'))
```




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


migraine
criterion
prevalence
patient
treatment
population
receive
ihs
need
year
chronic
evidence
high
report
acute
follow
international
care
consult
much
low
woman
base
disorder
rate
diagnose
diagnosis
unmet
medication
use
among
disability
migraine,
relate
health
include
participant
headache
psychological
study