```
library(twitteR)
library(tm)
library(wordcloud)
library(RColorBrewer)
library(ggplot2)
```

This is a report on text analysis of queries on Healthtap over the past two months.
We will look at the attributes *full_string* , *ip* and *timestamp* for this analysis.

```
search1<-read.csv("100Ksample_site_searches.csv")
search<-search1
search<-search[,which(names(search)%in%c('full_string','ip','timestamp'))]
```

The attribute "full_string" is the full search querry which should be processed.

```
a<-sub(".*?search_string=(.*?)&.*", "\\1", search$full_string)
c<- gsub("%2520|%2522|%252|%253|%2526|%20|%40|%92|%22|%27|%2|%3|%5|%D1%85%D1%86%D0%B3"," ",a
search$full_string<-c
```

Convert timestamp to to epoch time and create time intervals every "n" minutes

```
nmin<-4
search$timestamp<- unlist(strsplit(gsub("T"," ",search$timestamp),".",fixed=TRUE))[2*(1:leng
search$timestamp<-as.POSIXct(strptime(search$timestamp, "%Y-%m-%d %H:%M:%S"))
#sum(is.na(search$timestamp))
search$timeint<-unclass(search$timestamp)/(nmin*60)
```

We combine the records from same time interval and same ip address.

```
search$identifier<-apply(search,1,function(x) paste(x[2],x[4],sep=" ; "))
combine<-aggregate(full_string~identifier,search,function(x) paste(x, collapse = " ; ")) # c
combine$fs <- sapply(combine$full_string,function(x) paste(unique(strsplit(x," ")[[1]]),coll
combine$fs <- gsub("pregnant","pregnancy",combine$fs)
```

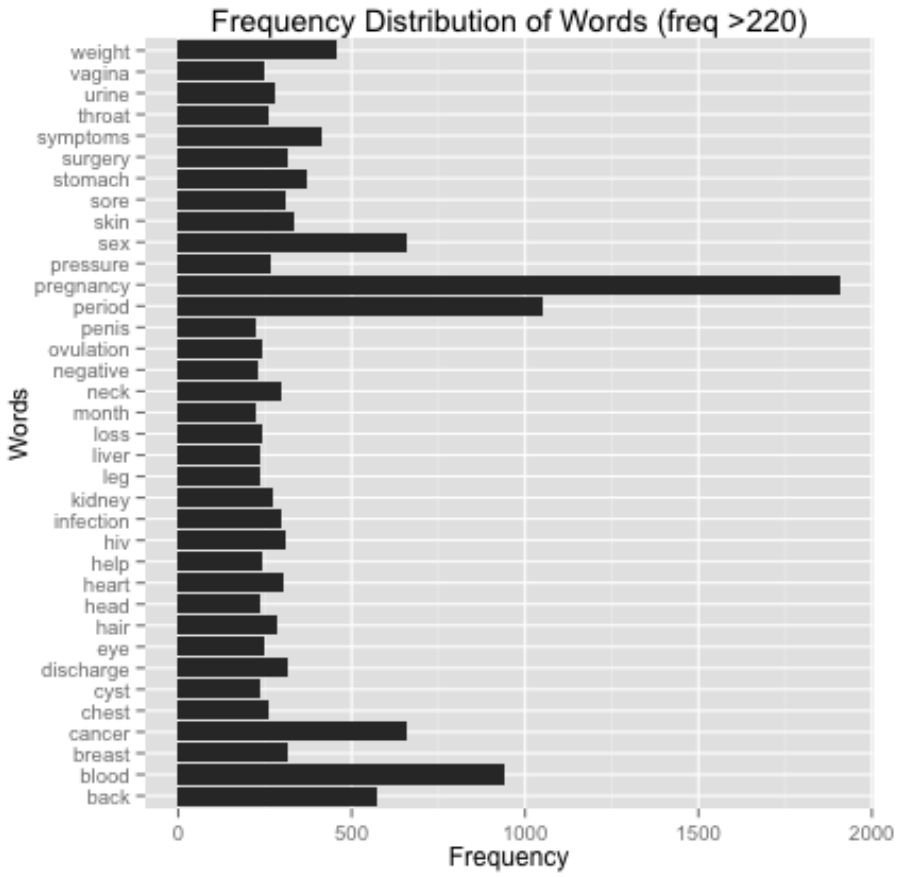
The number of records reduces from 100000 to 37455. Now we preprocess each
querry and find the most frequent words.

```
a<-Corpus(DataframeSource(data.frame(combine$fs)))
a <- tm_map(a, content_transformer(tolower))
a <- tm_map(a, removePunctuation)
mystopwords<- c('high','normal','lower','test','day','weeks','time','right','left','periods'
a <- tm_map(a, function(x) removeWords(x, c(stopwords("english"),mystopwords)))
corpus<- a
```

```
tdm <- TermDocumentMatrix(corpus)
m <- as.matrix(tdm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
d1<-d
d<-subset(d,d$freq>220)
n<- 25
bard<-d[1:n,]
```

The following figure shows the word distribution with frequency greater than 200,

```
qplot(x=d$word,y=as.numeric(d$freq),geom="bar",stat="identity")+coord_flip()+xlab("Words")+y
```



and the top 15 frequently appearing terms are as follows:

```
ylim<- c(0,1.1*max(bard$freq))
xx<-barplot(bard$freq,xaxt='n',xlab='',width=0.85,ylim=ylim,main="Top 25 Frequently Appearin
text(x=xx,y=bard$freq,label=bard$freq,pos=3,cex=0.6,col="red")
axis(1,at=xx,labels=bard$word,tick=FALSE,las=2,line=-0.5,cex.axis=0.9)
```