setwd("/Users/yiminwong/Project/KGC")  
getwd()

## [1] "/Users/yiminwong/Project/KGC"

data = read.csv('user\_interests.csv')  
str(data)

## 'data.frame': 178 obs. of 3 variables:  
## $ user\_id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ user : Factor w/ 176 levels "","ABHIK SEAL",..: 140 85 86 2 125 150 56 55 121 162 ...  
## $ interest: Factor w/ 171 levels "","  financial microdata and KG, Ontotext, reconciliation, false claims, updating KG, Lexus/Nexus, and Bloomberg l"| \_\_truncated\_\_,..: 166 160 166 166 162 159 164 166 163 170 ...

library(ggplot2);   
library(ggthemes);  
library(dplyr);

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

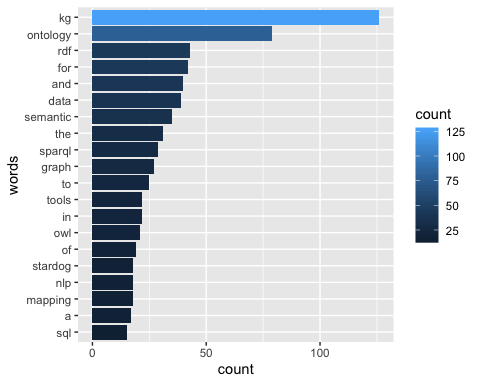
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidytext)

data$interest <- as.character(data$interest)

data%>%  
 unnest\_tokens(input = interest, output = word)%>%  
 select(word)%>%  
 group\_by(word)%>%  
 summarize(count = n())%>%  
 ungroup()%>%  
 arrange(desc(count))%>%  
 top\_n(20)%>%  
 ggplot(aes(x=reorder(word,count), y=count, fill=count))+  
 geom\_col()+  
 xlab('words')+  
 coord\_flip()

## Selecting by count



data%>%  
 unnest\_tokens(input = interest, output = word)%>%  
 select(word)%>%  
 anti\_join(stop\_words)%>%  
 group\_by(word)%>%  
 summarize(count = n())%>%  
 ungroup()%>%  
 arrange(desc(count))%>%  
 top\_n(25)

## Joining, by = "word"

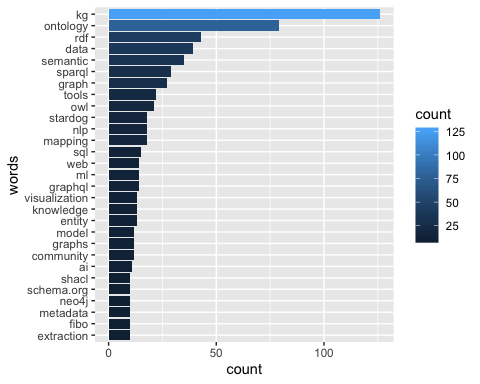
## Selecting by count

## # A tibble: 29 x 2  
## word count  
## <chr> <int>  
## 1 kg 126  
## 2 ontology 79  
## 3 rdf 43  
## 4 data 39  
## 5 semantic 35  
## 6 sparql 29  
## 7 graph 27  
## 8 tools 22  
## 9 owl 21  
## 10 mapping 18  
## # … with 19 more rows

data%>%  
 unnest\_tokens(input = interest, output = word)%>%  
 select(word)%>%  
 anti\_join(stop\_words)%>%  
 group\_by(word)%>%  
 summarize(count = n())%>%  
 ungroup()%>%  
 arrange(desc(count))%>%  
 top\_n(25)%>%  
 ggplot(aes(x=reorder(word,count), y=count, fill=count))+  
 geom\_col()+  
 xlab('words')+  
 coord\_flip()

## Joining, by = "word"

## Selecting by count



library(wordcloud)

## Loading required package: RColorBrewer

wordcloudData =   
data %>%  
 group\_by(user\_id) %>%  
 unnest\_tokens(output=word,input=interest)%>%  
 anti\_join(stop\_words)%>%  
 group\_by(word)%>%  
 summarize(freq = n())%>%  
 arrange(desc(freq))%>%  
 ungroup()%>%  
 data.frame()

## Joining, by = "word"

library(wordcloud)  
set.seed(617)  
wordcloud(words = wordcloudData$word,wordcloudData$freq,scale=c(2,0.5),max.words = 100,colors=brewer.pal(9,"Spectral"))

