Project 2:

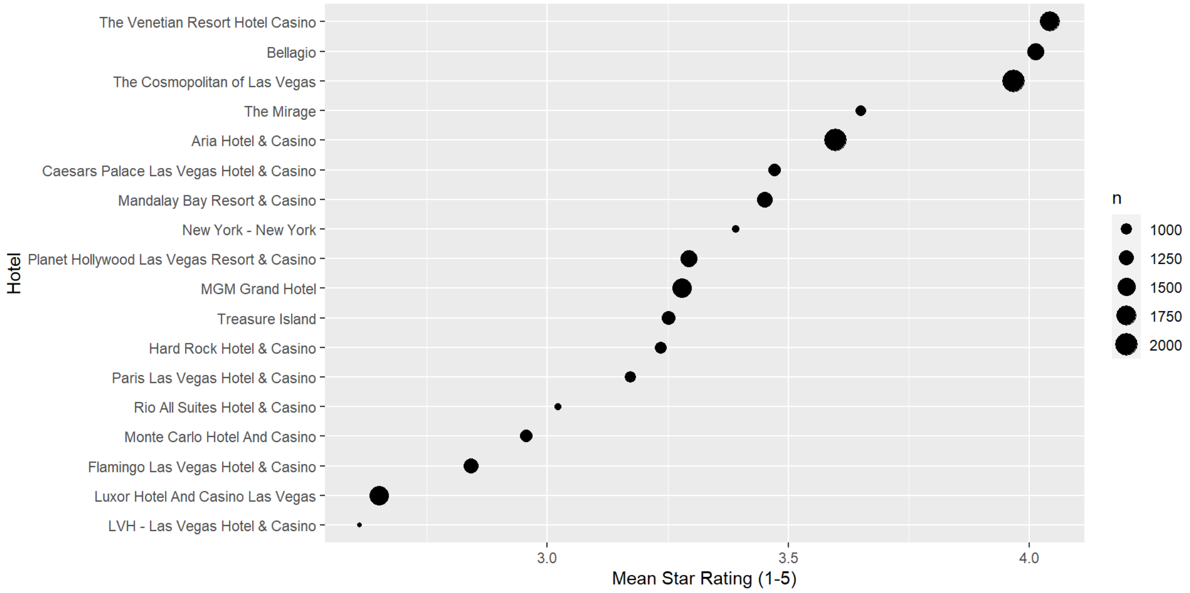
Yelp reviews of Las Vegas hotels (Data collected from Yelp: Courtesy of Prof Karsten Hansen)

```
reviews <- read_rds('data/vegas_hotel_reviews.rds')
business <- read_rds('data/vegas_hotels_info.rds')
```

This data contains customer reviews of 18 hotels in Las Vegas. In addition to text, each review also contains a star rating from 1 to 5.

Checking mean star rating:

```
library(tidyverse)
library(scales)
library(lubridate)

reviews %>%
  left_join(select(business,business_id,name),
             by='business_id') %>%
  group_by(name) %>%
  summarize(n = n(),
            mean.star = mean(as.numeric(stars))) %>%
  arrange(desc(mean.star)) %>%
  ggplot() +
  geom_point(aes(x=reorder(name,mean.star),y=mean.star,size=n))+
  coord_flip() +
  ylab('Mean Star Rating (1-5)') +
  xlab('Hotel')
```




```
## install packages
install.packages(c("wordcloud","tidytext"))  ## only run once
library(tidytext)
library(wordcloud)
```

Part 1:

We are interested in summarizing the reviews for the Aria hotel:

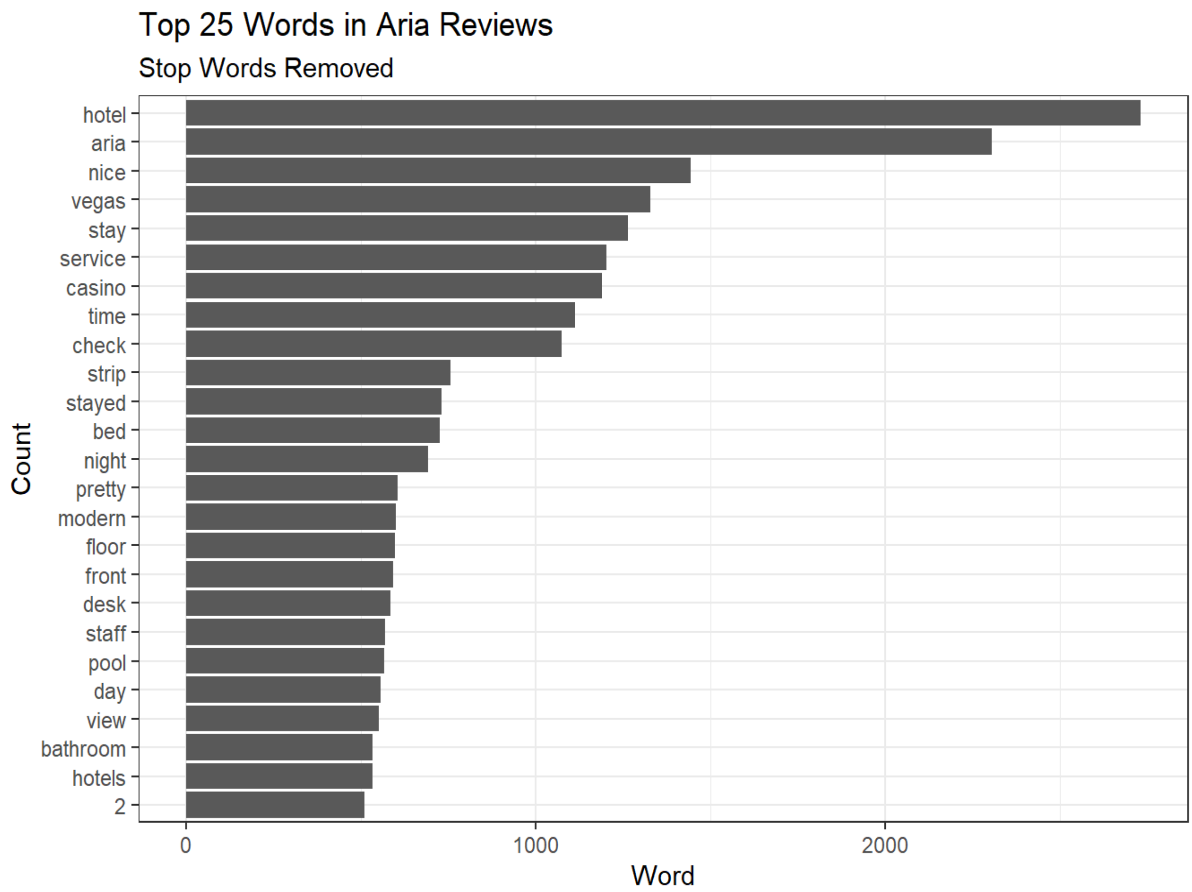
```
aria.id <-  filter(business,
        name=='Aria Hotel & Casino')$business_id
aria.reviews <- filter(reviews,
                       business_id==aria.id)


AriaTidy <- aria.reviews %>%
  select(review_id,text,stars) %>%
  unnest_tokens(word,text)

AriaFreqWords <- AriaTidy %>%
  count(word)
```

Plotting top words:

```
AriaFreqWords %>%
  top_n(25) %>%
  ggplot(aes(x=fct_reorder(word,n),y=n)) + geom_bar(stat='identity') +
  coord_flip() + theme_bw()+
  labs(title='Top 25 Words in Aria Reviews',
       x='Count',
       y= 'Word')
AriaFreqWords %>%
  top_n(25) %>%
  ggplot(aes(x=fct_reorder(word,n),y=n)) + geom_bar(stat='identity') +
  coord_flip() + theme_bw()+
  labs(title='Top 25 Words in Aria Reviews',
       x='Count',
       y= 'Word')
```

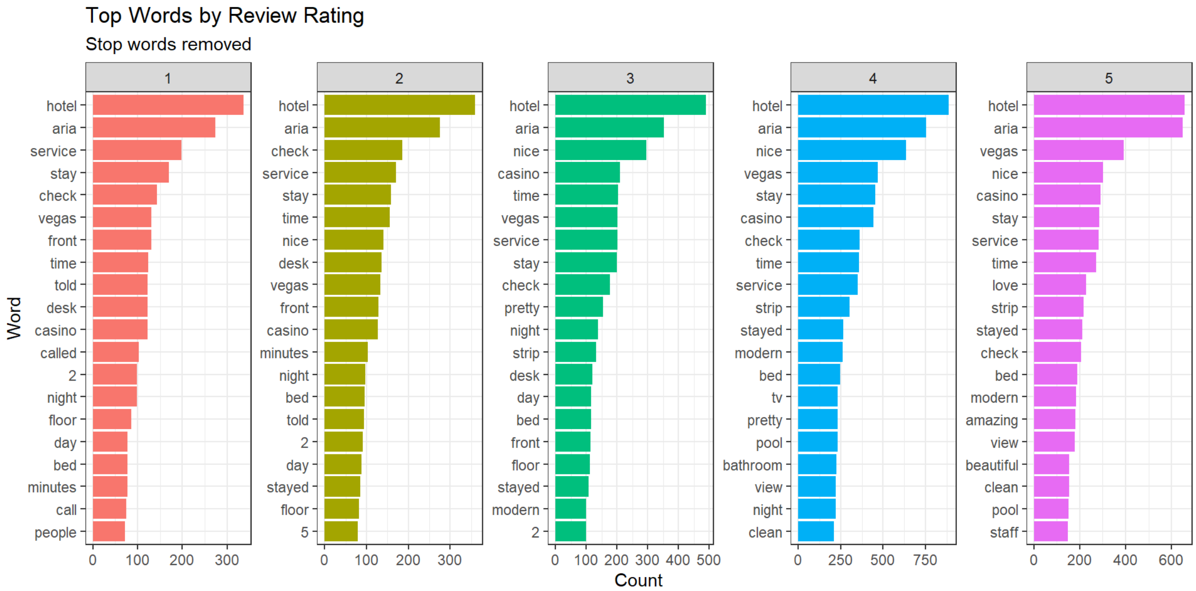
Problem with stop words. Let’s remove them:

```
AriaFreqWords %>%
  anti_join(stop_words) %>%
  top_n(25) %>%
  ggplot(aes(x=fct_reorder(word,n),y=n)) + geom_bar(stat='identity') +
  coord_flip() + theme_bw()+
  labs(title='Top 25 Words in Aria Reviews',
       subtitle = 'Stop Words Removed',
       x='Count',
       y= 'Word')
```



Top words vary with the rating of the underlying reviews:

```
AriaFreqWordsByRating <- AriaTidy %>%
  count(stars,word)
AriaFreqWordsByRating %>%
  anti_join(stop_words) %>%
  group_by(stars) %>%
  top_n(20) %>%
  ggplot(aes(x=reorder_within(word,n,stars),
             y=n,
             fill=stars)) +
  geom_bar(stat='identity') +
  coord_flip() +
  scale_x_reordered() +
  facet_wrap(~stars,scales = 'free',nrow=1) +
  theme_bw() +
  theme(legend.position = "none")+
  labs(title = 'Top Words by Review Rating',
       subtitle = 'Stop words removed',
       x = 'Word',
       y = 'Count')
```



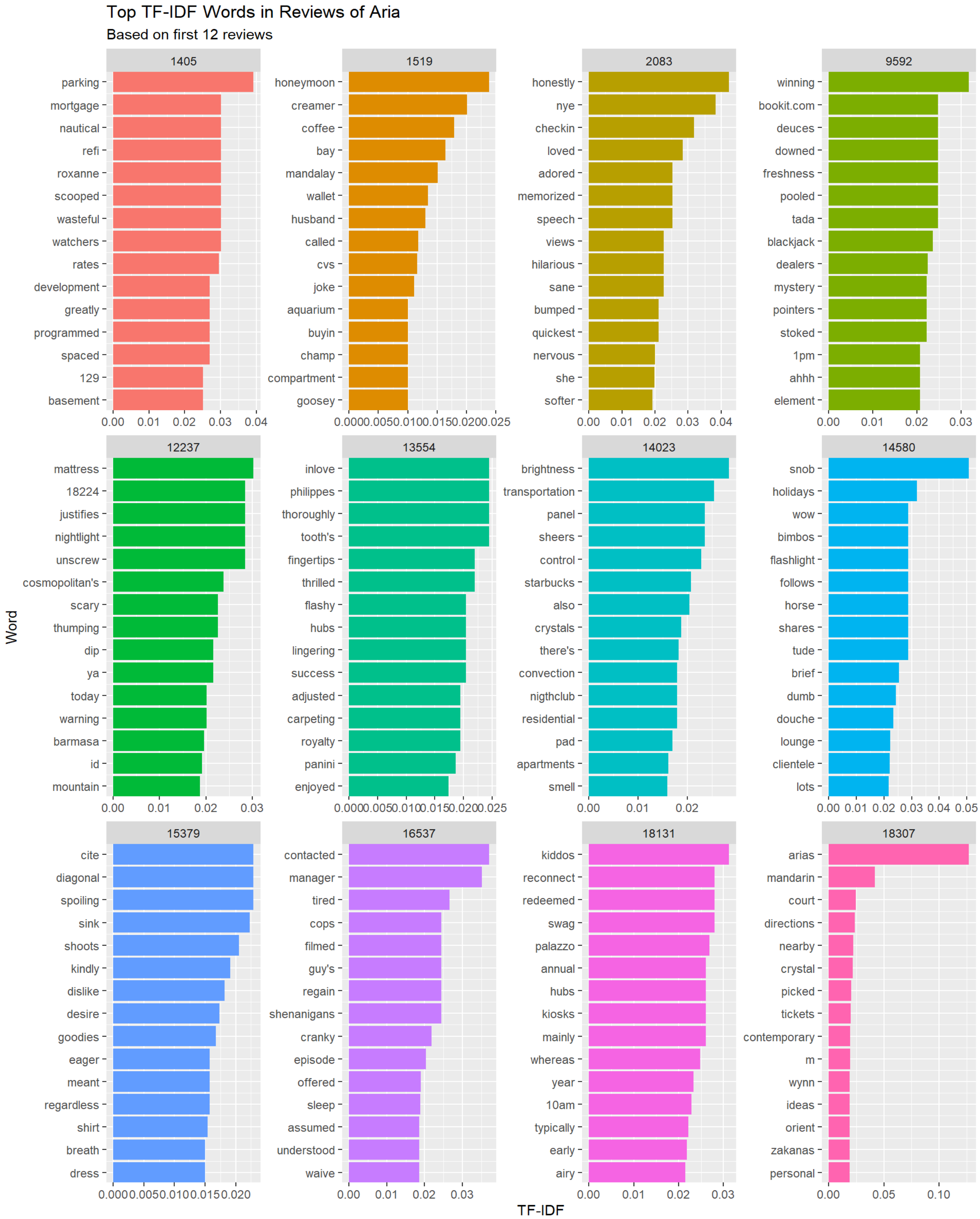
Words with high TF-IDF in a document will tend to be words that are rare (when compared to other documents) but not too rare. In this way they are informative about what the document is about! We can easily calculate TF-IDF.

```
tidyReviews <- aria.reviews %>%
  select(review_id,text) %>%
  unnest_tokens(word, text) %>%
  count(review_id,word)

minLength <- 200  # focus on long reviews
tidyReviewsLong <- tidyReviews %>%
  group_by(review_id) %>%
  summarize(length = sum(n)) %>%
  filter(length >= minLength)

tidyReviewsTFIDF <- tidyReviews %>%
  filter(review_id %in% tidyReviewsLong$review_id) %>%
  bind_tf_idf(word,review_id,n) %>%
  group_by(review_id) %>%
  arrange(desc(tf_idf)) %>%
  slice(1:15) %>% # get top 15 words in terms of tf-idf
  ungroup() %>%
  mutate(xOrder=n():1) %>%  # for plotting
  inner_join(select(aria.reviews,review_id,stars),by='review_id')  # get st
ar ratings
```

Here is a plot of the top TF-IDF words for 12 reviews:

```
nReviewPlot <- 12
plot.df <- tidyReviewsTFIDF %>%
  filter(review_id %in% tidyReviewsLong$review_id[1:nReviewPlot])

plot.df %>%
  mutate(review_id_n = as.integer(review_id)) %>%
  ggplot(aes(x=xOrder,y=tf_idf,fill=factor(review_id_n))) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  facet_wrap(~ review_id_n,scales='free') +
  scale_x_continuous(breaks = plot.df$xOrder,
                     labels = plot.df$word,
```

```
                      expand = c(0,0)) +
  coord_flip()+
  labs(x='Word',
       y='TF-IDF',
       title = 'Top TF-IDF Words in Reviews of Aria',
       subtitle = paste0('Based on first ',
                         nReviewPlot,
                         ' reviews'))+
  theme(legend.position = "none")
```



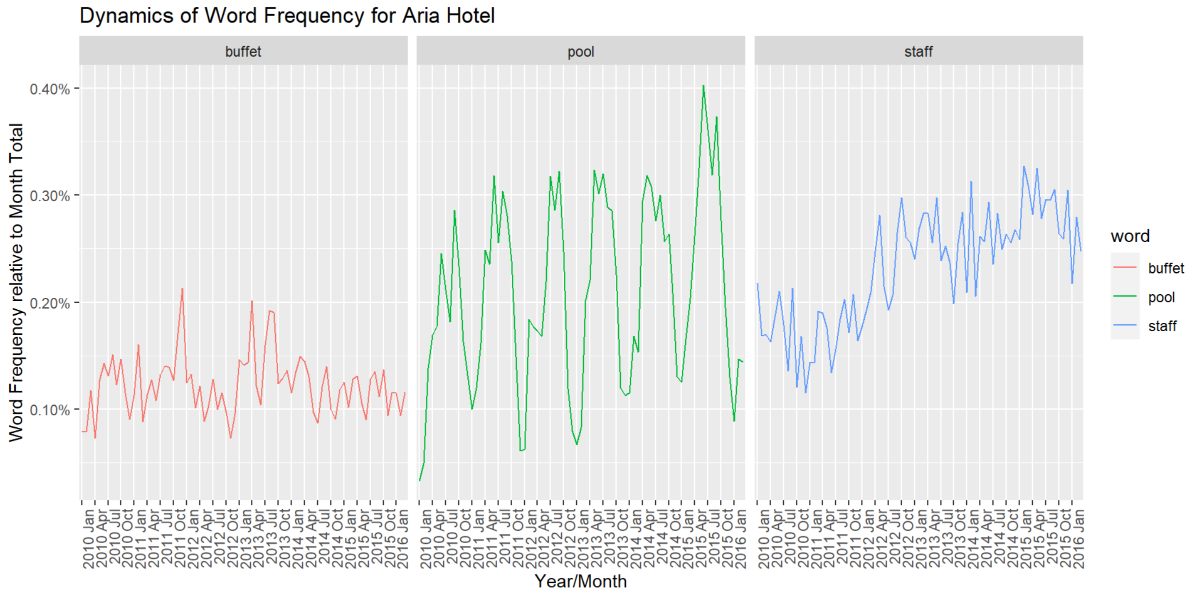
These can be used as keywords for each review.

Part 2:

How do word frequencies change over time(still using Aria data)?

We calculate the relative frequency of these three terms["buffet","pool" and "staff"] for each month (relative to the total number of terms used that month).

```
ariaTidy <- aria %>%
  select(reviewID,text) %>%
  unnest_tokens(word,text) %>%
  count(reviewID,word) %>%
  inner_join(meta.data,by="reviewID")

total.terms.time <- ariaTidy %>%
  group_by(year.month.group) %>%
  summarize(n.total=sum(n))

## for the legend
a <- 1:nrow(total.terms.time)
b <- a[seq(1, length(a), 3)]

ariaTidy %>%
  filter(word %in% c("pool","staff","buffet")) %>%
  group_by(word,year.month.group) %>%
  summarize(n = sum(n)) %>%
  left_join(total.terms.time, by='year.month.group') %>%
  ggplot(aes(x=year.month.group,y=n/n.total,color=word,group=word)) +
  geom_line() +
  facet_wrap(~word)+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  scale_x_discrete(breaks=as.character(total.terms.time$year.month.group[b]
))+
  scale_y_continuous(labels=percent)+xlab('Year/Month')+
  ylab('Word Frequency relative to Month Total')+
  ggtitle('Dynamics of Word Frequency for Aria Hotel')
```



We see three different patterns for the relative frequencies: "buffet" is used in a fairly stable manner over this time period, while "pool" displays clear seasonality, rising in popularity in the summer months. Finally, we see an upward trend in the use of "staff".

We can try a similar analysis where consider word frequency dynamics for different satisfaction segments:

```
aria.tidy2 <- ariaTidy %>%
  mutate(year = year(date),
         satisfaction = fct_recode(factor(reviewRating),
                                   "Not Satisfied"="1",
                                   "Not Satisfied"="2",
                                   "Neutral"="3",
                                   "Neutral"="4",
                                   "Satisfied"="5"))

total.terms.rating.year <- aria.tidy2 %>%
  group_by(satisfaction,year) %>%
  summarize(n.total = sum(n))


aria.tidy2 %>%
  filter(word %in% c("pool","staff","buffet","food","wait","casino","line",
"check","clean")) %>%
  group_by(satisfaction,year,word) %>%
  summarize(n = sum(n)) %>%
  left_join(total.terms.rating.year, by=c('year','satisfaction')) %>%
  ggplot(aes(x=year,y=n/n.total,color=satisfaction,group=satisfaction)) +
  geom_line(size=1,alpha=0.25) + geom_point() +
  facet_wrap(~word,scales='free')+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  scale_y_continuous(labels=percent)+xlab('Year')+
  ylab('Word Frequency relative to Month Total')+
  labs(title='Dynamics of Word Frequency for Aria Hotel',
       subtitle='Three Satisfaction Segments')
```