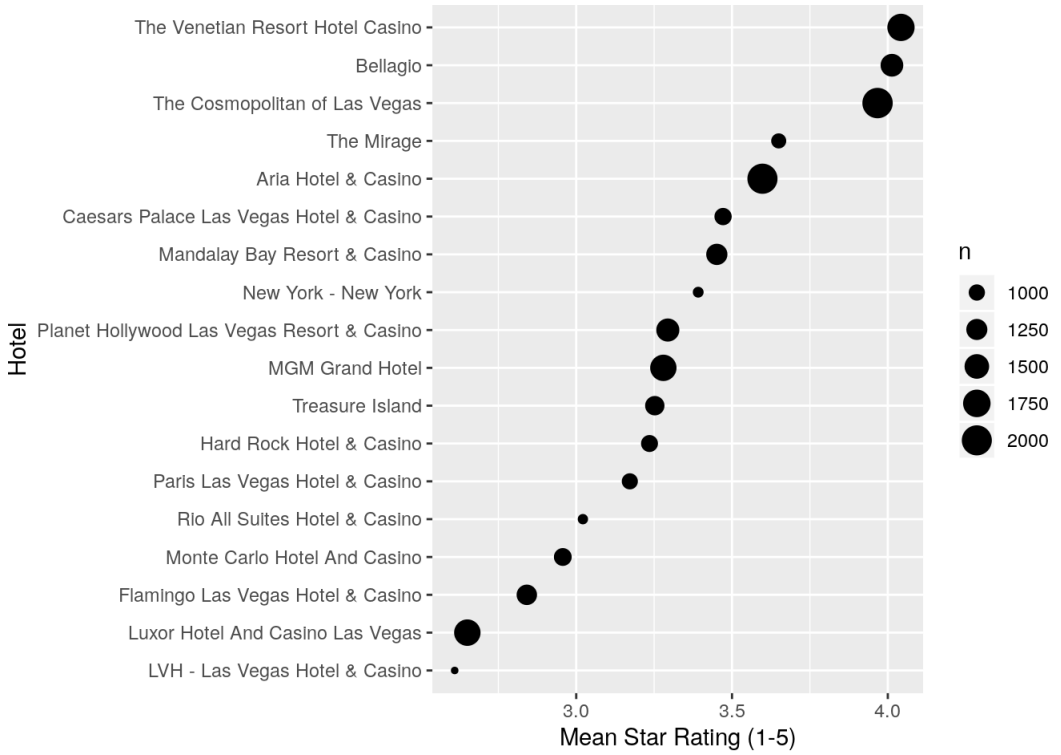
```
library(tidyverse)
library(scales)
library(forcats)
library(wordcloud)
library(tidytext)
library(lubridate)
```

```
## load review data for las vegas hotels
load('data/vegas_hotels.rda')
```

Find the most popular hotel in Las Vegas(most people go there and give high rating)

```
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')
```

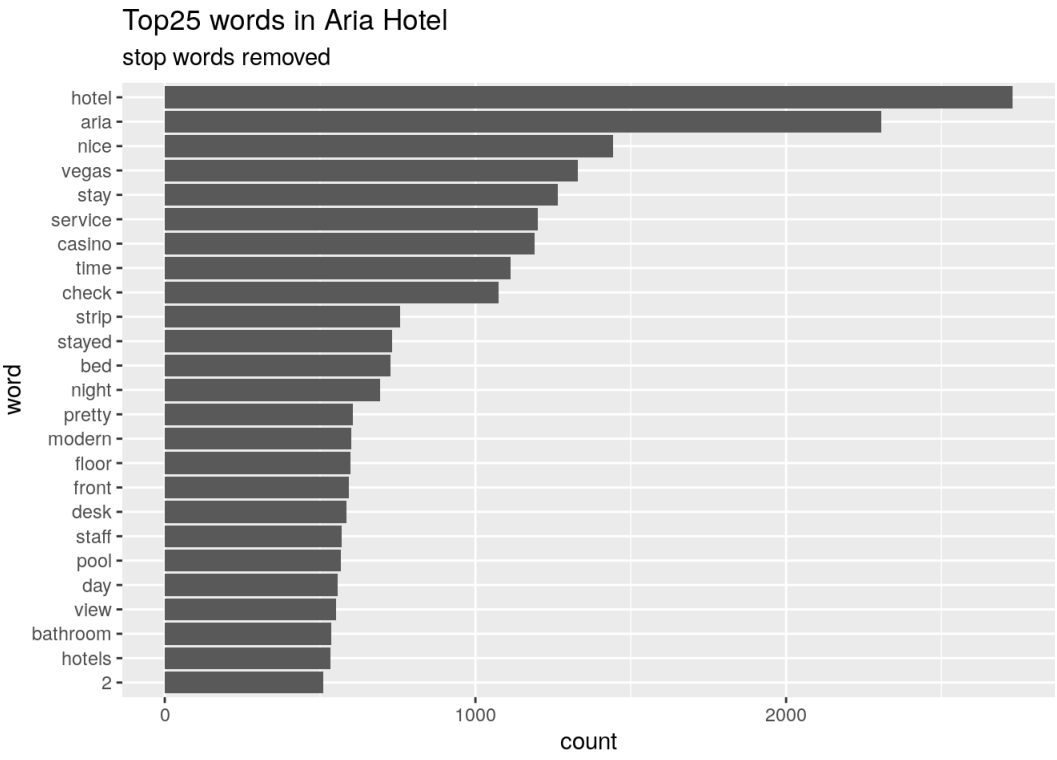


So The Venetian, Bellagio and The Cosmopolitan are clearly the highest rated hotels, while Luxor and LVH are the lowest rated. Ok, but what is behind these ratings? What are customers actually saying about these hotels? This is what we can hope to find through a text analysis.

Count words about one hotel(Aria):

```
## count each word in a document(sentences):split,count,antijoin
#exampleTidyNoStop= example%>%
  #unnest_tokens(word,text)%>%
  #count(doc_id,word) %>%
  #anti_join(stop_words,by='word')


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

## doc-term matrix - tidy

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

AriaFreqWords = AriaTidy%>%
  count(word)%>%
  anti_join(stop_words,by='word')

AriaFreqWords%>%
  top_n(25) %>%
  ggplot(aes(x=fct_reorder(factor(word),n),y=n))+geom_bar(stat='identity')+coord_flip()+
  labs(x='word',y="count",title='Top25 words in Aria Hotel',subtitle='stop words removed')
```

```
## Selecting by n
```



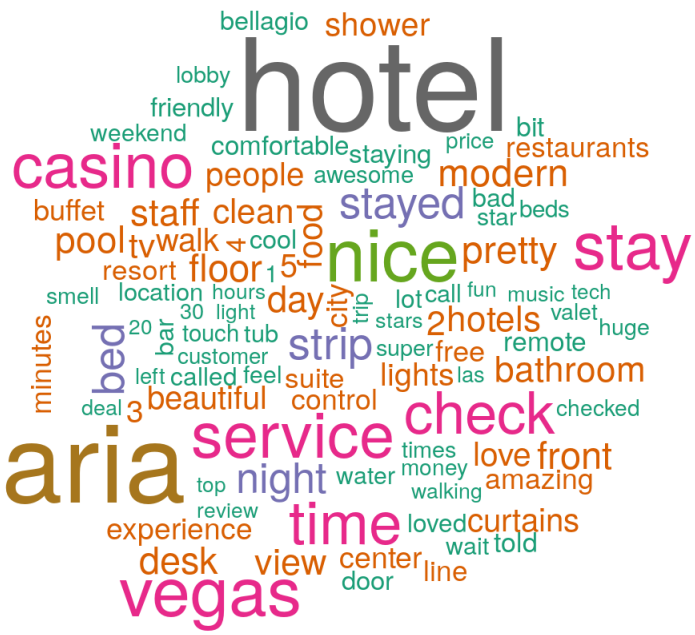
Word clouds:

```
#- visualizing a dtm

topWords <- AriaFreqWords %>%
  anti_join(stop_words) %>%
  top_n(100)
```

```
## Joining, by = "word"
```

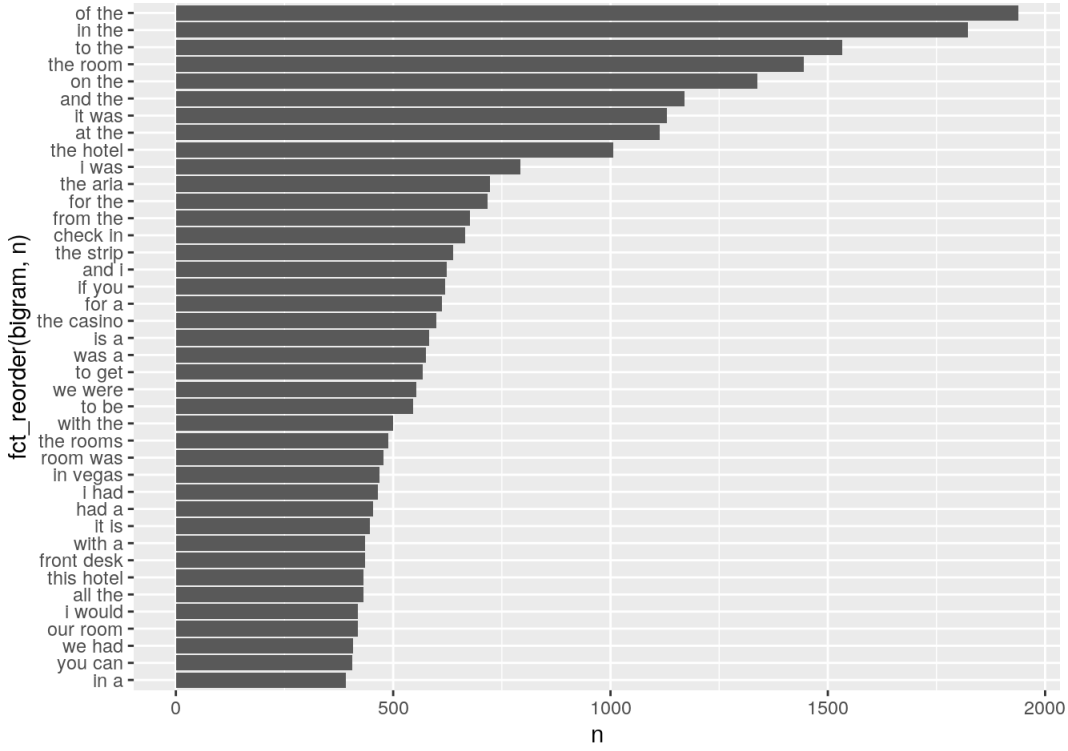
```
## Selecting by n
```

```
plot=wordcloud(topWords$word,
           topWords$n,
           scale=c(5,0.5),
           colors=brewer.pal(8,"Dark2"))
```




```
## repeat with bi-grams

aria.reviews %>%
  select(review_id,text) %>%
  unnest_tokens(bigram,text,token="ngrams",n=2) %>%
  count(bigram) %>%
  top_n(40) %>%
  ggplot(aes(x=fct_reorder(bigram,n),y=n)) + geom_bar(stat='identity') +
  coord_flip()
```

```
## Selecting by n
```



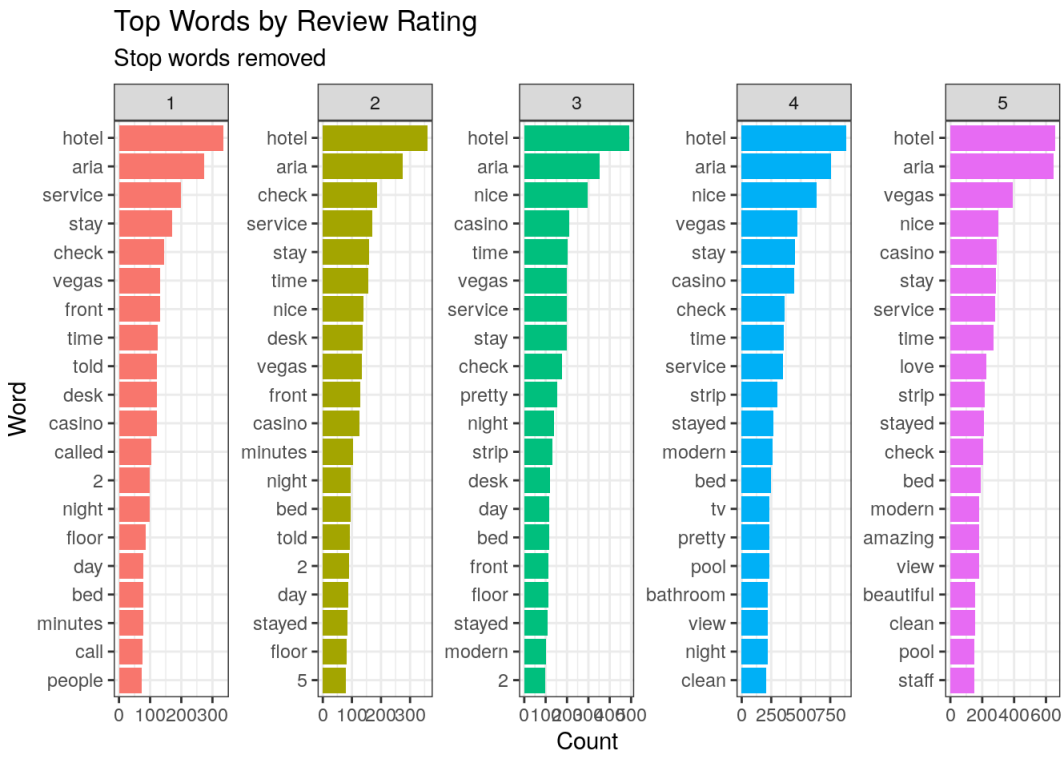
Clean pretty and buffet are the features that attract people. But some words like “bad” should also arouse the attention of the hotel.

Next, we see what people say in different ratings.

Count words in each star rating:

```
## Top words by rating

AriaFreqWordsByRating <- AriaTidy %>%
  count(stars,word)

## for plotting (from https://github.com/dgrtwo/drlib/blob/master/R/reorder_within.R)
##Reorder a column before plotting with faceting, such that the values are ordered within each facet.

reorder_within <- function(x, by, within, fun = mean, sep = "___", ...) {
  new_x <- paste(x, within, sep = sep)
  stats::reorder(new_x, by, FUN = fun)
}

scale_x_reordered <- function(..., sep = "___") {
  reg <- paste0(sep, ".+$")
  ggplot2::scale_x_discrete(labels = function(x) gsub(reg, "", x), ...)
}

#compare with the code that did't use the function. The y label are blanks because words and scales are diff
erent.
#bystar=AriaFreqWordsByRating %>%anti_join(stop_words,by='word')%>% group_by(stars) %>% top_n(20) %>% arrang
e(stars,desc(n))
#bystar %>% ggplot(aes(x=fct_reorder(factor(word),n),y=n))+facet_wrap(~stars)+geom_bar(stat='identity')+coor
d_flip()

AriaFreqWordsByRating %>%
  anti_join(stop_words,by='word') %>%
  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')
```

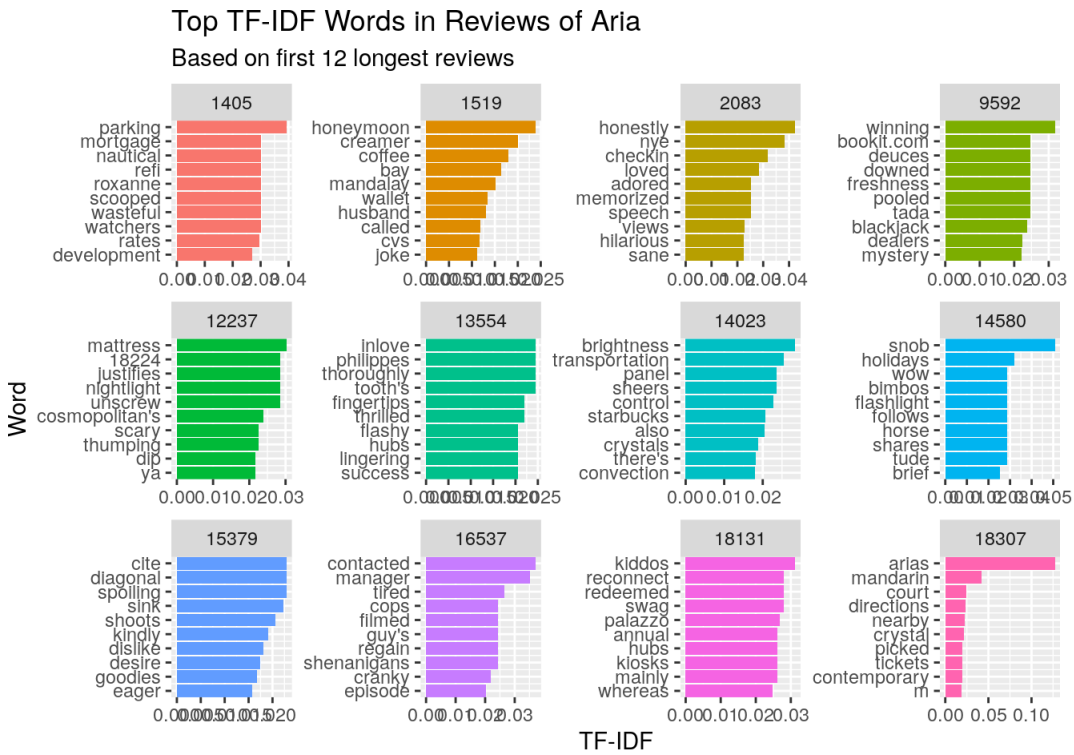
```
## Selecting by n
```



Obviously, check,service,font,minutes… These keys words appear most frequently in the low rating views. The customers were likely to complain about the wait time when checkin/checkout. This give the hotel some idea that it should improve its efficiency. Otherwise, it will worsen its rating on the website.

Pick top12 longest reviews. What's the feature of them?

```
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) %>% #word in each review 's n
  group_by(review_id) %>%
  arrange(desc(tf_idf)) %>%
  slice(1:10) %>% # get top 10 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 star ratings


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,
                         ' longest reviews'))+
  theme(legend.position = "none")
```

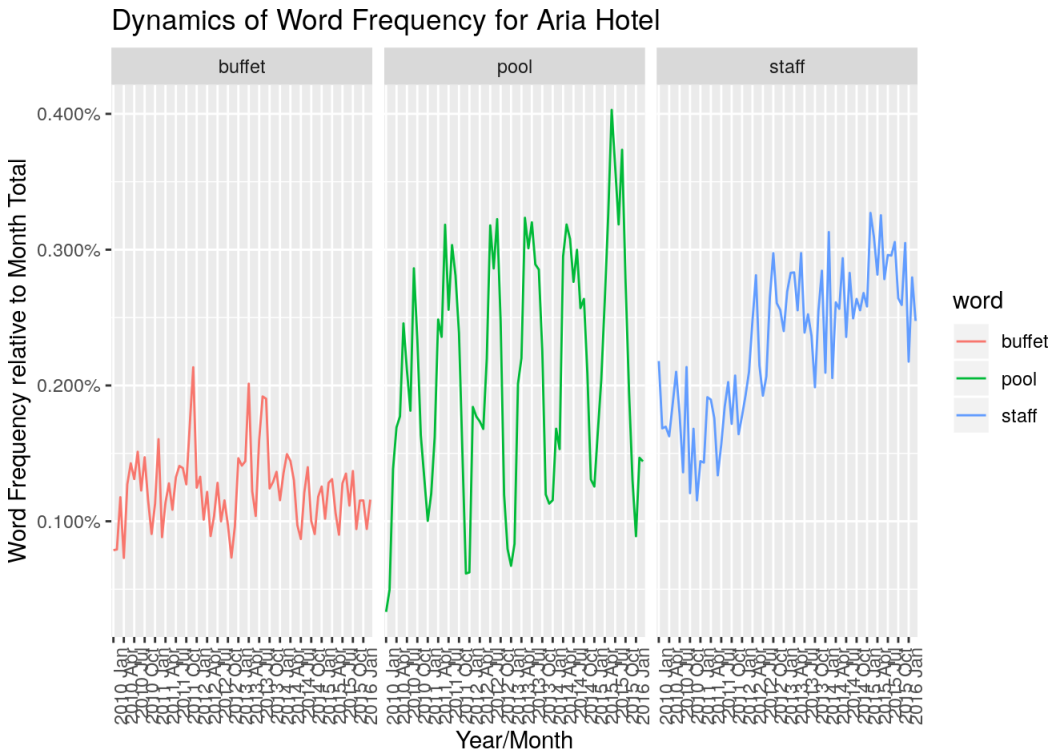


Now we get the keywords of longest views. We can find problems such as the flashlight and the staff. While

honeymoon,night club can be the words that the hotel could use more often to attract people.

Word frequencey change with time. This gives insights about the seasonality of branding.

```
## Aria on Tripadvisor reviews
aria <- read_rds('data/AriaReviewsTrip.rds') %>%
  rename(text = reviewText)

meta.data <- aria %>%
  select(reviewID,reviewRating,date,year.month.group)

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

## word frequency over time

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)]


words_want_know=c("pool","staff","buffet")
ariaTidy %>%
  filter(word %in% words_want_know) %>%
  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”.

Let’s see the trend for different key words. What do people care about now?

```
## same but 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))


words_want_know=c("pool","staff","buffet","food","wait","casino","line","check","clean")

t=aria.tidy2 %>%
  filter(word %in% words_want_know) %>%
  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))+xlab('Year')+
  scale_y_continuous(labels=percent)+
  ylab('Word Frequency relative to Month Total')+
  labs(title='Dynamics of Word Frequency for Aria Hotel',
       subtitle='Three Satisfaction Segments')
```

Buffet and casino appeared less frequently in satisfied comments. The hotel should consider branding its buffet or casino more effectively or have sth. new to attract customers. Obviously, the wait time in the check-in font is reduced over the years, which is good and proved that improving the service will be helpful to increase the hotel’s rating.