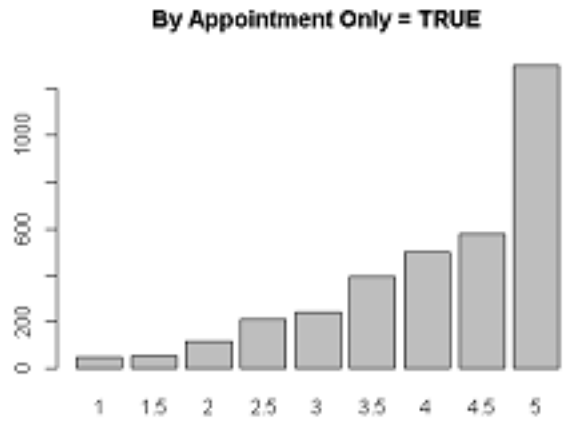
Analysis of Appointment Only Businesses in Yelp Dataset

Nick Lusk

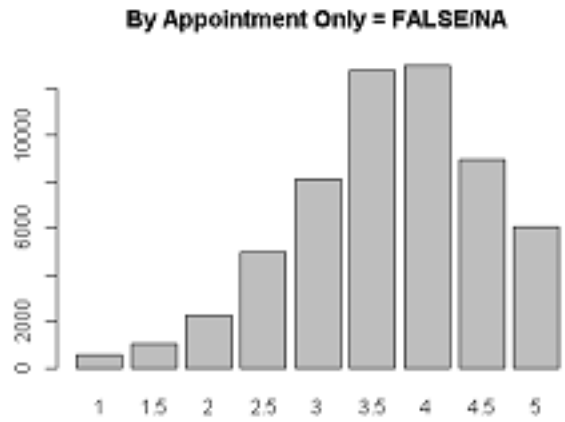
November 22, 2015

Introduction

Within the dataset provided for the Yelp Dataset Challenge a vast variety of business types can be found. During my initial exploration of the data I found that businesses operating as appointment only skewed very high in customer ratings.



In comparison, the distribution of ratings overall has a visibly wider distribution that peaks in the 3.5/4.0 range.



It is clear that a majority of these businesses have an average rating of five stars. Out of the 3440 businesses in this group 38% have a five star rating, which is 224% more than the next highest rating, which

is four-and-a-half. The full breakdown of percentages and totals compared to star rating can be seen in the table below.

Appointment Only Businesses		
Stars Avg.	Number	% of Total
1	47	01.36627906976744
1.5	53	01.5406976744186
2	115	03.34302325581395
2.5	209	06.07558139534884
3	242	07.03488372093023
3.5	394	11.45348837209302
4	500	14.53488372093023
4.5	580	16.86046511627907
5	1300	37.7906976744186

Can correlations be found among the appointment only businesses that have five star ratings which could be applied to other businesses to increase their chance of being more highly rated? Conversely, can patterns be established among the lower rated appointment only businesses which other businesses could avoid to increase their chance of being more highly rated?

Methods and Data

For this analysis I am going to use text mining techniques to identify common terms in both the reviews and tips json files provided in the dataset. They will be further broken down by star average. Additionally, I will identify the most common three word combinations (trigrams) in the reviews data. These will be broken down in the same way as the single word terms.

Below is a general breakdown of the R code that was used to generate these terms. Though it is incomplete the method of data processing can easily be extracted from what is included. For those who want to look at the entire code it can be found on GitHub here: appt-only-capstone.r

```
## read-in business json file
install.packages("jsonlite")
library("jsonlite")
busJsonOriginal <- stream_in(file("yelp_academic_dataset_business.json"))
## flatten business json data and convert categories to string
install.packages("dplyr")
library(dplyr)
busJsonOriginal.flat <- flatten(busJsonOriginal)
busJsonOriginal.cat <- mutate(busJsonOriginal.flat, cat2=toString(categories[[1]]))
## filter out only businesses that are appointment only and create barplot by stars
appointmentBIZ <- filter(busJsonOriginal.cat, `attributes.By Appointment Only` == TRUE)
stars.BIZ.appt.TRUE <- table(appointmentBIZ$stars)
barplot(stars.BIZ.appt.TRUE, main="By Appointment Only = TRUE")
## convert NA to false in appointment only category and filter all other businesses
## into seperate dataset and create barplot by stars
BIZ.appt.na.to.false <- busJsonOriginal.cat
BIZ.appt.na.to.false$`attributes.By Appointment Only`[is.na
    (BIZ.appt.na.to.false$`attributes.By Appointment Only`)] <- 'FALSE'
appt.BIZ.false <- filter(BIZ.appt.na.to.false, `attributes.By Appointment Only` != TRUE)
stars.BIZ.appt.FALSE <- table(appt.BIZ.false$stars)
barplot(stars.BIZ.appt.FALSE, main="By Appointment Only = FALSE/NA")
```

```
## create a subset of review dataset containing only reviews of appointment
## only businesses
apptBusinessIDs <- appointmentBIZ$business_id
revJsonOriginal <- stream_in(file("yelp_academic_dataset_review.json"))
apptReviews <- subset(revJsonOriginal, business_id %in% apptBusinessIDs)
apptReviewsText <- apptReviews$text
## begin data transformation and mining of appointment only review text
install.packages("tm")
install.packages("SnowballC")
library(tm)
library(SnowballC)
apptReviewsTextCorpus <- Corpus(VectorSource(apptReviewsText))
apptReviewsTextCorpus <- tm_map(apptReviewsTextCorpus, removePunctuation)
apptReviewsTextCorpus <- tm_map(apptReviewsTextCorpus, removeNumbers)
apptReviewsTextCorpus <- tm_map(apptReviewsTextCorpus, content_transformer(tolower))
apptReviewsTextCorpus <- tm_map(apptReviewsTextCorpus, removeWords, stopwords("english"))
apptReviewsTextCorpus <- tm_map(apptReviewsTextCorpus, stemDocument, language = "english")
apptReviewsTextCorpus <- tm_map(apptReviewsTextCorpus, stripWhitespace)
apptReviewsTextCorpus <- tm_map(apptReviewsTextCorpus, PlainTextDocument)
## Create document term matrix, sort frequent terms and write output to csv file
dtmReviews <- DocumentTermMatrix(apptReviewsTextCorpus)
freqReviewTerms <- sort(colSums(as.matrix(dtmReviews)), decreasing=TRUE)
write.csv(freqReviewTerms, file = "freqReviewTerms.csv")
## repeat freq terms process for tips data
## create subsets of all appointment only businesses by star rating
appointmentBIZoneStar <- appointmentBIZ[appointmentBIZ$stars == 1.0,]
appointmentBIZoneHalfStar <- appointmentBIZ[appointmentBIZ$stars == 1.5,]
appointmentBIZtwoStar <- appointmentBIZ[appointmentBIZ$stars == 2.0,]
appointmentBIZtwoHalfStar <- appointmentBIZ[appointmentBIZ$stars == 2.5,]
appointmentBIZthreeStar <- appointmentBIZ[appointmentBIZ$stars == 3.0,]
appointmentBIZthreeHalfStar <- appointmentBIZ[appointmentBIZ$stars == 3.5,]
appointmentBIZfourStar <- appointmentBIZ[appointmentBIZ$stars == 4.0,]
appointmentBIZfourHalfStar <- appointmentBIZ[appointmentBIZ$stars == 4.5,]
appointmentBIZfiveStar <- appointmentBIZ[appointmentBIZ$stars == 5.0,]
## repeat frequent term process this time subsetting each result by star rating
## identify frequent trigrams of the review text subsets
install.packages("RWeka")
library(RWeka)
# 3 word ngrams
TrigramTokenizer <- function(x) NGramTokenizer(x, Weka_control(min = 3, max = 3))
## one star business reviews frequent trigrams
apptReviewsTextOneStar <- readRDS("apptReviewsTextOneStar.rds")
apptReviewsTextOneStar <- Corpus(VectorSource(apptReviewsTextFiveStar))
apptReviewsTextOneStar <- tm_map(apptReviewsTextOneStar, removePunctuation)
apptReviewsTextOneStar <- tm_map(apptReviewsTextOneStar, removeNumbers)
apptReviewsTextOneStar <- tm_map(apptReviewsTextOneStar, content_transformer(tolower))
apptReviewsTextOneStar <- tm_map(apptReviewsTextOneStar, removeWords, stopwords("english"))
apptReviewsTextOneStar <- tm_map(apptReviewsTextOneStar, stemDocument, language = "english")
apptReviewsTextOneStar <- tm_map(apptReviewsTextOneStar, stripWhitespace)
apptReviewsTextOneStar <- tm_map(apptReviewsTextOneStar, PlainTextDocument)
dtmRevTrigramsOneStar <- DocumentTermMatrix(apptReviewsTextOneStar, control =
    list(tokenize = TrigramTokenizer))
freqDtmRevTrigramsOneStar <- sort(colSums(as.matrix(dtmRevTrigramsOneStar)), decreasing=TRUE)
```

```
write.csv(freqDtmRevTrigramsOneStar, file = "freqDtmRevTrigramsOneStar.csv")
## repeat frequent trigram process for remaining star ratings
```

Results

I have included here a sampling of the most frequent terms and phrases.

Trigrams from the review dataset for star ratings ranging from 1.0-2.5:

> never go back, sit wait room, dont wast time, peopl wait room, will never go, minut past appoint, wast time money, absolut worst experi, mani peopl wait, month later get, never receiv call, past appoint time, poor custom servic, will never return, act like bother, horribl custom servic, call sever time, wast time money, front desk staff, past appoint time, long wait time, go somewher els, wait hour see, wait wait room, dont feel like, hour wait room, made feel like, custom servic skill, front offic staff, make feel comfort, minut wait room, take time listen, ever go back, worst experi ever, didnt feel like,

Trigrams from the review dataset for star ratings ranging from 3.5-5.0:

> go anywher els, make feel comfort, feel much better, made feel comfort, definit go back, will definit go, will definit back, high recommend offic, love love love, high recommend anyon, will never go, go anyon els, cant say enough, will go back, wait go back, never go anywher, cant wait go, definit come back, high recommend place, make feel like, will definit return, great custom servic, look forward next, recommend anyon look, say enough good, front desk staff, enough good thing, enough good thing, time make sure, take time explain, will definit come, worth everi penni, staff friend help, everi time go, first time went

Frequent Trigrams from Tips Data Overall:

> go anywher els, great custom servic, definit come back, definit go back, make feel comfort, dont wast time, go somewher els, made feel comfort, never go anywher, super friend staff, will never go, wont go anywher, appoint make sure, im still wait, ive come year, make appoint get, past appoint time, servic friend staff, sit wait room, staff super friend, time make sure, worth everi penni, alway amaz job, cant wait go, everi time come, excel custom servic, friend clean profession, front desk staff, gave great advic, great friend staff, great servic friend, great staff great, keep come back, make feel better, say enough good, staff high recommend, take time make, trust anyon els, wait go back, wait time long, will definit back, will definit come, alway great job, alway look amaz, alway make feel, amaz custom servic, arriv min earli, best high recommend, cant wait next

Frequent Terms from Tips Data Overall:

> great, best, love, get, hair, place, time, staff, amazing, massage, good, will, nice, friendly, service, spa, today, awesome, always, just, new, getting, back, like, see, day, can, make, ask, wait

Frequent Terms from Reviews Data Overall:

> time, great, get, hair, just, like, back, one, will, place, massage, really, staff, room, appointment, good, also, can, spa, office, even, ive, first, going, nice, never, best, experience, got, day

Frequent Terms from Reviews with 1.0 Star Rating:

time, appointment, office, never, called, waiting, another, minutes, went, care, staff, people, need, first, asked, now, service, day, left, take, come, times, since, waited, want, later, wait, phone, rude, experience, finally, long, seen, visit, years, worst, customer, needed, bad, hours

Frequent Terms from Reviews with 5.0 Star Rating:

great, time, best, back, recommend, like, amazing, always, years, love, staff, feel, good, first, experience, friendly, professional, done, now, appointment, highly, well, went, care, day, job, definitely, every, comfortable, nice, right, since, service, better, awesome, happy, need, took, felt, wanted, anyone, super, wonderful, give, next, last, visit, everyone, thank, clean

Discussion

By looking at the words that come up most often in reviews and tips for appointment only businesses we can see that time is a very common recurring theme. People often include words to describe how they felt, or were treated by staff. We can infer they are describing the feeling the experience is giving them, their level of comfort.

On the low end of the star ratings people often describe having to wait, or having their time wasted. They also note poor customer service, rudeness and being made to feel like they were bothering the staff.

On the high end of the star ratings people are effusive about how great they were made to feel, often specifically noting a high level of comfort. Terms associated with time are almost universally positive, often specifically noting a desire to return to do more business or asserting that they will never go anywhere else.

We can extrapolate from this data that a major expectation customers have of appointment only businesses is that they do not waste their time, and that they are treated with a certain level of care and formality. This makes sense, at least superficially, when you consider that the customer is making the effort to pre-plan their patronage of a business by scheduling it ahead of time. Perhaps this gives the customer some feeling of ownership, or agency, in the business transaction. The businesses themselves have the ability to plan ahead, since they can expect a certain volume of business ahead of time, which could be what is allowing them to score such high ratings as a whole when compared to other types of businesses.

Ultimately, the concepts of not having your time wasted and not being treated indifferently are fairly universal. Appointment only businesses may very well be handling these aspects of customer service simply because it is built into the format of their business model. Truly, any business that deals with people directly could develop methods to identify appropriate lengths of time for transactions to take, and how to appropriately make their customers feel special or at very least appreciated. The terms and phrases identified in this subset of the Yelp dataset provide a highly specific set of terms related to these two concepts which could be used to quickly mine other datasets for specific reviews describing either positive or negative customer experiences. These experiences could be used to model new ways of approacing the business/customer interaction. It is very likely that this kind of coordinated approach to time efficiency and overall customer experience would result in an overall increase in positive sentiment towards a business.