

```
#####
# Association Rules for TripAdvisor
# Author: Ravi Makhija
# Version 1.1
#
# Description:
# We explore the TripAdvisor dataset using association rule mining.
#
# File Dependencies:
#   'data/tripadvisor_data.Rdata'
#
# How to run:
#   Source this script (no need to set wd beforehand if directory structure is
#   maintained as downloaded).
#
# References
#   1) Set working directory to the file path of a script:
#       http://stackoverflow.com/questions/13672720/r-command-for-setting-working-dire
#   2) Tutorial on association rules in R:
#       http://www.rdatamining.com/examples/association-rules
#   3) Renaming levels of a factor:
#       http://www.cookbook-r.com/Manipulating\_data/Renaming\_levels\_of\_a\_factor/
#   4) Installing package from a source file:
#       https://cran.r-project.org/web/packages/arules/index.html

require("arules")    # version 2.2 is needed, which required installing from source

## Loading required package: arules
## Warning: package 'arules' was built under R version 3.2.2
## Loading required package: Matrix
##
## Attaching package: 'arules'
##
## The following objects are masked from 'package:base':
##
##   %in%, abbreviate, write

require("arulesViz")

## Loading required package: arulesViz
```

```

## Warning: package 'arulesViz' was built under R version 3.1.3
## Loading required package: grid
##
## Attaching package: 'arulesViz'
##
## The following object is masked from 'package:arules':
##
##   abbreviate
##
## The following object is masked from 'package:base':
##
##   abbreviate

require("Hmisc")

## Loading required package: Hmisc
## Warning: package 'Hmisc' was built under R version 3.1.3
## Loading required package: lattice
## Loading required package: survival
## Loading required package: splines
## Loading required package: Formula
## Warning: package 'Formula' was built under R version 3.1.3
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:base':
##
##   format.pval, round.POSIXt, trunc.POSIXt, units

require("data.table")

## Loading required package: data.table
## Warning: package 'data.table' was built under R version 3.1.3

require("plyr")

## Loading required package: plyr
##
## Attaching package: 'plyr'
##
## The following objects are masked from 'package:Hmisc':

```

```

##
##      is.discrete, summarize

#####
# Load in data
#####

load("tripadvisor_data.Rdata")

#####
# Prep data for association rules
#####

# Create a new data.frame for the data set we want to use association rules on.
# We want to create categorical variables for this purpose.

tripadvisor_data_categorical <- data.frame(user_is_local = as.factor(tripadvisor_data

# Now, we bin some continuous variables and add to this new data set.

#####
# user_review_length
# We omit this for TripAdvisor, since the data is incomplete with some of the
# reviews being cut off (e.g. they end with the word "More").

#####
# user_rating

# Explore data
table(tripadvisor_data$user_rating)

# Bin and add to new data set:
# For the time being, we keep all five categories:
tripadvisor_data_categorical$user_rating <- as.factor(tripadvisor_data$user_rating)

#####

```

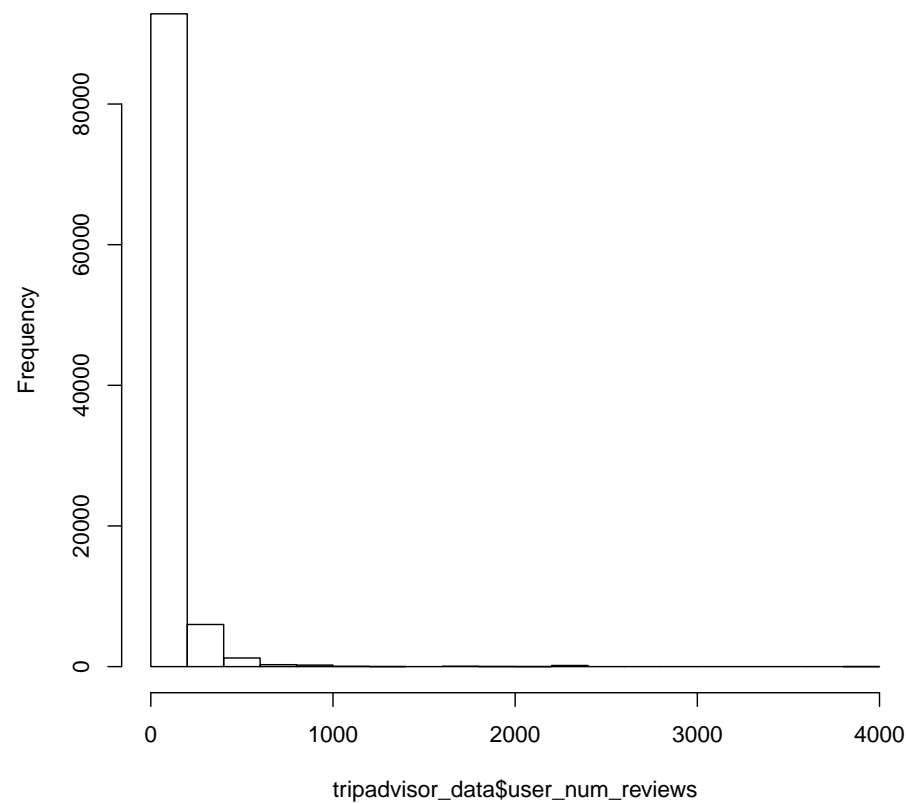
```
# user_num_reviews

# Explore data
summary(tripadvisor_data$user_num_reviews)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.00   17.00   43.00   80.21   93.00 3834.00

hist(tripadvisor_data$user_num_reviews)
```

Histogram of tripadvisor_data\$user_num_reviews



```
# Bin and add to new data set:
# low: [1 to 16)
```

```

# medium: [16 to 93)
# high: [83 and up)
tripadvisor_data_categorical$user_num_reviews <- cut2(x=tripadvisor_data$user_num_rev
                                                    cut=c(1, 16, 83))

#####
# Check out the new data set
head(tripadvisor_data_categorical)

##   user_is_local user_rating user_num_reviews
## 1          FALSE           5      [ 16,  83)
## 2          FALSE           5      [ 16,  83)
## 3          FALSE           5      [  1,  16)
## 4          FALSE           4      [ 16,  83)
## 5           TRUE           2      [  1,  16)
## 6          FALSE           5      [  1,  16)

#####
# Start association rules mining for tripadvisor
#####

attach(tripadvisor_data_categorical)

# Since a central question we are asking is whether or not local or non_local
# ratings are higher, we start association rule mining with the binary
# user_is_local on the right, to see if we can find any implications. We
# adjust the minimum support and confidence levels to obtain the most
# meaningful rule set. Just as we did for Yelp.

# A first look shows us that the higher user ratings 4 and 5 are associated
# with non_local reviews with a higher confidence than with local reviews.

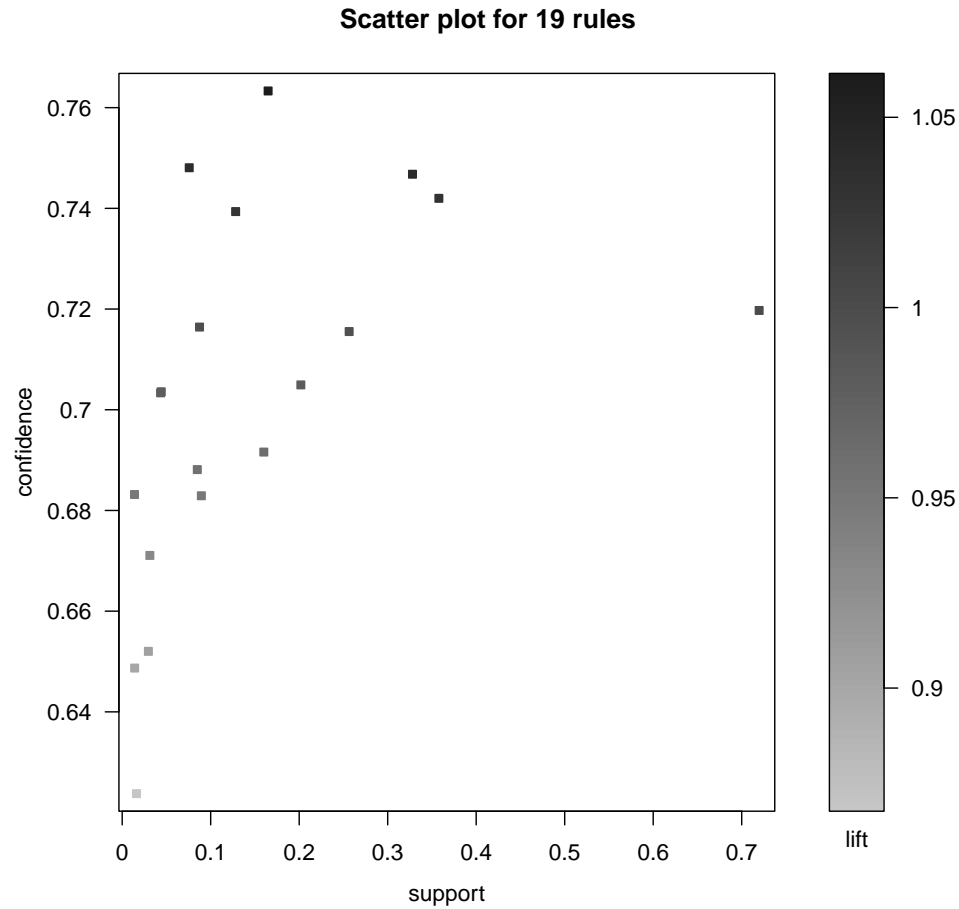
tripadvisor_rules_1 <- apriori(tripadvisor_data_categorical,
                               parameter = list(minlen=1, supp=.01, conf=.5),
                               appearance = list(rhs=c("user_is_local=FALSE", "user_i
                               control = list(verbose=F))

## Warning in match(x, table, nomatch = 0):  bytecode version mismatch;
## using eval
inspect(tripadvisor_rules_1)

```

##	lhs	rhs	support	confidence
## 1	{}	=> {user_is_local=FALSE}	0.71971913	0.7197191 1
## 2	{user_rating=1}	=> {user_is_local=FALSE}	0.01619542	0.6237586 0
## 3	{user_rating=2}	=> {user_is_local=FALSE}	0.02958416	0.6520219 0
## 4	{user_rating=3}	=> {user_is_local=FALSE}	0.08945662	0.6829194 0
## 5	{user_num_reviews=[1, 16]}	=> {user_is_local=FALSE}	0.16002023	0.6915988 0
## 6	{user_num_reviews=[83,3834]}	=> {user_is_local=FALSE}	0.20187244	0.7049351 0
## 7	{user_rating=4}	=> {user_is_local=FALSE}	0.25646874	0.7155308 0
## 8	{user_rating=5}	=> {user_is_local=FALSE}	0.32801420	0.7467769 1
## 9	{user_num_reviews=[16, 83]}	=> {user_is_local=FALSE}	0.35782646	0.7419899 1
## 10	{user_rating=2, user_num_reviews=[16, 83]}	=> {user_is_local=FALSE}	0.01396396	0.6831635 0
## 11	{user_rating=3, user_num_reviews=[1, 16]}	=> {user_is_local=FALSE}	0.01408297	0.6486980 0
## 12	{user_rating=3, user_num_reviews=[83,3834]}	=> {user_is_local=FALSE}	0.03126023	0.6710666 0
## 13	{user_rating=3, user_num_reviews=[16, 83]}	=> {user_is_local=FALSE}	0.04411342	0.7035748 0
## 14	{user_rating=4, user_num_reviews=[1, 16]}	=> {user_is_local=FALSE}	0.04340927	0.7033585 0
## 15	{user_rating=5, user_num_reviews=[1, 16]}	=> {user_is_local=FALSE}	0.08739376	0.7164228 0
## 16	{user_rating=4, user_num_reviews=[83,3834]}	=> {user_is_local=FALSE}	0.08485486	0.6881132 0
## 17	{user_rating=5, user_num_reviews=[83,3834]}	=> {user_is_local=FALSE}	0.07574060	0.7480654 1
## 18	{user_rating=4, user_num_reviews=[16, 83]}	=> {user_is_local=FALSE}	0.12820462	0.7393617 1
## 19	{user_rating=5, user_num_reviews=[16, 83]}	=> {user_is_local=FALSE}	0.16487985	0.7633150 1

plot(tripadvisor_rules_1)



```
# Narrowing down by increasing the minimum support shows again that higher
# ratings are associated with non_local reviewers. Of course, the non_local
# reviews are being prioritized with a higher support, due to the class
# imbalance with TripAdvisor data being in favor of non_local.

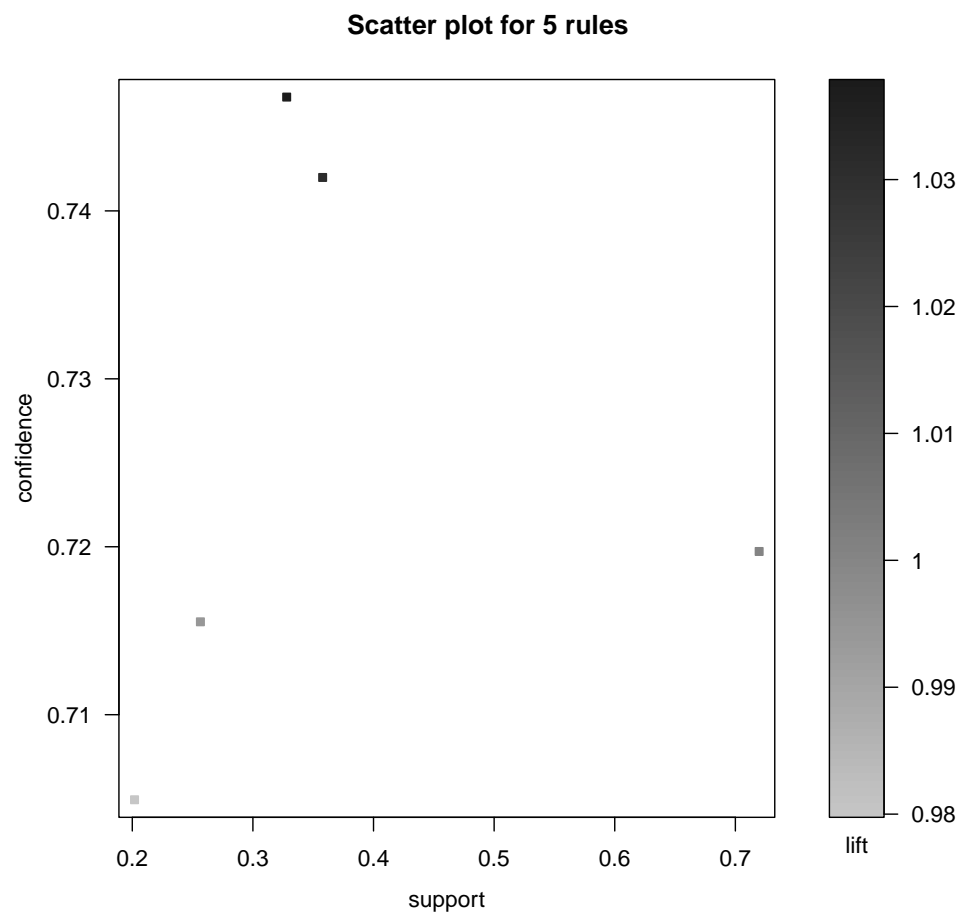
tripadvisor_rules_2 <- apriori(tripadvisor_data_categorical,
                               parameter = list(minlen=1, supp=.2, conf=.5),
                               appearance = list(rhs=c("user_is_local=FALSE", "user_i
                               control = list(verbose=F))

inspect(tripadvisor_rules_2)

##      lhs                                     rhs      support confidence
```

```
## 1 {} => {user_is_local=FALSE} 0.7197191 0.7197191 1.0
## 2 {user_num_reviews=[ 83,3834]} => {user_is_local=FALSE} 0.2018724 0.7049351 0.9
## 3 {user_rating=4} => {user_is_local=FALSE} 0.2564687 0.7155308 0.9
## 4 {user_rating=5} => {user_is_local=FALSE} 0.3280142 0.7467769 1.0
## 5 {user_num_reviews=[ 16, 83]} => {user_is_local=FALSE} 0.3578265 0.7419899 1.0

plot(tripadvisor_rules_2)
```



```
# Bringing the support level down to .1, but minimum confidence up to .7, we
# see again that higher ratings imply non_local reviews first. We also see that
# reviewers that have at least 16 reviews on TripAdvisor seem to also imply
# non_local reviews (as opposed to those who have very few reviews).
```



```

tripadvisor_rules_3 <- apriori(tripadvisor_data_categorical,
                               parameter = list(minlen=1, supp=.1, conf=.7),
                               appearance = list(rhs=c("user_is_local=FALSE", "user_i
                               control = list(verbose=F))

inspect(tripadvisor_rules_3)

##   lhs                                rhs          support confidence
## 1 {}                                => {user_is_local=FALSE} 0.7197191 0.7197191 1.0
## 2 {user_num_reviews=[ 83,3834]} => {user_is_local=FALSE} 0.2018724 0.7049351 0.9
## 3 {user_rating=4}                => {user_is_local=FALSE} 0.2564687 0.7155308 0.9
## 4 {user_rating=5}                => {user_is_local=FALSE} 0.3280142 0.7467769 1.0
## 5 {user_num_reviews=[ 16, 83]} => {user_is_local=FALSE} 0.3578265 0.7419899 1.0
## 6 {user_rating=4,
##   user_num_reviews=[ 16, 83]} => {user_is_local=FALSE} 0.1282046 0.7393617 1.0
## 7 {user_rating=5,
##   user_num_reviews=[ 16, 83]} => {user_is_local=FALSE} 0.1648798 0.7633150 1.0

plot(tripadvisor_rules_3)

```

