Yvonne Lee: MSDS 664 Week 5

Code

Assignment #5: Apriori Algorithm using bxBookRatings data.

Hide

```
head(bxBookRatings)
```

	User.ID <int>	ISBN <chr>	Book.Rating <int>
1	276725	034545104X	0
2	276726	0155061224	5
3	276727	0446520802	0
4	276729	052165615X	3
5	276729	0521795028	6
6	276733	2080674722	0

6 rows

Hide

```
head(bxUsers)
```

	User.ID <int>	Location <chr>	Age <chr>
1	1	nyc, new york, usa	NULL
2	2	stockton, california, usa	18
3	3	moscow, yukon territory, russia	NULL
4	4	porto, v.n.gaia, portugal	17
5	5	farnborough, hants, united kingdom	NULL
6	6	santa monica, california, usa	61

6 rows

Hide

```
head(bxBooks)
```

	ISBN <chr>
1	0195153448
2	0002005018
3	0060973129
4	0374157065
5	0393045218
6	0399135782

6 rows | 1-2 of 8 columns

Hide

```
str(bxBookRatings)
```

```
'data.frame':   1149780 obs. of  3 variables:
 $ User.ID    : int  276725 276726 276727 276729 276729 276733 276736 276737 276744 276745 ...
 $ ISBN       : chr  "034545104X" "0155061224" "0446520802" "052165615X" ...
 $ Book.Rating: int  0 5 0 3 6 0 8 6 7 10 ...
```

Hide

```
str(bxUsers)
```

```
'data.frame':   278858 obs. of  3 variables:
 $ User.ID : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Location: chr  "nyc, new york, usa" "stockton, california, usa" "moscow, yukon territory, russia" "porto, v.n.gaia, portu
gal" ...
 $ Age     : chr  "NULL" "18" "NULL" "17" ...
```

Hide

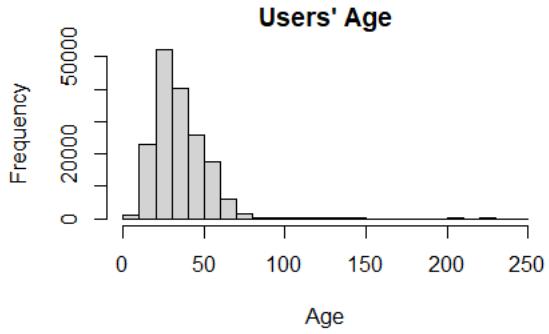
```
str(bxBooks)
```

```
'data.frame':   271379 obs. of  8 variables:
 $ ISBN               : chr  "0195153448" "0002005018" "0060973129" "0374157065" ...
 $ Book.Title         : chr  "Classical Mythology" "Clara Callan" "Decision in Normandy" "Flu: The Story of the Great Influe
nza Pandemic of 1918 and the Search for the Virus That Caused It" ...
 $ Book.Author        : chr  "Mark P. O. Morford" "Richard Bruce Wright" "Carlo D'Este" "Gina Bari Kolata" ...
 $ Year.Of.Publication: int  2002 2001 1991 1999 1999 1991 2000 1993 1996 2002 ...
 $ Publisher          : chr  "Oxford University Press" "HarperFlamingo Canada" "HarperPerennial" "Farrar Straus Giroux" ...
 $ Image.URL.S        : chr  "http://images.amazon.com/images/P/0195153448.01.THUMBZZZ.jpg" "http://images.amazon.com/image
s/P/0002005018.01.THUMBZZZ.jpg" "http://images.amazon.com/images/P/0060973129.01.THUMBZZZ.jpg" "http://images.amazon.com/ima
ges/P/0374157065.01.THUMBZZZ.jpg" ...
 $ Image.URL.M        : chr  "http://images.amazon.com/images/P/0195153448.01.MZZZZZZZ.jpg" "http://images.amazon.com/image
s/P/0002005018.01.MZZZZZZZ.jpg" "http://images.amazon.com/images/P/0060973129.01.MZZZZZZZ.jpg" "http://images.amazon.com/ima
ges/P/0374157065.01.MZZZZZZZ.jpg" ...
 $ Image.URL.L        : chr  "http://images.amazon.com/images/P/0195153448.01.LZZZZZZZ.jpg" "http://images.amazon.com/image
s/P/0002005018.01.LZZZZZZZ.jpg" "http://images.amazon.com/images/P/0060973129.01.LZZZZZZZ.jpg" "http://images.amazon.com/ima
ges/P/0374157065.01.LZZZZZZZ.jpg" ...
```

Hide

```
hist(as.numeric(bxUsers$Age),
     main="Users' Age",
     xlab = "Age")
```

```
NAs introduced by coercion
```



Hide

```
library(dplyr)
```

```
package 'dplyr' was built under R version 4.0.5
Attaching package: 'dplyr'

The following objects are masked from 'package:arules':

    intersect, recode, setdiff, setequal,
    union

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

Hide

```
library(ggplot2)
```

```
package 'ggplot2' was built under R version 4.0.5
```

Hide

```
library(arules)
library(arulesViz)
```

```
package 'arulesViz' was built under R version 4.0.5
```

Hide

```
top20<-bxBookRatings %>% group_by(ISBN) %>% summarise(n=n()) %>% top_n(n=20) %>% arrange(n)
```

```
Selecting by n
```

Hide

```
top20<-merge(top20,bxBooks[,c("ISBN","Book.Title")])
ggplot(top20,aes(x=reorder(Book.Title,n),n))+ geom_bar(stat='identity')+ theme(axis.text.x=element_text(angle=90, hjust=1))+
coord_flip() + labs(x = "Book Title",y="Number of Ratings")
```

Hide

```
rules <- apriori(bxBookRatings, parameter=list(support=0.000005, confidence = 0.3, target='rules'))
```

```
Column(s) 1, 2, 3 not logical or factor. Applying default discretization (see '? discretizeDF').The calculated breaks are:
0, 0, 5, 10
  Only unique breaks are used reducing the number of intervals. Look at ? discretize for details.
```

```
Apriori

Parameter specification:
```

confidence <dbl>	minval <dbl>	smax <dbl>	arem <chr>	aval <lgl>	originalSupport <lgl>	maxtime <dbl>	support <dbl>	minlen <int>
0.3	0.1	1	none	FALSE	TRUE	5	5e-06	1

1 row | 1-10 of 12 columns

```
Algorithmic control:
```

filter <dbl>	tree <lgl>	heap <lgl>	memopt <lgl>	load <lgl>	sort <int>	verbose <lgl>
0.1	TRUE	TRUE	FALSE	TRUE	2	TRUE

1 row

```
Absolute minimum support count: 5

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[340556 item(s), 1149780 transaction(s)] done [2.15s].
sorting and recoding items ... [34762 item(s)] done [0.10s].
creating transaction tree ... done [0.94s].
checking subsets of size 1 2 3 done [5.16s].
writing ... [102890 rule(s)] done [1.78s].
creating S4 object  ... done [0.51s].
```

Hide

```
rules.sorted <- sort(rules, by = "lift")
```

Hide

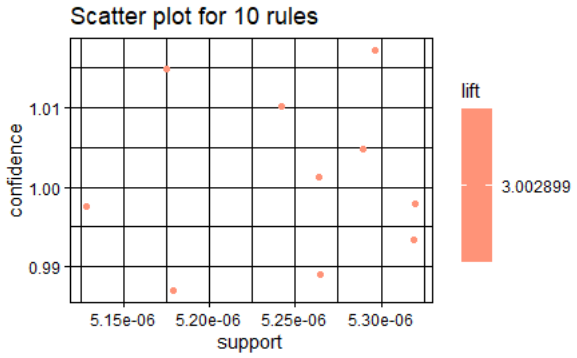
```
inspect(rules.sorted[1:1])
```

```
    lhs                  rhs                         support confidence    coverage     lift count
[1] {ISBN=1556610599} => {User.ID=[2,9.49e+04)} 5.21839e-06          1 5.21839e-06 3.002899     6
```

Hide

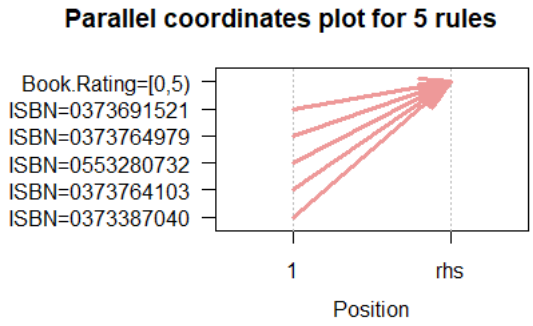
```
plot(rules.sorted[1:10])
```

```
To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```



Hide

```
plot(rules[1:10], method="paracoord", control=list(reorder=TRUE))
```



It seems that for the amount of data being processed, the minimum value for support and confidence needs to be pretty low. Using the initial suppor = 0.05 and confidence = 0.7 did not result in any rules to be produced. After changing support to 0.00005 and confidence to 0.3 did I finally receive results. The scatterplot seems like the most helpful showing support, confidence, and lift.

References:

Implementing Apriori algorithm in R. DataScience+. (n.d.). Retrieved November 23, 2021, from https://datascienceplus.com/implementing-apriori-algorithm-in-r/ (https://datascienceplus.com/implementing-apriori-algorithm-in-r/).

SAKSHIKULSHRESHTHA. (2021, August 20). Apriori algorithm in R programming. GeeksforGeeks. Retrieved November 23, 2021, from https://www.geeksforgeeks.org/apriori-algorithm-in-r-programming/# (https://www.geeksforgeeks.org/apriori-algorithm-in-r-programming/#):~:text='apriori()'%20function%20is%20in,for%20finding%20the%20association