Question 1: Chapter 6. Exercise 2

Recap

$$support = \frac{(X \cup Y).count}{n}$$
$$confidence = \frac{(X \cup Y).count}{X.count}$$

CustomerID	TransactionID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

a. Compute the support for itemsets {e}, {b, d}, and {b, d, e} by treating each transaction ID as a market basket

Itemsets	Transaction	Support
{e}	0001, 0024, 0012, 0031,	8/10 = 0.8 = 80%
	0015, 0022, 0033, 0038	
{b,d}	0012, 0022	2/10 = 0.2 = 20%
{b, d, e}	0012, 0022	2/10 = 0.2 = 20%

b. Use the results in part(a) to compute the confidence for the association rules {b,d} -> {e} and {e} -> {b, d}. Is confidence a symmetric measure?
 Recap

Confidence {b, d} -> {e} =
$$\frac{\{b,d,e\}.count}{\{b,d\}.count} = \frac{2}{2} = 100\%$$

Confidence {e} -> {b, d} =
$$\frac{\{e,b,d\}.count}{\{e\}.count} = \frac{2}{8} = 25 \%$$

Based on the result above, confidence is NOT a symmetric measure.

c. Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise).

Reconstruct the table we have:

CustomerID	a	b	С	d	e
1	1	1	1	1	1
2	1	1	1	1	1
3	0	1	1	1	1
4	1	1	1	1	0
5	1	1	0	1	1

Itemsets	CustomerID	Support
<i>{e}</i>	1,2,3,5	4/5 = 0.8 = 80%
<i>{b,d}</i>	1,2,3,4,5	5/5 = 1 = 100%
{b, d, e}	1,2,3,5	4/5 = 0.8 = 80%

d. Use the results in part (c) to compute the confidence for the association rules {b, d} -> {e} and

$$\{e\} -> \{b, d\}$$

Confidence {b, d} -> {e} =
$$\frac{\{b,d,e\}.count}{\{b,d\}.count} = \frac{4}{5} = 80\%$$

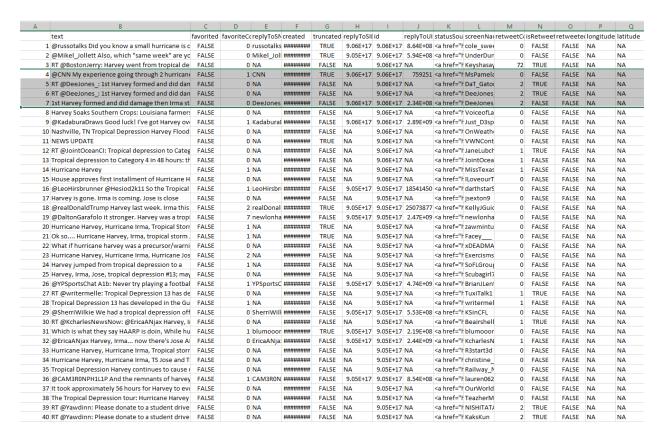
Confidence {e} -> {b, d} =
$$\frac{\{e,b,d\}.count}{\{e\}.count} = \frac{4}{4} = 100 \%$$

e. Suppose s₁ and c₁ are the support and confidence values of an association rule *r* when treating each transaction ID as a market basket. Also, let s₂ and c₂ be the support and confidence values of *r* when treating each customer ID as a market basket. Discuss whether there are any relationships between s₁ and s₂ or c₁ and c₂

It seems that s₂ is the generalized version of s₁ when we group transaction by customerID. Because there is no clear relationship between CustomerID and TransactionID in terms of items. So there is no clear relationship between s₁, c₁ and s₂, c₂

Question 3:

Our original dataset contains **5,312 records**



First: We want to remove all retweets because they contribute no additional information. As shown in the shaded region above. This step can be done by filtering and get all isRetweet = FALSE.

Our dataset in this step shrinks down to 1,718 records. We name this dataset is TwitterData

xt	
russotalks Did you know a small hurricane is coming here too? Or some tropical depression or something I don't knoae, https://t.co/GdfyDz5luK	
Mikel_Jollett Also, which "same week" are you referring too? Because Irma formed 9 DAYS after Harvey was in the Gu… https://t.co/UVoBwI3NB1	
CNN My experience going through 2 hurricanes and a stagnant tropical depression in Houston (NOT inc Harvey): FIVE 6€ https://t.co/k2nYWvl8Qm	
t Harvey formed and did damage then Irma started to form and Jose is right behind Irma as a tropical depression	
rvey Soaks Southern Crops: Louisiana farmers took the brunt of the rain and crop damage. https://t.co/Pixe07Ro42	
KadaburaDraws Good luck! I've got Harvey overhead here as a tropical depression, and it's still a pain	
shville, TN Tropical Depression Harvey Flooding Closes I40 - 8/31/2017 https://t.co/LCH7OGXF2L https://t.co/fFTC1MRfeO	
WS UPDATE	
opical depression to Category 4 in 48 hours: the science behind #HurricaneHarvey's rapid intensification https://t.co/wsMlokR8Bn	
rricane Harvey	
use approves first installment of Hurricane Harvey disaster aid https://t.co/cVFUWc6JVC via @usatoday	
LeoHirsbrunner @Hesiod2k11 So the Tropical Depression in the Gulf got upgraded and got a namethat's where Harvey was born too.	
rvey is gone. Irma is coming. Jose is close behind. And now there's another one!	
realDonaldTrump Harvey last week. Irma this week. Jose in the Atlantic right behind Irma. New tropical depressionåe¦ https://t.co/wcbLMDG2qB	
DaltonGarafolo it stronger. Harvey was a tropical depression and turned into a category 5 within under 48 hours	
rricane Harvey, Hurricane Irma, Tropical Storm Jose, Tropical Depression THIRTEEN. Mother nature is angry!… https://t.co/4eluhpifSN	
so Hurricane Harvey, Irma, tropical storm Josî, & NOW tropical depression 13 has formed in the gulf, like-riâ€; https://t.co/CBfWMxiOON	
hat if hurricane harvey was a precursor/warning of the storms to come(irma, jose and a tropical depression forming in the gulf)	

Data is good now we need to clean them. First, we must remove non-ansi characters as highlighted above. This step can be done in R with the following command

data <- iconv(TwitterData\$text, "latin1", "ASCII", sub="") //we remove non-ansi characters and copy to data

New data below

```
[2] "@Mikel_Jollett Also, which \"same week\" are you referring too? Because Irma formed 9 DAYS after Harvey was in the Gu https://t.co/UVOBWI3NB1"
[3] "@CNN My experience going through 2 hurricanes and a stagnant tropical depression in Houston (NOT inc Harvey): FIVE https://t.co/k2nYWv18Qm"
[4] "1st Harvey formed and did damage then Irma started to form and Jose is right behind Irma as a tropical depression"
[5] "MakadaburaDraws Good luck! I've got Harvey overhead here as a tropical depression, and it's still a pain"
[7] "Nashville, TN Tropical Depression Harvey Flooding Closes I40 - 8/31/2017 https://t.co/LCH70GXF2L https://t.co/FTCIMRfe0"
[8] "NeWS UPDATE\INTOpical Depression Harvey\n\nThe death toll across the affected areas has risen to at least 70 - it coul https://t.co/x8STRyjvaD"
[9] "Tropical depression to Category 4 in 48 hours: the science behind #HurricaneHarveys rapid intensification https://t.co/wswlokR8Bm"
[10] "Hurricane Harvey\n\nTexas under water\n\nHurricane Harvey disaster aid https://t.co/cVFUWcG3VC via @usatoday"
[11] "House approves first installment of Hurricane Harvey disaster aid https://t.co/cVFUWcG3VC via @usatoday"
[12] "@LeoHirsbrunner @Hesiod2kl1 So the Tropical Depression in the Gulf got uppraded and got a name...that's where Harvey was born too."
[13] "Harvey is gone. Irma is coming. Jose is close behind. And now there's another one\n\nhttps://t.co/WBLG7wcC7"
[14] "@realDonaldTrump Harvey last week. Irma this week. Jose in the Atlantic right behind Irma. New tropical depression https://t.co/wbLMDG2qB"
[15] "@QultonGarafolo it stronger. Harvey was a tropical depression and turned into a category S within under 48 hours"
[16] "Hurricane Harvey, Hurricane Irma, Tropical Storm Jose, Komp; NOW tropical Depression 11 has formed in the gulf, like ri https://t.co/eLMDfSqB"
[17] "Ok so... Hurricane harvey was a precursor/warning of the storms to come(irma, jose and a tropical depression is supposed to hit mexico. were all dead."
[18] "Hurricane Harvey, Hurricane Irma, Hurricane Jose is f
```

Next we have to:

- remove all URLs (ie https://t.co/k2nYWvl8Qm)
 - gsub("http[^[:space:]]*", "", x)
- remove all people (ie @Mike)
 - gsub("@[^[:space:]]*", "", x)
- remove all punctuation
 - gsub("[^[:alpha:][:space:]]*", "", x)
- remove all stopwords
 - removeWords(x, c(stopwords("english"),"at","via","amp"))
- remove all whitespace
 - stripWhitespace(x)
- remove all new lines
 - gsub("\n", "", x)
- stemming the document
 - stemDocument(x)

The result of this step is shown as below

```
Console ~/  

[272] "hurricaneharvey weaken tropic depress death toll rise"
[273] "latest offici rais harveyrel death toll houston ap latest tropic depress"
[274] "latest atlant tropic outlook nhc tropic depress harvey public advisori number"
[275] "hour am pm multipl radarmultipl sensor mrms rainfal estim tropic depress harvey"
[276] "tropic storm harvey photo mustse pictur storm"
[277] "tropic depress harvey public advisori number"
[278] "tropic depress harvey public advisori number"
[279] "tropic depress harvey advisori kt mph wind mb n w move ne kt mph tropic"
[280] "noaa tropic depress harvey public advisori number"
[281] "nhcatlant tropic depress harvey advisori now avail nhc websit"
[282] "noaa tropic depress harvey public advisori number"
[283] "harvey slog inland dump heavi rain amid flood threat memphi tenn ap tropic depress harvey s"
[284] "tropic depress harvey public advisori number"
[285] "tropic depress harvey public advisori number"
[286] "tropic depress harvey public advisori number"
[287] "tropic depress harvey public advisori number the nhc issu final advisori system public ad"
[288] "nhcatlant tropic depress harvey advisori number the nhc issu final advisori system public ad"
[288] "hocatlant tropic depress harvey public advisori number"
[289] "tropic depress harvey public advisori number"
[280] "unfortun will stare tropic depress harvey pass direct bowl green"
[290] "unfortun will stare tropic depress harvey pass direct bowl green"
```

After cleaning the data, there are a lot of duplications, we need to remove them all and keep unique tweets. This can be done by the following command

cleaneddata <-unique(data)

Now, our final cleaned data has 1.025 tweets

Next step, we save the results as a csv file for association rules mining. We name the data as "basket"

write(cleaneddata,"basket")

Now we read the data as a transaction

trans = read.transactions("basket", format = "basket", sep=" ", rm.duplicates = TRUE);
inspect(trans)

The output of the transaction is:

```
{also,day,form,gu,harvey,irma,refer,week}
[3]
        {depress, experi, five, go, harvey, houston, hurrican, inc, stagnant, tropic}
[4]
[5]
[6]
       {behind,damag,depress,form,harvey,irma,jose,right,st,start,tropic}
        {brunt,crop,damag,farmer,harvey,louisiana,rain,soak,southern,took}
       {depress,good,got,harvey,ive,luck,overhead,pain,still,tropic}
[7]
[8]
        {close, depress, flood, harvey, nashvill, tn, tropic}
       {across,affect,area,coul,death,depress,harveyth,least,news,risen,toll,updatetrop}
[9]
        {behind,categori,depress,hour,hurricaneharvey,intensif,rapid,scienc,tropic}
[10]
       {dacacalifornia,depress,fire,harveytexa,hurrican,irmatrop,josetrop,katia,storm,tbc,waterhurrican}
[11]
        {aid,approv,disast,first,harvey,hous,hurrican,instal}
[12]
       {born,depress,got,gulf,harvey,namethat,tropic,upgrad}
[13]
        {anoth, behind, close, come, gone, harvey, irma, jose, now, one, there}
       {atlant,behind,depress,harvey,irma,jose,last,new,right,tropic,week} {categori,depress,harvey,hour,stronger,tropic,turn,within}
[14]
[15]
[16]
        {angri,depress,harvey,hurrican,irma,jose,mother,natur,storm,thirteen,tropic}
[17]
        {depress, form, gulf, harvey, hurrican, irma, jos, like, now, ok, ri, storm, tropic}
[18]
        {comeirmajos,depress,form,gulf,harvey,hurrican,precursorwarn,storm,tropic}
[19]
        {dead,depress,form,harvey,hit,hour,hurrican,irma,jose,mexico,suppos,tropic}
       {cat,categori,day,depress,fl,floridan,gtfo,harvey,hour,hurrican,jump,tropic,warn}
```

Now we are good to run the algorithm

First, we want to find all rules with default values: minsupport = 0.1, minconfidence = 0.8

rules <-apriori(trans)

We got totally 85 rules

```
Console ~/ 🔗
Parameter specification:
 confidence minval smax arem aval originalSupport maxtime support minlen maxlen target
        0.8
               0.1
                      1 none FALSE
                                                                0.1
                                                                               10 rules FALSE
                                                                         1
Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE FALSE TRUE
                                 2
                                      TRUF
Absolute minimum support count: 101
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[1597 item(s), 1018 transaction(s)] done [0.00s].
sorting and recoding items \dots [11 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [85 rule(s)] done [0.00s].
creating S4 object ... done [0.00s].
```

Inspect the rules : *inspect(rules)* we got

```
> inspect(rules)
          1hs
                                                                      rhs
                                                                                           support
                                                                                                                confidence lift
                                                                                                                                                          count
                                                                => {depress} 0.8870334 0.8870334 1.000000 903
[1]
          {}
[2]
                                                                => {tropic} 0.9037328 0.9037328 1.000000 920
          {}
[3]
          {}
                                                               => {harvey} 0.9341847 0.9341847
                                                                                                                                      1.000000 951
                                           => {harvey} 0.1041257 0.9614615 1.036025 165
=> {depress} 0.1218075 0.9185185 1.035495 124
=> {tropic} 0.1257367 0.9481481 1.049147 128
=> {harvey} 0.1247544 0.9407407 1.007018 127
=> {depress} 0.1031434 0.9545455 1.076110 105
=> {tropic} 0.1041257 0.9636364 1.066285 106
=> {harvey} 0.1070727 0.9909091 1.060721 109
=> {depress} 0.1257367 0.8951049 1.009099 128
=> {tropic} 0.1326130 0.9440559 1.044618 135
=> {harvey} 0.1345776 0.9580420 1.025538 137
=> {depress} 0.1296660 0.9295775 1.047962 132
=> {tropic} 0.1355599 0.9718310 1.075352 138
=> {harvey} 0.1365422 0.9788732 1.047837 139
=> {depress} 0.1611002 0.9479769 1.068705 164
[4]
                                                               => {harvey} 0.1041257 0.9814815 1.050629 106
         {move}
[5]
         {weaken}
         {weaken}
[6]
[7]
         {weaken}
[8]
         {rain}
[9] {rain}
[10] {rain}
[11] {hurrican}
[12] {hurrican}
[13] {hurrican}
[14] {flood}
[15] {flood}
[16] {flood}
                                                               => {depress} 0.1611002 0.9479769 1.068705 164
[17] {now}
```

Given a minsupport =0.2 and minconf = 0.2 we run the algorithm rules 2 < -apriori(trans,parameter = list(supp = 0.2, conf = 0.2)) we got total 13 rules that satisfy the conditions.

```
> inspect(rules2)
     1hs
                            rhs
                                                   confidence lift
                                                                          count
                                        support
                        => {downgrad} 0.2023576 0.2023576 1.000000 206
[1]
     {}
[2]
     {}
                        => {depress} 0.8870334 0.8870334
                                                               1.000000 903
[3]
                        => {tropic}
                                        0.9037328 0.9037328
                                                               1.000000 920
     {}
[4]
     {}
                        => {harvey}
                                        0.9341847 0.9341847
                                                               1.000000 951
                 => {tropic} 0.8703340 0.9811739

=> {depress} 0.8703340 0.9630435

=> {harvey} 0.8467583 0.9545958

=> {depress} 0.8467583 0.9064143
[5]
     {depress}
                                                               1.085690 886
[6]
     {tropic}
                                                               1.085690 886
     {depress}
[7]
                                                               1.021849 862
[8]
     {harvey}
                                                               1.021849 862
[9]
     {tropic}
                        => {harvey}
                                        0.8693517 0.9619565
                                                               1.029728 885
                        => {tropic}
[10] {harvey}
                                        0.8693517 0.9305994
                                                               1.029728 885
[11] {depress,tropic} => {harvey}
                                        0.8408644 0.9661400
                                                               1.034207 856
[12] {depress,harvey} => {tropic}
                                        0.8408644 0.9930394
                                                               1.098820 856
[13] {harvey,tropic} => {depress} 0.8408644 0.9672316
                                                               1.090412 856
```