Grocery associate rule mining

Soumik Choudhuri

# ASSOCIATE RULE MINING

## Treatment of the data

In this step, we first took the Groceries dataset where each row represented the items bought by a user We then created a dataframe to have only 2 columns - user\_id and product where user\_id is the row number for groceries datasets and grouped the products by user\_id

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.3 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.3 ✓ stringr 1.4.0  
## ✓ readr 2.0.1 ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

##   
## Attaching package: 'arules'

## The following object is masked from 'package:dplyr':  
##   
## recode

## The following objects are masked from 'package:base':  
##   
## abbreviate, write

##   
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':  
##   
## smiths

## Summary of the products distribution

## transactions as itemMatrix in sparse format with  
## 15296 rows (elements/itemsets/transactions) and  
## 169 columns (items) and a density of 0.01677625   
##   
## most frequent items:  
## whole milk other vegetables rolls/buns soda   
## 2513 1903 1809 1715   
## yogurt (Other)   
## 1372 34055   
##   
## element (itemset/transaction) length distribution:  
## sizes  
## 1 2 3 4   
## 3485 2630 2102 7079   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 2.000 3.000 2.835 4.000 4.000   
##   
## includes extended item information - examples:  
## labels  
## 1 abrasive cleaner  
## 2 artif. sweetener  
## 3 baby cosmetics  
##   
## includes extended transaction information - examples:  
## transactionID  
## 1 1  
## 2 2  
## 3 3

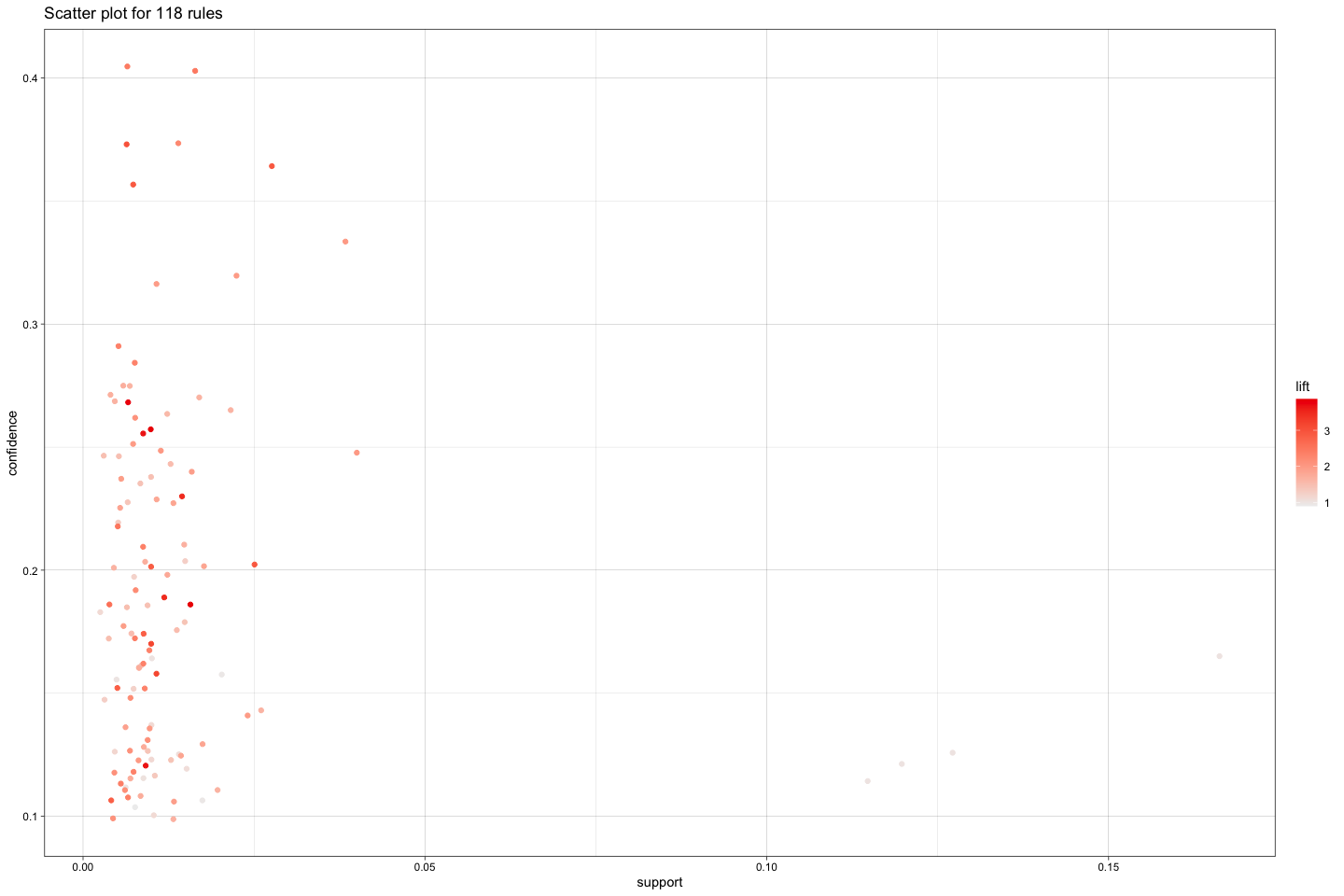
## OBSERVATIONS

In the summary of the prod\_dist, we find that “whole milk” was the highest bought item followed by “other vegetables” Also, we found that in majority (around 7079 times), 4 items were bought in a single transaction.

## Running the apriori function to get the support, confidence and lift

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.1 0.1 1 none FALSE TRUE 5 0.005 1  
## maxlen target ext  
## 5 rules TRUE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 76   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[169 item(s), 15296 transaction(s)] done [0.00s].  
## sorting and recoding items ... [101 item(s)] done [0.00s].  
## creating transaction tree ... done [0.00s].  
## checking subsets of size 1 2 3 done [0.00s].  
## writing ... [118 rule(s)] done [0.00s].  
## creating S4 object ... done [0.00s].

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



## OBSERVATIONS

From the scatter plot we found that there are a few items with support less than 0.05 but have a very high confidence above 0.3 which indicated that there is a good association between these items So we decided to further inspect these items.

## Inspection steps and Observations

We started with the inspection of product combination with confidence above 0.3

## lhs rhs support   
## [1] {onions} => {other vegetables} 0.007452929  
## [2] {curd} => {whole milk} 0.012617678  
## [3] {butter} => {whole milk} 0.014382845  
## [4] {root vegetables} => {other vegetables} 0.025366109  
## [5] {root vegetables} => {whole milk} 0.022620293  
## [6] {other vegetables} => {whole milk} 0.040860356  
## [7] {other vegetables,root vegetables} => {whole milk} 0.008172071  
## [8] {root vegetables,whole milk} => {other vegetables} 0.008172071  
## [9] {other vegetables,yogurt} => {whole milk} 0.006341527  
## confidence coverage lift count  
## [1] 0.3737705 0.01993985 3.004306 114   
## [2] 0.3683206 0.03425732 2.241875 193   
## [3] 0.4036697 0.03563023 2.457036 220   
## [4] 0.3619403 0.07008368 2.909216 388   
## [5] 0.3227612 0.07008368 1.964566 346   
## [6] 0.3284288 0.12441161 1.999064 625   
## [7] 0.3221649 0.02536611 1.960937 125   
## [8] 0.3612717 0.02262029 2.903842 125   
## [9] 0.3991770 0.01588651 2.429690 97

Upon inspection, we found that “onions” and “other vegetables” have the highest lift indicating a strong tendency of other vegetables being bought with onions Same is the case with “root vegetables” and “other vegetables” or vegetables along with milk which have lift close to 3 Among these, we find the highest count of 625 for the combination of “other vegetables” and “whole milk”.

Next, we inspect the product combinations with count > 150 and lift > 1.5

## lhs rhs support confidence  
## [1] {curd} => {whole milk} 0.01261768 0.3683206   
## [2] {butter} => {whole milk} 0.01438285 0.4036697   
## [3] {whipped/sour cream} => {whole milk} 0.01144090 0.2482270   
## [4] {pip fruit} => {tropical fruit} 0.01268305 0.2607527   
## [5] {tropical fruit} => {pip fruit} 0.01268305 0.1879845   
## [6] {pip fruit} => {other vegetables} 0.01091789 0.2244624   
## [7] {pip fruit} => {whole milk} 0.01255230 0.2580645   
## [8] {pastry} => {rolls/buns} 0.01019874 0.1782857   
## [9] {citrus fruit} => {tropical fruit} 0.01248692 0.2346437   
## [10] {tropical fruit} => {citrus fruit} 0.01248692 0.1850775   
## [11] {citrus fruit} => {other vegetables} 0.01281381 0.2407862   
## [12] {other vegetables} => {citrus fruit} 0.01281381 0.1029953   
## [13] {sausage} => {rolls/buns} 0.01078713 0.1785714   
## [14] {sausage} => {other vegetables} 0.01261768 0.2088745   
## [15] {other vegetables} => {sausage} 0.01261768 0.1014188   
## [16] {bottled water} => {soda} 0.01464435 0.2060718   
## [17] {soda} => {bottled water} 0.01464435 0.1306122   
## [18] {tropical fruit} => {root vegetables} 0.01098326 0.1627907   
## [19] {root vegetables} => {tropical fruit} 0.01098326 0.1567164   
## [20] {tropical fruit} => {other vegetables} 0.01549425 0.2296512   
## [21] {other vegetables} => {tropical fruit} 0.01549425 0.1245402   
## [22] {tropical fruit} => {whole milk} 0.01830544 0.2713178   
## [23] {whole milk} => {tropical fruit} 0.01830544 0.1114206   
## [24] {root vegetables} => {other vegetables} 0.02536611 0.3619403   
## [25] {other vegetables} => {root vegetables} 0.02536611 0.2038886   
## [26] {root vegetables} => {whole milk} 0.02262029 0.3227612   
## [27] {whole milk} => {root vegetables} 0.02262029 0.1376840   
## [28] {yogurt} => {whole milk} 0.02425471 0.2704082   
## [29] {whole milk} => {yogurt} 0.02425471 0.1476323   
## [30] {other vegetables} => {whole milk} 0.04086036 0.3284288   
## [31] {whole milk} => {other vegetables} 0.04086036 0.2487067   
## coverage lift count  
## [1] 0.03425732 2.241875 193   
## [2] 0.03563023 2.457036 220   
## [3] 0.04609048 1.510895 175   
## [4] 0.04864017 3.864800 194   
## [5] 0.06746862 3.864800 194   
## [6] 0.04864017 1.804191 167   
## [7] 0.04864017 1.570774 192   
## [8] 0.05720450 1.507495 156   
## [9] 0.05321653 3.477820 191   
## [10] 0.06746862 3.477820 191   
## [11] 0.05321653 1.935400 196   
## [12] 0.12441161 1.935400 196   
## [13] 0.06040795 1.509911 165   
## [14] 0.06040795 1.678898 193   
## [15] 0.12441161 1.678898 193   
## [16] 0.07106433 1.837944 224   
## [17] 0.11212082 1.837944 224   
## [18] 0.06746862 2.322805 168   
## [19] 0.07008368 2.322805 168   
## [20] 0.06746862 1.845898 237   
## [21] 0.12441161 1.845898 237   
## [22] 0.06746862 1.651444 280   
## [23] 0.16429132 1.651444 280   
## [24] 0.07008368 2.909216 388   
## [25] 0.12441161 2.909216 388   
## [26] 0.07008368 1.964566 346   
## [27] 0.16429132 1.964566 346   
## [28] 0.08969665 1.645907 371   
## [29] 0.16429132 1.645907 371   
## [30] 0.12441161 1.999064 625   
## [31] 0.16429132 1.999064 625

Upon putting the above condition of count and lift, we find that “pip fruit” and “tropical fruit” have a lift close to 4 indicating high association between them

We find that people generally buy whole milk and other vegetables together upon inspecting for lift > 1 and support > 0.03

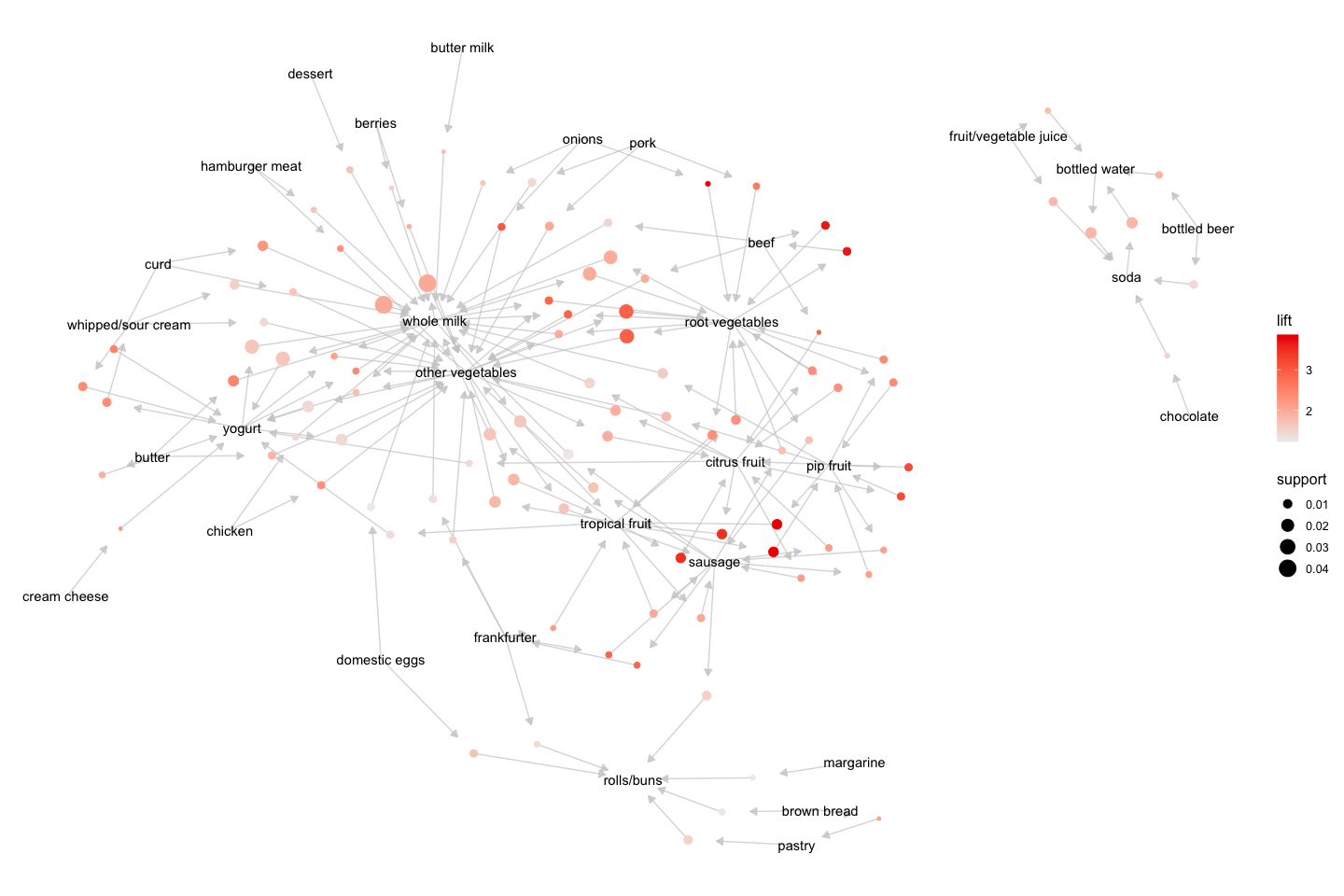
## lhs rhs support confidence coverage   
## [1] {other vegetables} => {whole milk} 0.04086036 0.3284288 0.1244116  
## [2] {whole milk} => {other vegetables} 0.04086036 0.2487067 0.1642913  
## lift count  
## [1] 1.999064 625   
## [2] 1.999064 625

Upon putting the condition of support > 0.10, we find that these are values for individual items like “soda” or “rolls/buns”

## lhs rhs support confidence coverage lift count  
## [1] {} => {soda} 0.1121208 0.1121208 1 1 1715   
## [2] {} => {rolls/buns} 0.1182662 0.1182662 1 1 1809   
## [3] {} => {other vegetables} 0.1244116 0.1244116 1 1 1903   
## [4] {} => {whole milk} 0.1642913 0.1642913 1 1 2513

## NETWORK GRAPHS

We then plot the network graph to see the association between the different items bought.



## OBSERVATIONS

1)We get to see that other vegetables and whole milk are the items which are highest associated with other items i.e. they have the highest network.

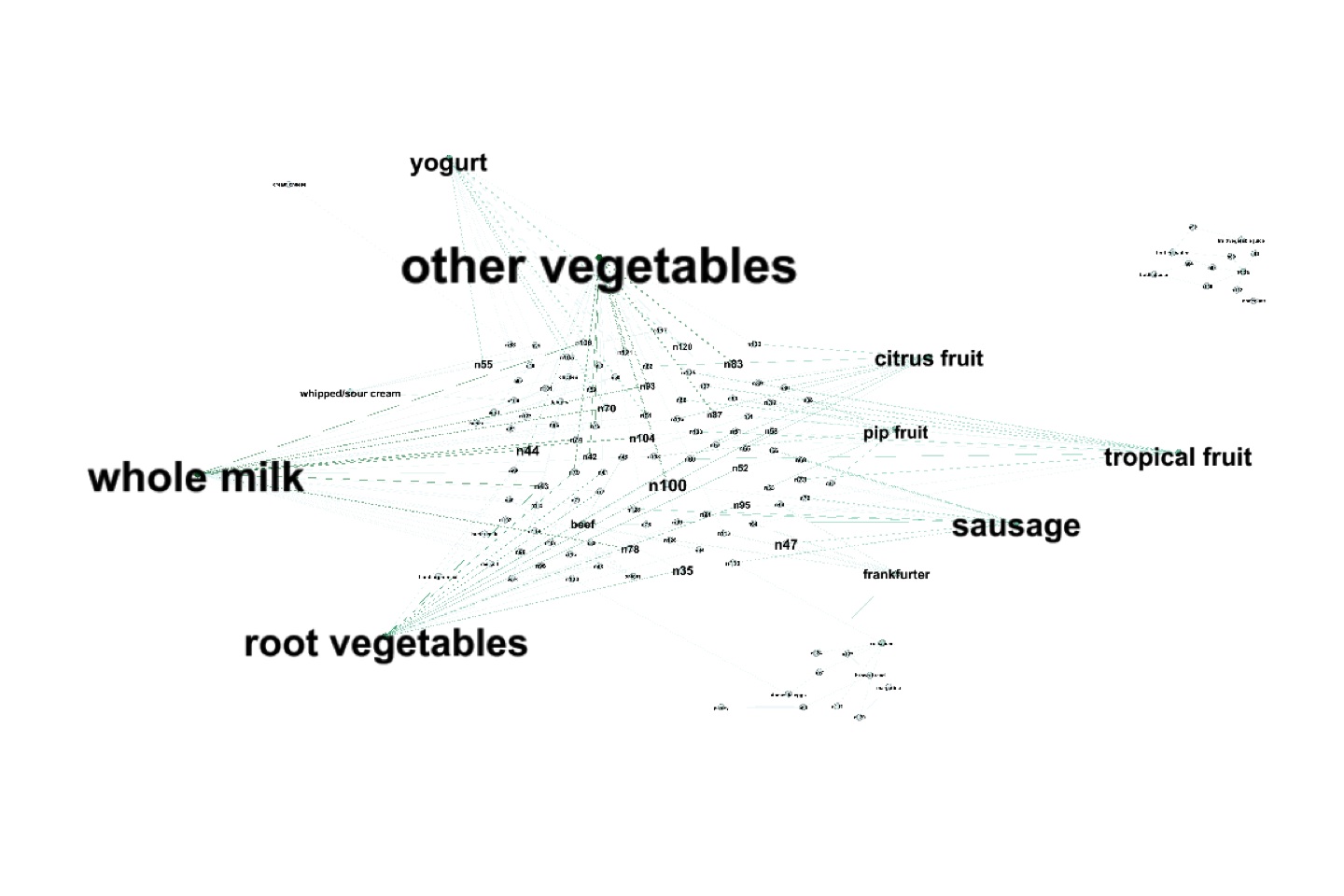
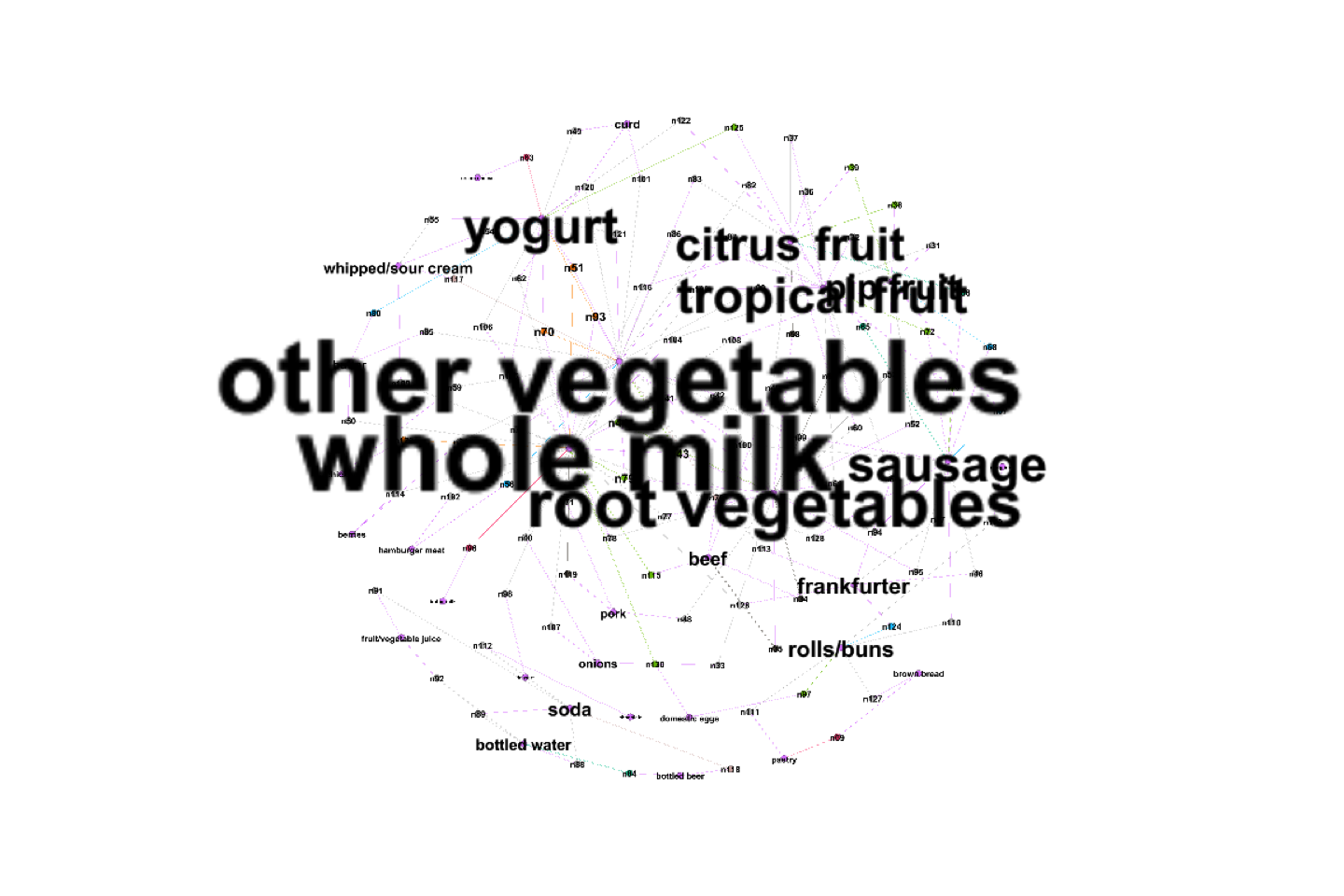
1. Also, we find bottled water, beer, soda, chocolate and juice belong to a completely different cluster indicating they are not associated with vegetables or whole milk or other regular items

## GEPHI PLOTS

Finally, we use the Gephi software to visualize the association between the items.

##   
## Attaching package: 'EBImage'

## The following object is masked from 'package:purrr':  
##   
## transpose



## OBSERVATIONS

1. Here we see that “other vegetables” and “whole milk” have the biggest font followed by “root vegetables”, “yogurt”, “sausage” and others.
2. The font of an item is proportional to the degree of the node which in turn is related to the association of that item. The Gephi plot also validates the earlier network plot indicating that vegetables and whole milk have highest association and are often bought with other items
3. In the second Gephi plot we find that there are 2 clusters validating our earlier inference that bottled water, beer, soda are not associated with the regular items and are mostly bought separately.