STA 380, Part 2: Exercises

Andrew Gillock

2022-08-15

LINK TO GITHUB REPO: https://github.com/andrewgillock/groupExercises

Question 8: The Reuters Corpus

Question

Similar to our example in class, we hope to dissect an author's documents and try our best to group them by their contents. More specifically, we'd like to answer the question of what terms appear most frequently in author Heather Scoffield's documents, as well as how related these documents are in space.

Approach

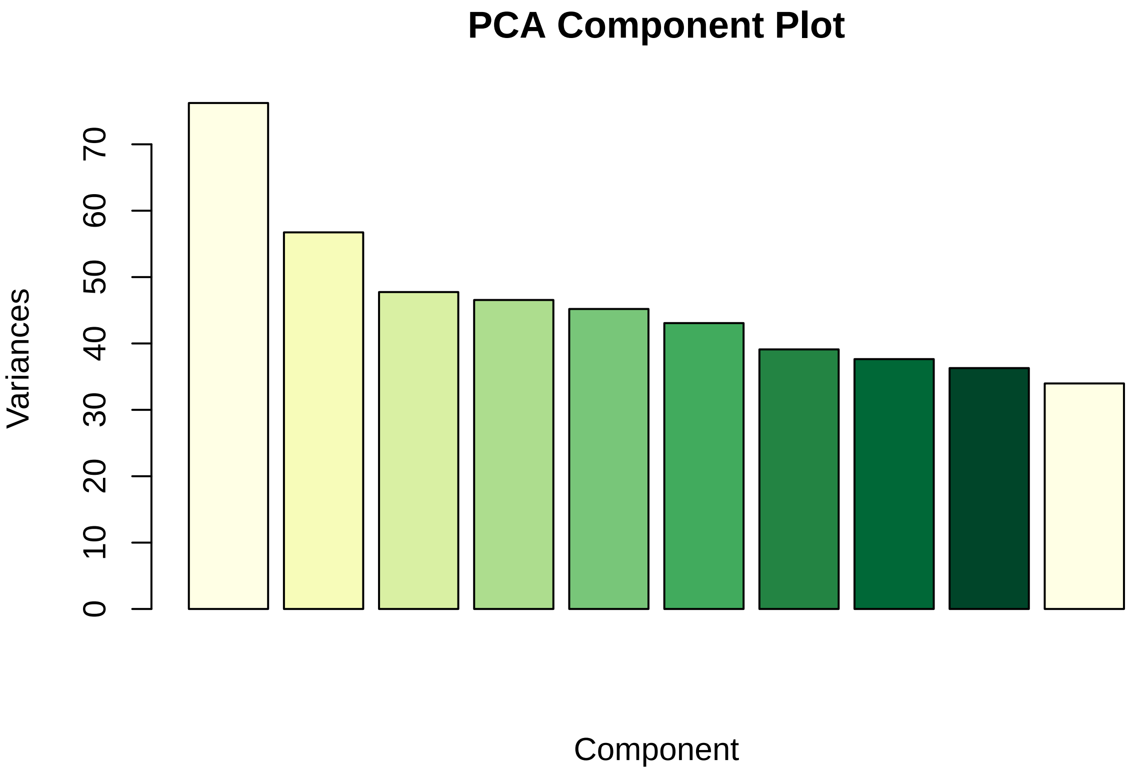
To answer this question, we will start by reading each of Heather Scoffield's documents into a file list. To make interpretation and recognition easier, we can clean up the original document names and applied them to the file list. We then plan to follow general tokenization steps, such as removing numbers, punctuation, white spaces, and common stop words. This whole process ensures we are analyzing relevant data and removes unnecessary words from each document.

The next step is to create a document-term matrix (DTM), which contains the frequency of terms in each of the documents. The corresponding inverse document frequency (IDF) assigns higher weights to uncommon terms across the corpus as a whole. Finally, a term frequency-inverse document frequency (TF-IDF) will be obtained by combining the DTM and IDF.

Finally, we will perform principal component analysis (PCA) to see if we can summarize each document within the corpus given its qualities. Additionally, we will perform clustering of the documents to see which ones are most similar.

Results

After completing PCA, we have summarized each document in the corpus into a single pair of values in the real coordinate space. In the plot below, we can see that we were able to maintain a majority of the variation between documents within the first 2 principal components. Words that appeared 75 or more times within the DTM are listed below. Furthermore, examining the loadings will give us an idea of which words have the strongest influence on each component.



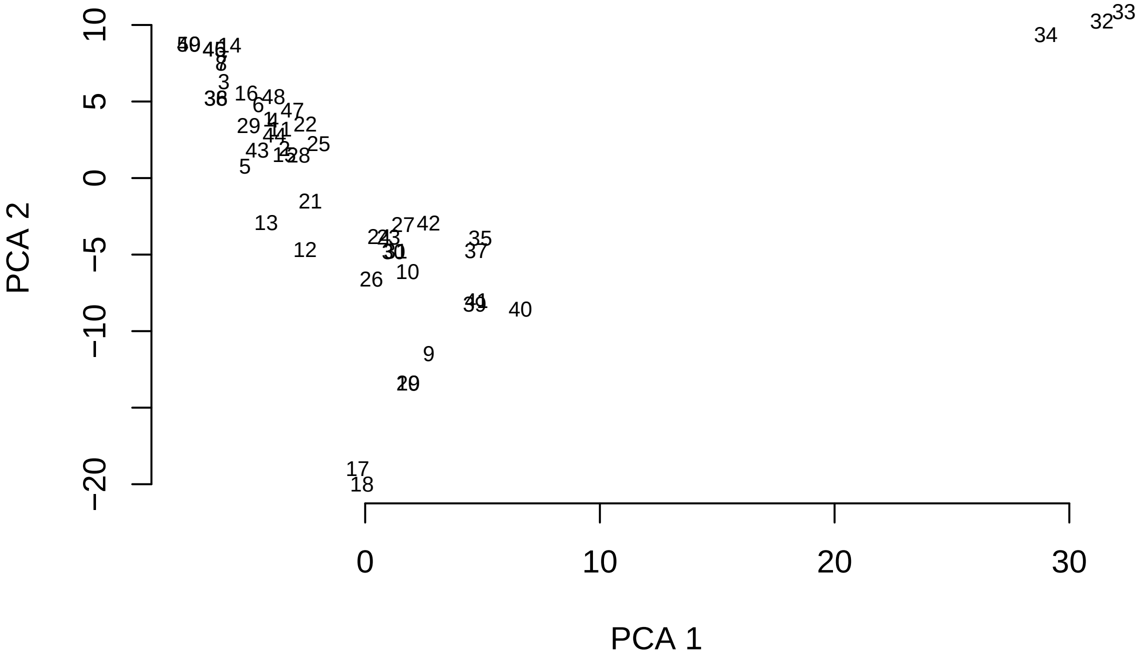
```
##  [1] "character"   "companies"   "company"     "exploration" "million"
##  [6] "mining"      "percent"     "said"        "will"        "year"
## [11] "analysts"    "corp"        "gold"        "government"  "stock"
## [16] "barrick"     "brex"        "busang"      "indonesian"
```

```
##    deadline      blocks     busangs      clinch     clocked    crossing
##   0.1037899   0.1034938   0.1034938   0.1034938   0.1034938   0.1034938
##    directed discouraged     dotting    expedite
##   0.1034938   0.1034938   0.1034938   0.1034938
```

```
##        brex     minorca  indonesian    suhartos   contracts       horst
## -0.09406477 -0.08760043 -0.08734735 -0.08572991 -0.08527286 -0.08497220
##      roland   indonesia    partners         son
## -0.08460859 -0.08315992 -0.08307677 -0.08111043
```

Now we can examine the magnitude of the first 2 principal components for each document. Furthermore, we can see how the documents were placed in the real coordinate space and try to see what qualities they have. From the plot below, we can see that documents 32, 33, and 34 share similar characteristics, as well as documents 17 and 18. Let's try and get a summary of their content.

```
##
## Docs       PC1        PC2
##   1  -4.111420  3.8544729
##   2  -3.438514  1.9221374
##   3  -6.019950  6.3068271
##   4  -3.917225   3.7813299
##   5  -5.122433   0.7673308
##   6  -4.549808   4.7863933
##   7  -6.102428   7.5450820
##   8  -6.136413   7.5013545
##   9   2.715296 -11.4886497
##   10  1.808049  -6.1213339
```



```
#view first group of docs
content(heather[[32]])[1:3]
```

```
## [1] "No final deal was in sight Wednesday for Bre-X Minerals Ltd. and Barrick Gold Corp., which are
## [2] "As a Wednesday deadline slid by, Bre-X and Barrick said they were still trying to hammer out se
## [3] "\"Several points remain outstanding,\" said Barrick spokesman Vince Borg. \"An overall deal has
```

```
content(heather[[33]])[1:3]
```

```
## [1] "Investors were on edge on Wednesday, anxiously awaiting the outcome of talks between Canada's B
## [2] "As a Dec. 4 deadline slid by, Bre-X and Barrick said they were still trying to work out several
## [3] "\"Several points remain outstanding,\" Barrick spokesman Vince Borg said. \"An overall deal has
```

```
content(heather[[34]])[1:3]
```

```
## [1] "No final deal was in sight Wednesday for Bre-X Minerals Ltd. and Barrick Gold Corp., which are
## [2] "As a Wednesday deadline slid by, Bre-X and Barrick said they were still trying to hammer out se
## [3] "A few issues remain to be solved, and Bre-X will have more news on the negotiations \"shortly,\
```

Documents 32, 33, and 34 discuss a business deal between Bre-X Minerals Ltd, Barrick Gold Corp., and Indonesia's Busang gold deposits. These documents appear to be recording relevant updates between mining companies and the Indonesian government.

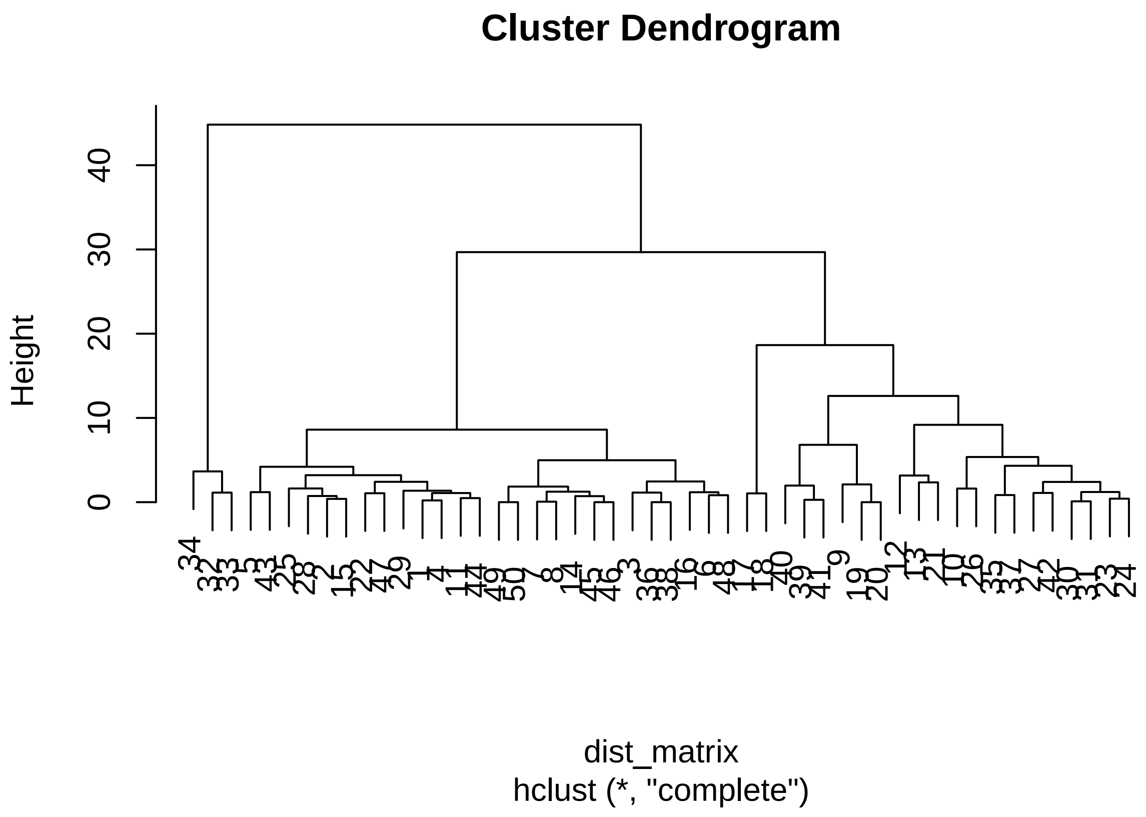
```
#view second group of docs
content(heather[[17]])[1:3]
```

```
## [1] "Canadian mining company Bre-X Minerals Ltd has been in hiding since it announced 17 days ago tha
## [2] "Despite a constant whirl of rumors and persistent questions that have sent the company's shares
## [3] "The Calgary-based company that controls one of the world's biggest gold prospects in Indonesia I
```

```
content(heather[[18]])[1:3]
```

```
## [1] "Bre-X Minerals Ltd. has been silent since it said last month that it formed a partnership with t
## [2] "Despite a whirl of rumours and persistent questions that have sent the Canadian mining company's
## [3] "The Calgary-based company that controls one of the world's biggest gold prospects in Indonesia I
```

Alternatively, documents 17 and 18 focus more on Bre-X Minerals Ltd and their financials. We can now perform clustering using the PCA scores obtained above to see if we can uncover additional relationships between the documents.



```
#looking at the 5th cluster, we see that it placed documents similarly
which(pruned_heather == 5)
```

```
## 32 33 34
## 32 33 34
```

```
#same with 4th cluster
which(pruned_heather == 4)
```

```
## 17 18
## 17 18
```

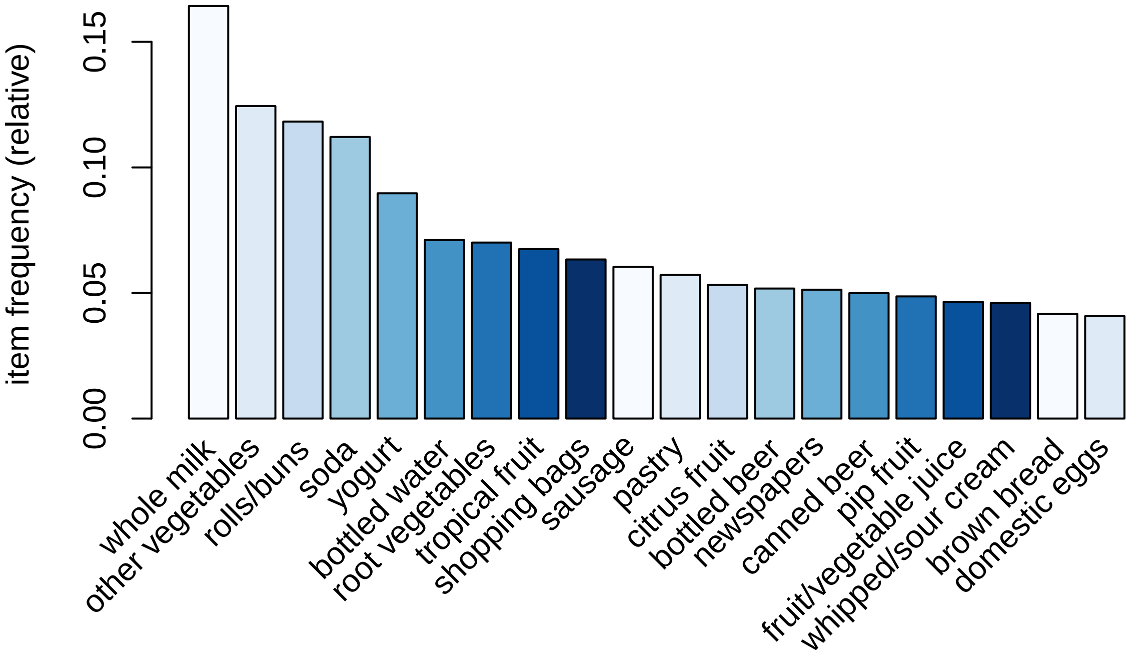
Clustering has resulted in similar groupings to what we saw in PCA. Documents 32, 33, and 34 are grouped together, and so are documents 17 and 18 (as expected). Although the results appear to be essentially the same as before, examining a dendrogram can be a simpler way to see how the documents are related.

Conclusion

In conclusion, it's clear that PCA and clustering are optimal for organizing the documents within the corpus. PCA allowed us to summarize each document into a single pair of numbers while still maintaining a majority of the variation in the dataset. While it may be difficult to understand the contents of these documents from our perspective, it's likely that this grouping can be advantageous for companies looking to organize articles/documents regarding their performance and business deals. This example demonstrates how we can take the composition of words within a document and use them to make informed decisions.

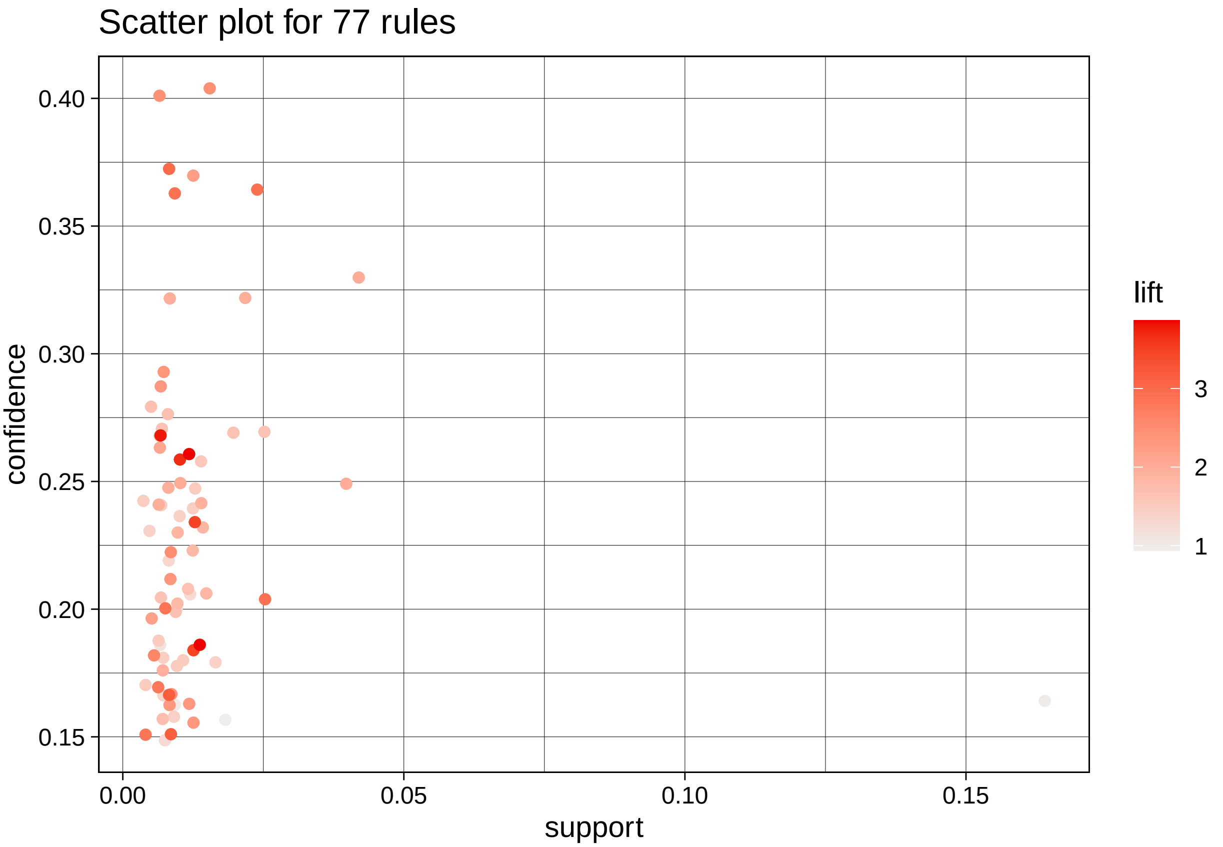
Question 9: Association Rule Mining

Using data regarding different customer's baskets at the grocery store allows us to predict what will be purchased by future customers. We began by reading in the dataset and performing data wrangling to get it in the correct format. This was done by creating a new ID variable for each customer, then pivoting longer to get each customer's basket items into their own observation. We can now split the data into baskets organized by customer and prepare to feed it into the Apriori algorithm. First, let's take a look at the most frequent cart items within the network.

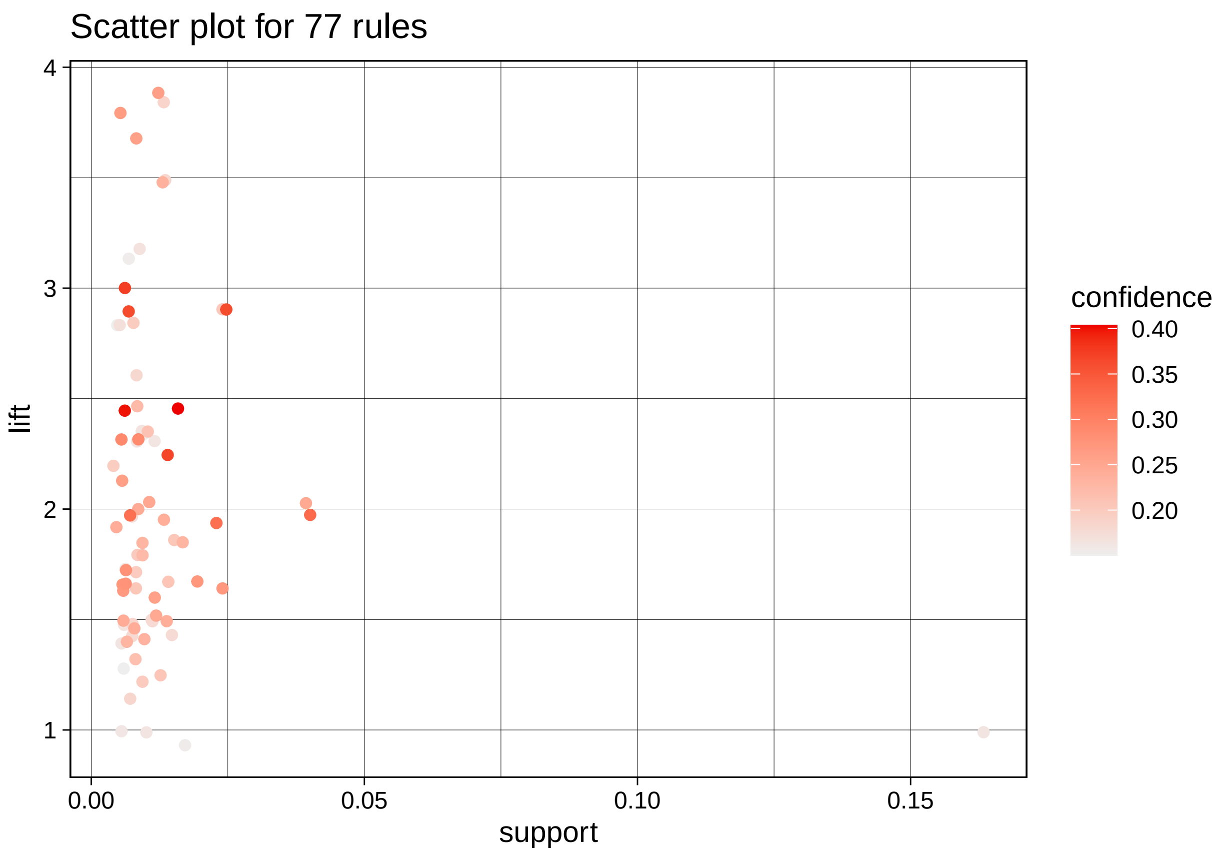


We can see that the item most commonly in our customer's baskets is whole milk, followed by other vegetables, rolls/buns, and soda. Now that we can visualize each item's relative frequency, we can now create our association rules. When determining the parameters for the algorithm, we first started with a low threshold for each rule (support >= 0.005, confidence >= 0.15, maxlen = 5). This is to ensure we are only selecting rules that will give us relative information about the network. Seen below is a scatterplot of each generated rule, which helps us visualize the support, confidence, and lift for each rule.

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.15    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 ... [77 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```



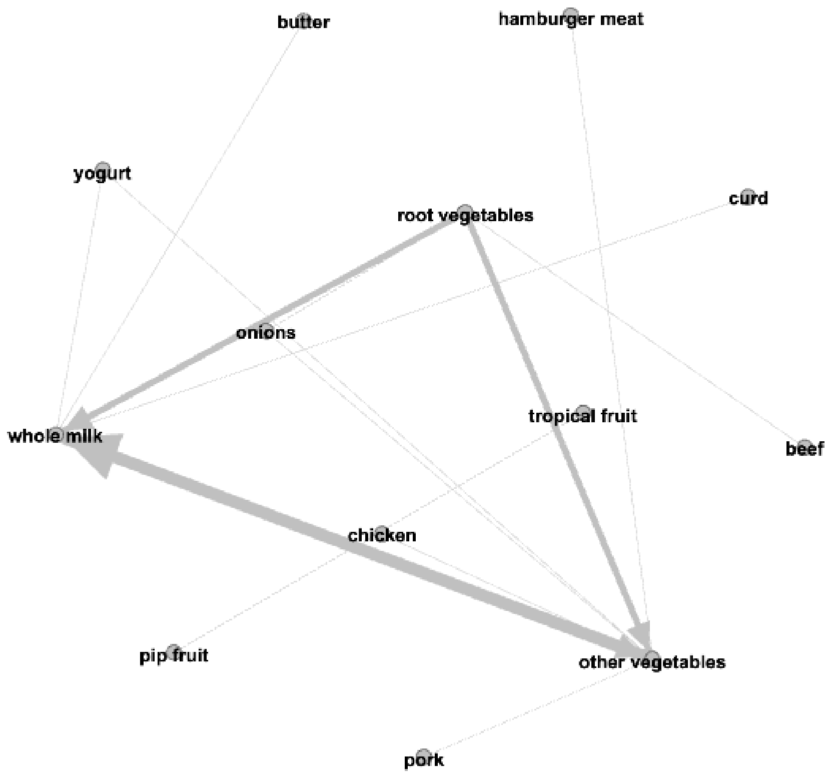
The plot supports the idea that rules with high lift tend to have low support. We put lift on the y-axis to get a different perspective:



This plot helps us choose stricter parameters for the graph we will export to Gephi. We chose a subset of rules with lift greater than 1.75 and confidence greater than 0.25, which filters for the most important rules in the network. The resulting 16 rules can be seen below.

```
##      lhs                                   rhs                support
## [1]  {onions}                           => {root vegetables}  0.005295502
## [2]  {onions}                           => {other vegetables} 0.007452929
## [3]  {hamburger meat}                   => {other vegetables} 0.006210774
## [4]  {chicken}                          => {other vegetables} 0.007975941
## [5]  {beef}                             => {root vegetables}  0.008695084
## [6]  {curd}                             => {whole milk}       0.012617678
## [7]  {butter}                           => {whole milk}       0.014382845
## [8]  {pork}                             => {other vegetables} 0.009283473
## [9]  {pip fruit}                        => {tropical fruit}   0.012683054
## [10] {root vegetables}                  => {other vegetables} 0.025366109
## [11] {root vegetables}                  => {whole milk}       0.022620293
## [12] {other vegetables}                 => {whole milk}       0.040860356
## [13] {other vegetables, root vegetables} => {whole milk}       0.008172071
## [14] {root vegetables, whole milk}      => {other vegetables} 0.008172071
## [15] {other vegetables, yogurt}         => {whole milk}       0.006341527
## [16] {whole milk, yogurt}               => {other vegetables} 0.006341527
##      confidence coverage   lift     count
## [1]  0.2655738  0.01993985 3.789381  81
## [2]  0.3737705  0.01993985 3.004306 114
## [3]  0.2905199  0.02137814 2.335151  95
## [4]  0.2890995  0.02758891 2.323734 122
## [5]  0.2577519  0.03373431 3.677774 133
## [6]  0.3683206  0.03425732 2.241875 193
## [7]  0.4036697  0.03563023 2.457036 220
## [8]  0.2504409  0.03706851 2.013003 142
## [9]  0.2607527  0.04864017 3.864800 194
## [10] 0.3619403  0.07008368 2.909216 388
## [11] 0.3227612  0.07008368 1.964566 346
## [12] 0.3284288  0.12441161 1.999064 625
## [13] 0.3221649  0.02536611 1.960937 125
## [14] 0.3612717  0.02262029 2.903842 125
## [15] 0.3991770  0.01588651 2.429690  97
## [16] 0.2614555  0.02425471 2.101536  97
```

We can see that our highest lift rule is from onions to root vegetables, which suggests that customers who add onions to their cart are more likely to add root vegetables to their basket. Finally, we export the network as a .graphml file so that it can be viewed and interpreted in Gephi.



The resulting network, although appearing quite simple, explains some relationships between the items in the basket of each customer. We can see that customers who add root/other vegetables to their baskets often add whole milk as well. Furthermore, we can see other cold goods such as yogurt and butter that are connected to whole milk. These relationships are easily interpretable since since our network consists of only a few grocery items. We expect customers who buy whole milk to add other cold items next, which is why we see yogurt and butter connected to milk. Additionally, the larger arrows pointing to whole milk leads us to believe that most people begin their shopping trip at the front of the grocery store where produce is often located before adding other items to their cart.

With the addition of more customer's baskets and more grocery items, this network could become much more elaborate and offer further interesting relationships regarding customer's grocery baskets. The reason we see the greatest magnitude of vegetables and milk is likely due to the fact that these are some of the most commonly purchased groceries for all customers.