CSC 466 Lab 2: Apriori Algorithm for Finding Frequent Itemsets and Mining Associations

By Kaanan Kharwa and Laura McGann

Abstract

In this lab, we implemented Apriori's algorithm to mine skyline frequent itemsets and associations in three bakery goods datasets (synthetic) and one book bingo dataset (real). Experimenting with different minimum relative supports yielded 1-2% for the goods datasets and 6.5% for the bingo dataset. Similar experimentation for minimum confidences yielded 95-96% for the goods datasets and 65% for the bingo dataset. For the Bakery data sets, 10 skyline frequent associations were found and for the Fantasy Bingo dataset, 6 skyline frequent associations were found.

Introduction

The purpose of this lab is to understand and practice frequent itemset and association mining via the apriori algorithm. The key to the efficiency of this algorithm is not in the worst-case performance – which is still exponential – but in that it prunes candidate frequent itemsets of length k based on the already calculated supports of itemsets of length k-1, saving time by not considering itemsets containing subsets you already know to be infrequent. We exercised this algorithm on four datasets. The first three datasets are of different sizes (5000, 20000, 75000) but are all synthetic and consist of bakery goods purchased together on different receipts. The fourth dataset is a real dataset collected from 243 individuals who participated in a Fantasy Bingo challenge, reading about 25 books in a year and recording the names of the authors of the books read. We ran Apriori's algorithm over a range of min support values (the threshold for itemsets being considered "frequent") for each dataset to determine which value yielded enough but not an overwhelming amount of skyline frequent itemsets of length >= 2 (needed to search for associations). Upon choosing a minimum relative support, we performed a similar series of tests – iterating over min confidence (the threshold for association existence) – to find association rules among the frequent itemsets. For association mining, we looked only among the *skyline* frequent itemsets and further simplified the problem by only searching for rules with a single item on the right side of the rule. As such, our association mining algorithm is a simple loop iterating over the k-1 subsets of each skyline frequent itemset, using the stored supports to calculate a confidence value for each rule and then comparing that value to the given min confidence.

Commented [LM1]: how much data description here and in the abstract vs in the Data Description section?

Research Questions

Research concerning market baskets, frequent itemsets, and association rule mining concern two main questions, rephrased here for each of the two dataset types:

For the Bakery Datasets:

1. What items *tend* to be frequently bought together?
2. What relationships between frequently bought items exist?

For the Fantasy Bingo Dataset:

3. What authors *tend* to be frequently read together?
4. What relationships between frequently read authors exist?

While sounding similar and even identical, these questions are subtly distinct. Answering question 1 requires a single, concrete, empirical frequency count, but that count cannot necessarily answer question 2, since the frequency relationship may not be symmetric. In other words, X might imply Y while Y does not imply X to the required degree of confidence even if XY is frequent.

As such, we set out to find which itemsets – bakery goods and read authors – occurred frequently and what association rules exist between them.

Dataset Description

We used four datasets. The first three datasets are of different sizes – 5000, 20000, and 75000 data points – but are all synthetic and consist of purchases of bakery goods. The fourth dataset is a real dataset collected from 243 individuals who participated in a Fantasy Bingo challenge, reading about 25 books in a year and recording the names of the authors of the books read. We used a sparse vector representation of each of these datasets, so each data point represented a single receipt or reader, each consisting of a list of ids which correspond to a named bakery good or author. The id-name mapping is provided outside of the main dataset.

Methods

Since we used a sparse vector data representation, we stored each dataset in a list of lists (of ids). Our apriori function took in this data structure, as well as a list of the ids in the item universe. The function stores frequent itemsets in a dictionary (frequency indexed by itemset tuple) which is initialized with the frequent singletons, a subset of the item universe list determined by calculating the supports in the dataset. The function then enters the main loop which generates candidate itemsets of length k from the frequent itemsets of length k-1, eliminates candidates after calculating their supports (reading row by row first to minimize disk reads), then adds the resulting frequent itemsets to the dictionary. During this process, frequent itemsets are flagged as being in the skyline when initially found as frequent and unflagged when one of their supersets is found as frequent. Given a list of frequent skyline itemsets, we found associations among them, calculating confidences based on stored support values and only searching for rules with a single item on the right side of the rule. With our algorithms working, we then tested sets of parameters (min relative support and min confidence) for each of the given datasets.

For the Bakery datasets, to determine which minRSup was optimal, we ran apriori's algorithm with various minRSup values. The values we tested were .005, .006, .007, .008, .009, .01, .02, .03, .04, and .05. We used the total number of itemsets in the skyline for each minRSup, as well as the count of singletons, pairs, triples, quadruples, etc. found as frequent to select an optimal minRSup that yielded enough itemsets of longer length (needed for associations) but not an overwhelming amount. Once we found the best minRSup, we used that value as the minRSup when iterating over minConf. We iterated over minConf in the exact same manner as for minRSup and similarly chose the optimal minConf value by considering the number of associations found. For minConf, the values we iterated over were .9, .91, .92, .93, .94, .95, .96, .97, .98, .99, and 1.0.

We used the exact same approach when calculating the minRSup and minConf values for the Bingo Baskets dataset. For minRSup, we tested 11 values evenly spaced in the range 0.03 to 0.1, inclusive, as well as 0.027, 0.028, and 0.029. For minConf, we tested .6, .65, .7, .75, .8, .85, .9, .95, 1.0.

Results

MinRSup Test Results for Size 5000 Bakery Dataset:

MinRSup Value	**Number of Itemsets in Skyline**
.005	171
.006	64
.007	32
.008	26
.009	26
.01	26
.02	26
.03	35
.04	38
.05	37

As shown above, 26 seems to be the local minimum of the number of frequent itemsets in the skyline. For this reason, .02 was chosen to be the minRSup as it is the maximum value that produces 26 items in the skyline.

MinConf Test Results for Size 5000 Bakery Dataset (Calculated with .02 MinRSup):

MinConf	**Number of Association Rules in Skyline**
.90	19
.91	18
.92	16
.93	16
.94	13
.95	11
.96	11
.97	10
.98	10
.99	10
1.0	8

The number of skyline frequent association rules for the tested range of confidence values seems to hover around 10-11 rules before the count quickly grows to a saturating amount as minimum confidence decreases. However, upon closer inspection of the data, all values seem to have confidence higher than .97 except one which is around .96. Therefore, we removed that one observation and chose .97 as our minConf, yielding 10 association rules.

MinRSup Test Results for Size 20,000 Bakery Dataset:

MinRSup Value	Number of Itemsets in Skyline
.005	98
.006	33
.007	26
.008	26
.009	26
.01	26
.02	26
.03	39
.04	42
.05	37

Once again, 26 seems to be the local minimum for the number of itemsets in the skyline and .02 is the maximum minRSup value that produces this many itemsets which is why we chose it.

MinConf Test Results for Size 20,000 Bakery Dataset (Calculated with .02 MinRSup):

MinConf	Number of Association Rules in Skyline
.90	19
.91	19
.92	16
.93	14
.94	13
.95	10
.96	10
.97	10
.98	10
.99	9
1.0	2

Similar to the 5000 dataset, the number of skyline frequent association rules hovers around 10. We decided to choose the lowest confidence which produces 10 rules in the skyline which ended up being 95%.

MinRSup Test Results for Size 75,000 Bakery Dataset:

MinRSup Value	Number of Item Sets in Skyline
.005	85
.006	28
.007	26
.008	26
.009	26
.01	26
.02	26

.03	38
.04	42
.05	36

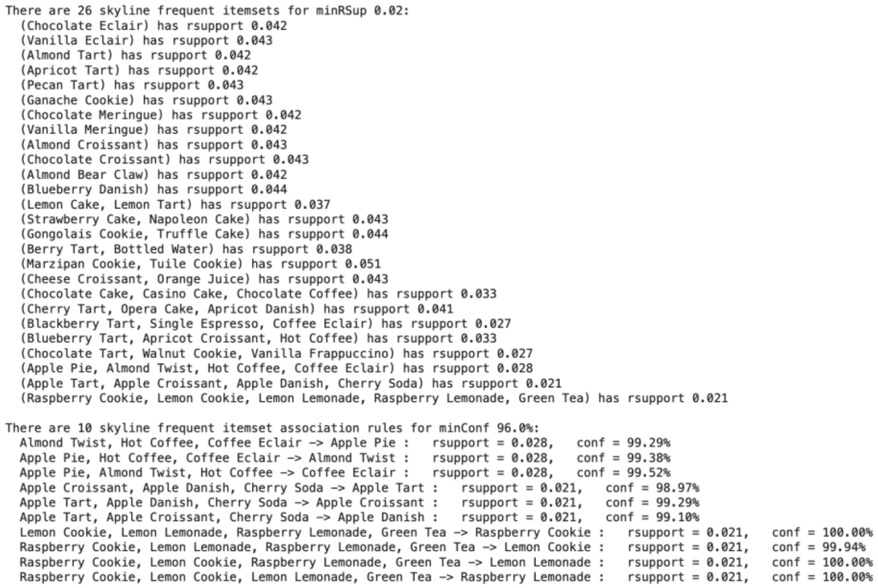
Like the other two Bakery datasets, 26 is the local minimum and .02 was selected as the minRSup since it is the maximum min rsupport at which that number of skyline frequent itemsets is produced.

MinConf Test Results for 75,000 Bakery Dataset (Calculated with .02 MinRSup):

MinConf	**Number of Association Rules in Skyline**
.90	20
.91	20
.92	20
.93	17
.94	13
.95	11
.96	10
.97	10
.98	10
.99	9
1.0	3

The trend we noticed on the 75,000 Bakery dataset was as the confidence increases, the number of association rules in the skyline decreases until it stabilizes around 10 to 11. We initially wanted to go with .95 as it indicated the beginning of the stabilization, but upon further inspection of the data, there was only one rule included by a 95% confidence, so we decided to go with a minConf of .96 instead.

Skyline Frequent Itemsets and Skyline Frequent Association Rules for Size 75,000 Bakery Dataset:



```
There are 26 skyline frequent itemsets for minRSup 0.02:
  (Chocolate Eclair) has rsupport 0.042
  (Vanilla Eclair) has rsupport 0.043
  (Almond Tart) has rsupport 0.042
  (Apricot Tart) has rsupport 0.042
  (Pecan Tart) has rsupport 0.043
  (Ganache Cookie) has rsupport 0.043
  (Chocolate Meringue) has rsupport 0.042
  (Vanilla Meringue) has rsupport 0.042
  (Almond Croissant) has rsupport 0.043
  (Chocolate Croissant) has rsupport 0.043
  (Almond Bear Claw) has rsupport 0.042
  (Blueberry Danish) has rsupport 0.044
  (Lemon Cake, Lemon Tart) has rsupport 0.037
  (Strawberry Cake, Napoleon Cake) has rsupport 0.043
  (Gongolais Cookie, Truffle Cake) has rsupport 0.044
  (Berry Tart, Bottled Water) has rsupport 0.038
  (Marzipan Cookie, Tuile Cookie) has rsupport 0.051
  (Cheese Croissant, Orange Juice) has rsupport 0.043
  (Chocolate Cake, Casino Cake, Chocolate Coffee) has rsupport 0.033
  (Cherry Tart, Opera Cake, Apricot Danish) has rsupport 0.041
  (Blackberry Tart, Single Espresso, Coffee Eclair) has rsupport 0.027
  (Blueberry Tart, Apricot Croissant, Hot Coffee) has rsupport 0.033
  (Chocolate Tart, Walnut Cookie, Vanilla Frappuccino) has rsupport 0.027
  (Apple Pie, Almond Twist, Hot Coffee, Coffee Eclair) has rsupport 0.028
  (Apple Tart, Apple Croissant, Apple Danish, Cherry Soda) has rsupport 0.021
  (Raspberry Cookie, Lemon Cookie, Lemon Lemonade, Raspberry Lemonade, Green Tea) has rsupport 0.021

There are 10 skyline frequent itemset association rules for minConf 96.0%:
  Almond Twist, Hot Coffee, Coffee Eclair -> Apple Pie :   rsupport = 0.028,   conf = 99.29%
  Apple Pie, Hot Coffee, Coffee Eclair -> Almond Twist :   rsupport = 0.028,   conf = 99.38%
  Apple Pie, Almond Twist, Hot Coffee -> Coffee Eclair :   rsupport = 0.028,   conf = 99.52%
  Apple Croissant, Apple Danish, Cherry Soda -> Apple Tart :   rsupport = 0.021,   conf = 98.97%
  Apple Tart, Apple Danish, Cherry Soda -> Apple Croissant :   rsupport = 0.021,   conf = 99.29%
  Apple Tart, Apple Croissant, Cherry Soda -> Apple Danish :   rsupport = 0.021,   conf = 99.10%
  Lemon Cookie, Lemon Lemonade, Raspberry Lemonade, Green Tea -> Raspberry Cookie :   rsupport = 0.021,   conf = 100.00%
  Raspberry Cookie, Lemon Lemonade, Raspberry Lemonade, Green Tea -> Lemon Cookie :   rsupport = 0.021,   conf = 99.94%
  Raspberry Cookie, Lemon Cookie, Raspberry Lemonade, Green Tea -> Lemon Lemonade :   rsupport = 0.021,   conf = 100.00%
  Raspberry Cookie, Lemon Cookie, Lemon Lemonade, Green Tea -> Raspberry Lemonade :   rsupport = 0.021,   conf = 100.00%
```

MinRSup Test Results for Fantasy Bingo Dataset:

MinRSup	**Number of Itemsets in Skyline**	**Difference From Previous**
.027	1353	-
.028	1353	0
.029	1022	321
.03	1022	0
.037	744	333
.044	489	255
.051	329	160
.058	229	100
.065	200	29
.072	162	38
.079	127	45
.086	116	9
.093	92	24
.100	69	23

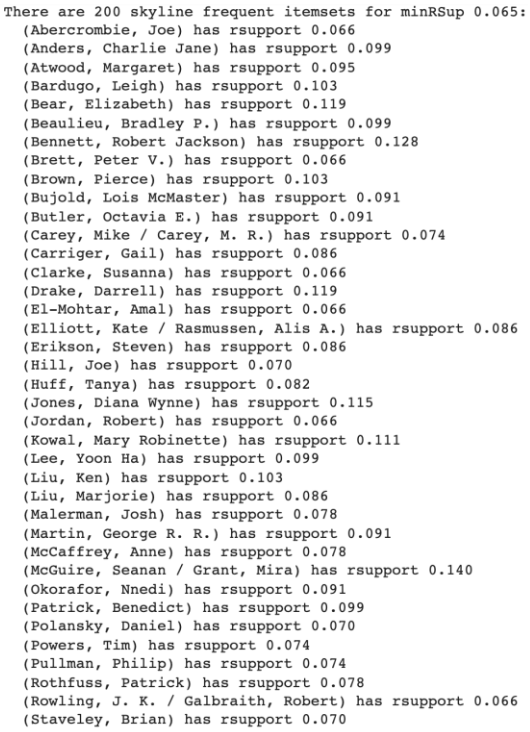
We noticed as the values for MinRSup increase, the difference in the number of itemsets in the skyline begins to decrease then stabilize around the values between 20 and 30. We decided to choose .065 for our MinRSup because it is the first value where this stabilization occurs.

MinConf Test Results for Fantasy Bingo Dataset (Calculated with .065 MinRSup):

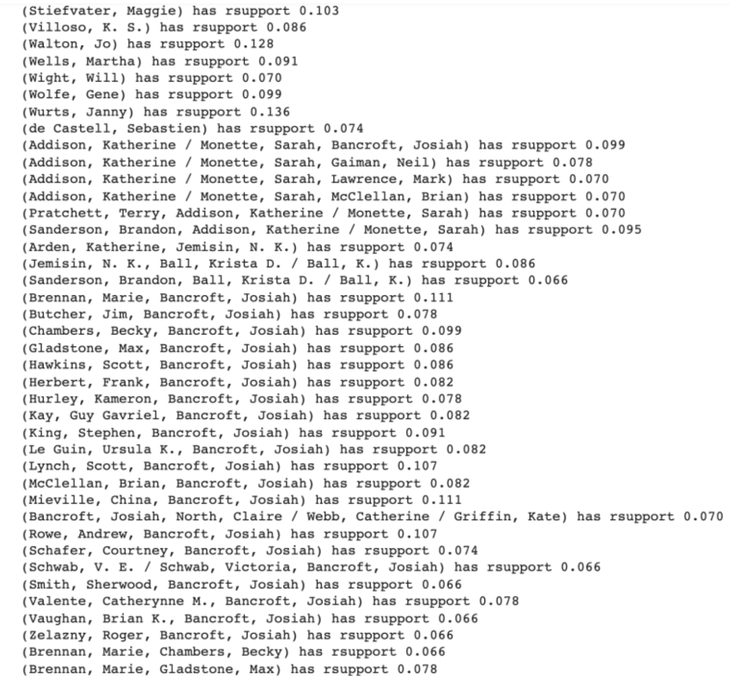
MinConf	Number of Association Rules in Skyline
.60	19
.65	6
.70	2
.75	1
.80	0
.85	0
.90	0
.95	0
1.0	0

The trend visible is that the number of rules declines with increasing minConf until it stabilizes around 1-2. However, with further inspection of the data, we noticed that an association between two items was left out in the .7 minConf. Namely, when using 0.65 as minConf instead of 0.65, we noticed 3 rules that elaborated on the relationship between authors Mark Lawrence (on the left) and Josiah Bancroft (on the right). We thought the occurrence of this association and the slight extra detail was interesting enough to include, so we decided to go with .65 as our minConf.

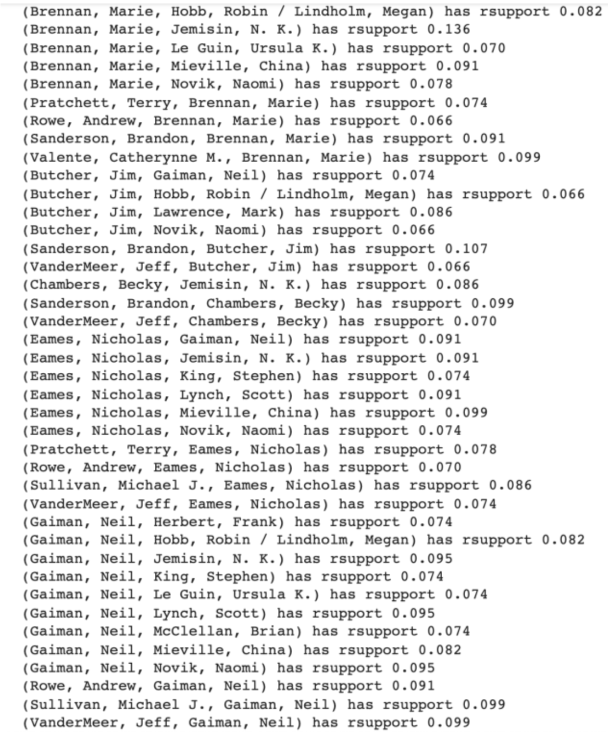
Skyline Frequent Itemsets and Skyline Frequent Association Rules for Fantasy Bingo Dataset:



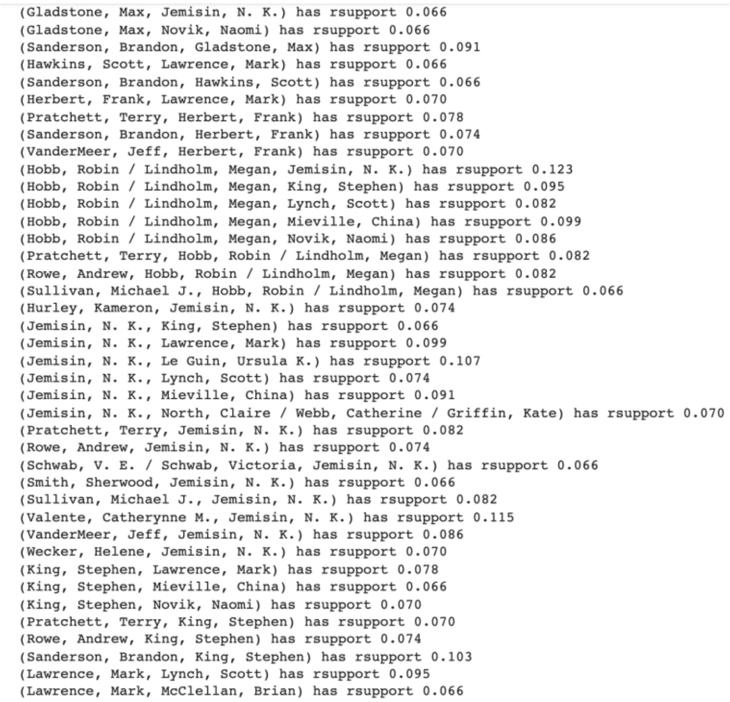
```
There are 200 skyline frequent itemsets for minRSup 0.065:
  (Abercrombie, Joe) has rsupport 0.066
  (Anders, Charlie Jane) has rsupport 0.099
  (Atwood, Margaret) has rsupport 0.095
  (Bardugo, Leigh) has rsupport 0.103
  (Bear, Elizabeth) has rsupport 0.119
  (Beaulieu, Bradley P.) has rsupport 0.099
  (Bennett, Robert Jackson) has rsupport 0.128
  (Brett, Peter V.) has rsupport 0.066
  (Brown, Pierce) has rsupport 0.103
  (Bujold, Lois McMaster) has rsupport 0.091
  (Butler, Octavia E.) has rsupport 0.091
  (Carey, Mike / Carey, M. R.) has rsupport 0.074
  (Carriger, Gail) has rsupport 0.086
  (Clarke, Susanna) has rsupport 0.066
  (Drake, Darrell) has rsupport 0.119
  (El-Mohtar, Amal) has rsupport 0.066
  (Elliott, Kate / Rasmussen, Alis A.) has rsupport 0.086
  (Erikson, Steven) has rsupport 0.086
  (Hill, Joe) has rsupport 0.070
  (Huff, Tanya) has rsupport 0.082
  (Jones, Diana Wynne) has rsupport 0.115
  (Jordan, Robert) has rsupport 0.066
  (Kowal, Mary Robinette) has rsupport 0.111
  (Lee, Yoon Ha) has rsupport 0.099
  (Liu, Ken) has rsupport 0.103
  (Liu, Marjorie) has rsupport 0.086
  (Malerman, Josh) has rsupport 0.078
  (Martin, George R. R.) has rsupport 0.091
  (McCaffrey, Anne) has rsupport 0.078
  (McGuire, Seanan / Grant, Mira) has rsupport 0.140
  (Okorafor, Nnedi) has rsupport 0.091
  (Patrick, Benedict) has rsupport 0.099
  (Polansky, Daniel) has rsupport 0.070
  (Powers, Tim) has rsupport 0.074
  (Pullman, Philip) has rsupport 0.074
  (Rothfuss, Patrick) has rsupport 0.078
  (Rowling, J. K. / Galbraith, Robert) has rsupport 0.066
  (Staveley, Brian) has rsupport 0.070
```



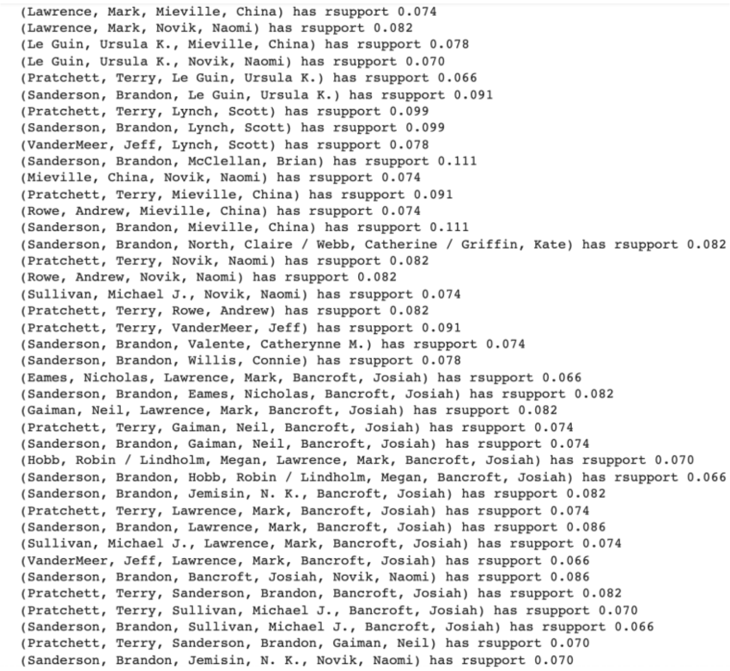
```
(Stiefvater, Maggie) has rsupport 0.103
(Villoso, K. S.) has rsupport 0.086
(Walton, Jo) has rsupport 0.128
(Wells, Martha) has rsupport 0.091
(Wight, Will) has rsupport 0.070
(Wolfe, Gene) has rsupport 0.099
(Wurts, Janny) has rsupport 0.136
(de Castell, Sebastien) has rsupport 0.074
(Addison, Katherine / Monette, Sarah, Bancroft, Josiah) has rsupport 0.099
(Addison, Katherine / Monette, Sarah, Gaiman, Neil) has rsupport 0.078
(Addison, Katherine / Monette, Sarah, Lawrence, Mark) has rsupport 0.070
(Addison, Katherine / Monette, Sarah, McClellan, Brian) has rsupport 0.070
(Pratchett, Terry, Addison, Katherine / Monette, Sarah) has rsupport 0.070
(Sanderson, Brandon, Addison, Katherine / Monette, Sarah) has rsupport 0.095
(Arden, Katherine, Jemisin, N. K.) has rsupport 0.074
(Jemisin, N. K., Ball, Krista D. / Ball, K.) has rsupport 0.086
(Sanderson, Brandon, Ball, Krista D. / Ball, K.) has rsupport 0.066
(Brennan, Marie, Bancroft, Josiah) has rsupport 0.111
(Butcher, Jim, Bancroft, Josiah) has rsupport 0.078
(Chambers, Becky, Bancroft, Josiah) has rsupport 0.099
(Gladstone, Max, Bancroft, Josiah) has rsupport 0.086
(Hawkins, Scott, Bancroft, Josiah) has rsupport 0.086
(Herbert, Frank, Bancroft, Josiah) has rsupport 0.082
(Hurley, Kameron, Bancroft, Josiah) has rsupport 0.078
(Kay, Guy Gavriel, Bancroft, Josiah) has rsupport 0.082
(King, Stephen, Bancroft, Josiah) has rsupport 0.091
(Le Guin, Ursula K., Bancroft, Josiah) has rsupport 0.082
(Lynch, Scott, Bancroft, Josiah) has rsupport 0.107
(McClellan, Brian, Bancroft, Josiah) has rsupport 0.082
(Mieville, China, Bancroft, Josiah) has rsupport 0.111
(Bancroft, Josiah, North, Claire / Webb, Catherine / Griffin, Kate) has rsupport 0.070
(Rowe, Andrew, Bancroft, Josiah) has rsupport 0.107
(Schafer, Courtney, Bancroft, Josiah) has rsupport 0.074
(Schwab, V. E. / Schwab, Victoria, Bancroft, Josiah) has rsupport 0.066
(Smith, Sherwood, Bancroft, Josiah) has rsupport 0.066
(Valente, Catherynne M., Bancroft, Josiah) has rsupport 0.078
(Vaughan, Brian K., Bancroft, Josiah) has rsupport 0.066
(Zelazny, Roger, Bancroft, Josiah) has rsupport 0.066
(Brennan, Marie, Chambers, Becky) has rsupport 0.066
(Brennan, Marie, Gladstone, Max) has rsupport 0.078
```



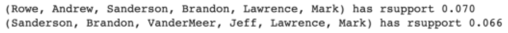
```
(Brennan, Marie, Hobb, Robin / Lindholm, Megan) has rsupport 0.082
(Brennan, Marie, Jemisin, N. K.) has rsupport 0.136
(Brennan, Marie, Le Guin, Ursula K.) has rsupport 0.070
(Brennan, Marie, Mieville, China) has rsupport 0.091
(Brennan, Marie, Novik, Naomi) has rsupport 0.078
(Pratchett, Terry, Brennan, Marie) has rsupport 0.074
(Rowe, Andrew, Brennan, Marie) has rsupport 0.066
(Sanderson, Brandon, Brennan, Marie) has rsupport 0.091
(Valente, Catherynne M., Brennan, Marie) has rsupport 0.099
(Butcher, Jim, Gaiman, Neil) has rsupport 0.074
(Butcher, Jim, Hobb, Robin / Lindholm, Megan) has rsupport 0.066
(Butcher, Jim, Lawrence, Mark) has rsupport 0.086
(Butcher, Jim, Novik, Naomi) has rsupport 0.066
(Sanderson, Brandon, Butcher, Jim) has rsupport 0.107
(VanderMeer, Jeff, Butcher, Jim) has rsupport 0.066
(Chambers, Becky, Jemisin, N. K.) has rsupport 0.086
(Sanderson, Brandon, Chambers, Becky) has rsupport 0.099
(VanderMeer, Jeff, Chambers, Becky) has rsupport 0.070
(Eames, Nicholas, Gaiman, Neil) has rsupport 0.091
(Eames, Nicholas, Jemisin, N. K.) has rsupport 0.091
(Eames, Nicholas, King, Stephen) has rsupport 0.074
(Eames, Nicholas, Lynch, Scott) has rsupport 0.091
(Eames, Nicholas, Mieville, China) has rsupport 0.099
(Eames, Nicholas, Novik, Naomi) has rsupport 0.074
(Pratchett, Terry, Eames, Nicholas) has rsupport 0.078
(Rowe, Andrew, Eames, Nicholas) has rsupport 0.070
(Sullivan, Michael J., Eames, Nicholas) has rsupport 0.086
(VanderMeer, Jeff, Eames, Nicholas) has rsupport 0.074
(Gaiman, Neil, Herbert, Frank) has rsupport 0.074
(Gaiman, Neil, Hobb, Robin / Lindholm, Megan) has rsupport 0.082
(Gaiman, Neil, Jemisin, N. K.) has rsupport 0.095
(Gaiman, Neil, King, Stephen) has rsupport 0.074
(Gaiman, Neil, Le Guin, Ursula K.) has rsupport 0.074
(Gaiman, Neil, Lynch, Scott) has rsupport 0.095
(Gaiman, Neil, McClellan, Brian) has rsupport 0.074
(Gaiman, Neil, Mieville, China) has rsupport 0.082
(Gaiman, Neil, Novik, Naomi) has rsupport 0.095
(Rowe, Andrew, Gaiman, Neil) has rsupport 0.091
(Sullivan, Michael J., Gaiman, Neil) has rsupport 0.099
(VanderMeer, Jeff, Gaiman, Neil) has rsupport 0.099
```



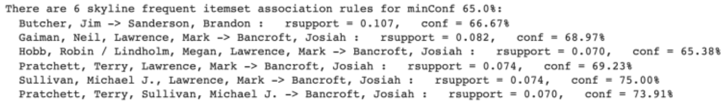
```
(Gladstone, Max, Jemisin, N. K.) has rsupport 0.066
(Gladstone, Max, Novik, Naomi) has rsupport 0.066
(Sanderson, Brandon, Gladstone, Max) has rsupport 0.091
(Hawkins, Scott, Lawrence, Mark) has rsupport 0.066
(Sanderson, Brandon, Hawkins, Scott) has rsupport 0.066
(Herbert, Frank, Lawrence, Mark) has rsupport 0.070
(Pratchett, Terry, Herbert, Frank) has rsupport 0.078
(Sanderson, Brandon, Herbert, Frank) has rsupport 0.074
(VanderMeer, Jeff, Herbert, Frank) has rsupport 0.070
(Hobb, Robin / Lindholm, Megan, Jemisin, N. K.) has rsupport 0.123
(Hobb, Robin / Lindholm, Megan, King, Stephen) has rsupport 0.095
(Hobb, Robin / Lindholm, Megan, Lynch, Scott) has rsupport 0.082
(Hobb, Robin / Lindholm, Megan, Mieville, China) has rsupport 0.099
(Hobb, Robin / Lindholm, Megan, Novik, Naomi) has rsupport 0.086
(Pratchett, Terry, Hobb, Robin / Lindholm, Megan) has rsupport 0.082
(Rowe, Andrew, Hobb, Robin / Lindholm, Megan) has rsupport 0.082
(Sullivan, Michael J., Hobb, Robin / Lindholm, Megan) has rsupport 0.066
(Hurley, Kameron, Jemisin, N. K.) has rsupport 0.074
(Jemisin, N. K., King, Stephen) has rsupport 0.066
(Jemisin, N. K., Lawrence, Mark) has rsupport 0.099
(Jemisin, N. K., Le Guin, Ursula K.) has rsupport 0.107
(Jemisin, N. K., Lynch, Scott) has rsupport 0.074
(Jemisin, N. K., Mieville, China) has rsupport 0.091
(Jemisin, N. K., North, Claire / Webb, Catherine / Griffin, Kate) has rsupport 0.070
(Pratchett, Terry, Jemisin, N. K.) has rsupport 0.082
(Rowe, Andrew, Jemisin, N. K.) has rsupport 0.074
(Schwab, V. E. / Schwab, Victoria, Jemisin, N. K.) has rsupport 0.066
(Smith, Sherwood, Jemisin, N. K.) has rsupport 0.066
(Sullivan, Michael J., Jemisin, N. K.) has rsupport 0.082
(Valente, Catherynne M., Jemisin, N. K.) has rsupport 0.115
(VanderMeer, Jeff, Jemisin, N. K.) has rsupport 0.086
(Wecker, Helene, Jemisin, N. K.) has rsupport 0.070
(King, Stephen, Lawrence, Mark) has rsupport 0.078
(King, Stephen, Mieville, China) has rsupport 0.066
(King, Stephen, Novik, Naomi) has rsupport 0.070
(Pratchett, Terry, King, Stephen) has rsupport 0.070
(Rowe, Andrew, King, Stephen) has rsupport 0.074
(Sanderson, Brandon, King, Stephen) has rsupport 0.103
(Lawrence, Mark, Lynch, Scott) has rsupport 0.095
(Lawrence, Mark, McClellan, Brian) has rsupport 0.066
```



```
(Lawrence, Mark, Mieville, China) has rsupport 0.074
(Lawrence, Mark, Novik, Naomi) has rsupport 0.082
(Le Guin, Ursula K., Mieville, China) has rsupport 0.078
(Le Guin, Ursula K., Novik, Naomi) has rsupport 0.070
(Pratchett, Terry, Le Guin, Ursula K.) has rsupport 0.066
(Sanderson, Brandon, Le Guin, Ursula K.) has rsupport 0.091
(Pratchett, Terry, Lynch, Scott) has rsupport 0.099
(Sanderson, Brandon, Lynch, Scott) has rsupport 0.099
(VanderMeer, Jeff, Lynch, Scott) has rsupport 0.078
(Sanderson, Brandon, McClellan, Brian) has rsupport 0.111
(Mieville, China, Novik, Naomi) has rsupport 0.074
(Pratchett, Terry, Mieville, China) has rsupport 0.091
(Rowe, Andrew, Mieville, China) has rsupport 0.074
(Sanderson, Brandon, Mieville, China) has rsupport 0.111
(Sanderson, Brandon, North, Claire / Webb, Catherine / Griffin, Kate) has rsupport 0.082
(Pratchett, Terry, Novik, Naomi) has rsupport 0.082
(Rowe, Andrew, Novik, Naomi) has rsupport 0.082
(Sullivan, Michael J., Novik, Naomi) has rsupport 0.074
(Pratchett, Terry, Rowe, Andrew) has rsupport 0.082
(Pratchett, Terry, VanderMeer, Jeff) has rsupport 0.091
(Sanderson, Brandon, Valente, Catherynne M.) has rsupport 0.074
(Sanderson, Brandon, Willis, Connie) has rsupport 0.078
(Eames, Nicholas, Lawrence, Mark, Bancroft, Josiah) has rsupport 0.066
(Sanderson, Brandon, Eames, Nicholas, Bancroft, Josiah) has rsupport 0.082
(Gaiman, Neil, Lawrence, Mark, Bancroft, Josiah) has rsupport 0.082
(Pratchett, Terry, Gaiman, Neil, Bancroft, Josiah) has rsupport 0.074
(Sanderson, Brandon, Gaiman, Neil, Bancroft, Josiah) has rsupport 0.074
(Hobb, Robin / Lindholm, Megan, Lawrence, Mark, Bancroft, Josiah) has rsupport 0.070
(Sanderson, Brandon, Hobb, Robin / Lindholm, Megan, Bancroft, Josiah) has rsupport 0.066
(Sanderson, Brandon, Jemisin, N. K., Bancroft, Josiah) has rsupport 0.082
(Pratchett, Terry, Lawrence, Mark, Bancroft, Josiah) has rsupport 0.074
(Sanderson, Brandon, Lawrence, Mark, Bancroft, Josiah) has rsupport 0.086
(Sullivan, Michael J., Lawrence, Mark, Bancroft, Josiah) has rsupport 0.074
(VanderMeer, Jeff, Lawrence, Mark, Bancroft, Josiah) has rsupport 0.066
(Sanderson, Brandon, Bancroft, Josiah, Novik, Naomi) has rsupport 0.086
(Pratchett, Terry, Sanderson, Brandon, Bancroft, Josiah) has rsupport 0.082
(Pratchett, Terry, Sullivan, Michael J., Bancroft, Josiah) has rsupport 0.070
(Sanderson, Brandon, Sullivan, Michael J., Bancroft, Josiah) has rsupport 0.066
(Pratchett, Terry, Sanderson, Brandon, Gaiman, Neil) has rsupport 0.070
(Sanderson, Brandon, Jemisin, N. K., Novik, Naomi) has rsupport 0.070
```



```
(Rowe, Andrew, Sanderson, Brandon, Lawrence, Mark) has rsupport 0.070
(Sanderson, Brandon, VanderMeer, Jeff, Lawrence, Mark) has rsupport 0.066
```



```
There are 6 skyline frequent itemset association rules for minConf 65.0%:
  Butcher, Jim -> Sanderson, Brandon :   rsupport = 0.107,   conf = 66.67%
  Gaiman, Neil, Lawrence, Mark -> Bancroft, Josiah :   rsupport = 0.082,   conf = 68.97%
  Hobb, Robin / Lindholm, Megan, Lawrence, Mark -> Bancroft, Josiah :   rsupport = 0.070,   conf = 65.38%
  Pratchett, Terry, Lawrence, Mark -> Bancroft, Josiah :   rsupport = 0.074,   conf = 69.23%
  Sullivan, Michael J., Lawrence, Mark -> Bancroft, Josiah :   rsupport = 0.074,   conf = 75.00%
  Pratchett, Terry, Sullivan, Michael J. -> Bancroft, Josiah :   rsupport = 0.070,   conf = 73.91%
```

Final Parameter Choices

Dataset	**minRSup**	**minConf**	**# Association Rules Induced**
Goods 5000	0.02	0.95	10
Goods 20000	0.02	0.95	10
Goods 75000	0.02	0.97	10
Bingo	.065	.65	6

Discussion and Conclusion

Bakery Dataset

In the Bakery dataset, there are three groups of items that tend to be bought together. The first group consists of Almond Twist, Hot Coffee, Apple Pie, and Coffee Éclair. Hot Coffee never appears on the right side whereas all the pastries do. This shows that buying all the pastries does not imply that a customer will buy a coffee, rather that if a customer has all but one of the pastries and a coffee, they will buy the missing pastry. This pattern appears again in the second group which consists of the Apple Tart, Apple Danish, Apple Croissant, and Cherry Soda. Like with the previous group, all pastries appear on the right-hand side, but the drink, in this case Cherry Soda, only appears on the left-hand side. This similarly implies that if a customer has two out of the three pastries, and a drink (Cherry Soda), they are probably going to buy the third pastry. Once more, this pattern is visible in the third group which consists of a Raspberry Cookie, Raspberry Lemonade, Lemon Cookie, and Green Tea. All the items in the group except for Green Tea appear on the right-hand side, once again implying that if a customer has bought three out of the four items in the group, one of which is a Green Tea, they will probably buy the fourth pastry.

Fantasy Bingo Dataset

Based on the found associations in the Fantasy Bingo dataset, we see readers often read books by Josiah Bancroft as a result of reading combinations of books by Mark Lawrence, Michael J. Sullivan, and Terry Pratchett. Since Bancroft only appears on the right side of found associations not the left, we can infer he is not a well-known author for readers entering the genre of his writing. Rather, Lawrence, Sullivan and Pratchett may be well-known authors and are thus the first authors people hear of and read when getting into the genre. Upon liking those authors, readers then begin searching for similar authors, at which point Bancroft is discovered. There is one more association rule present in the Fantasy Bingo set which is between Jim Butcher and Bandon Sanderson. Since these names only appear once, it is hard to say much about this rule. However, since Brandon Sanderson only appears on the right, we can imply that readers began to read his work after reading Jim Butcher's work, implying that they are similar stories or in the same genre, but Butcher may be more well-known.