CS 634 Data Mining Midterm Project

APRIORI ALGORITHM

Submitted by: Sowmya Kothapalli

UCID: sk272

Email: sk272@njit.edu

Date: 03/28/2021

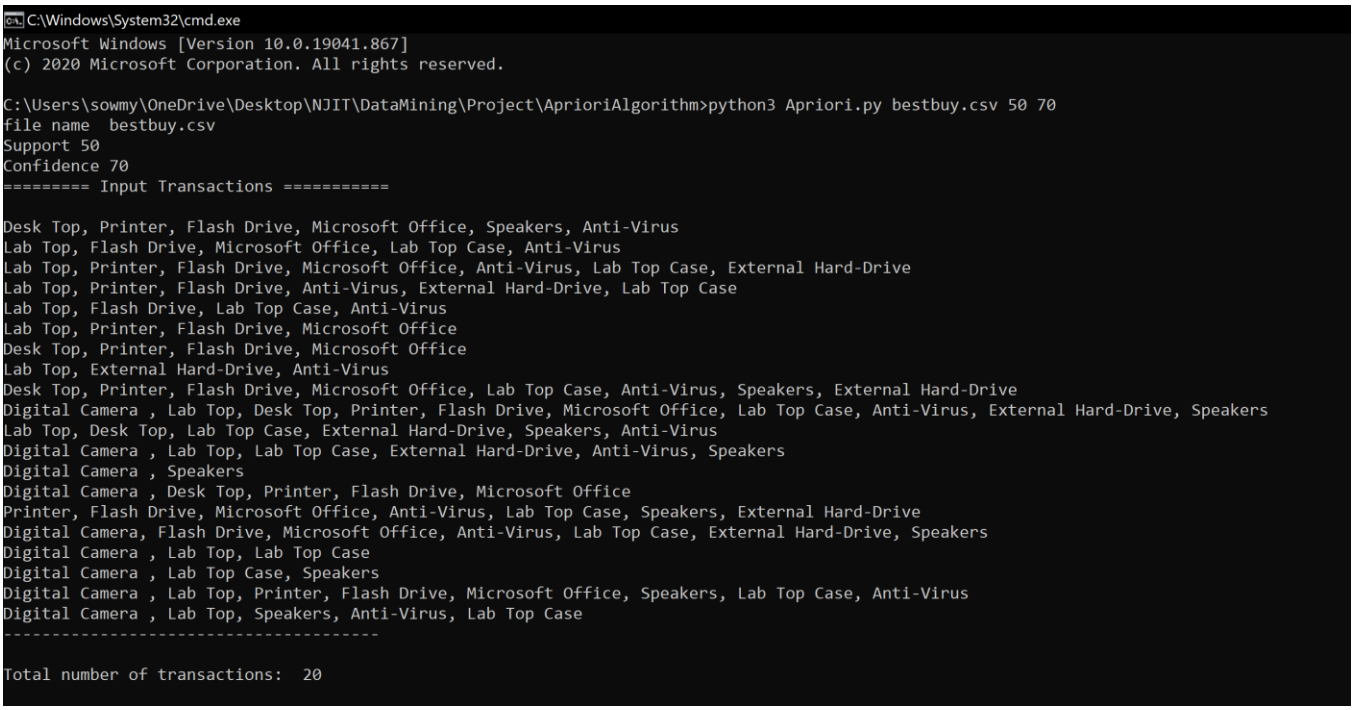
Programming Language: Python

How to run the file:

python Apriori.py datasetname {minimum support (in percentage)} {minimum confidence (in percentage)}

E.g.: python3 Apriori.py bestbuy.csv 50 70

Input: For dataset Best buy



Output: For dataset Best buy

```
C:\Windows\System32\cmd.exe
Total number of transactions:  20

Frequent sets for k= 1
+------------------+-----------+
|    Item sets     | Frequency |
+------------------+-----------+
|     Printer      |     10    |
|   Flash Drive    |     13    |
| Microsoft Office |     11    |
|     Speakers     |     11    |
|    Anti-Virus    |     14    |
|     Lab Top      |     12    |
|   Lab Top Case   |     14    |
+------------------+-----------+

Frequent set for k= 2
+-------------------------------------+-----------+
|              Item sets              | Frequency |
+-------------------------------------+-----------+
|      ('Printer', 'Flash Drive')     |     10    |
| ('Flash Drive', 'Microsoft Office') |     11    |
|    ('Flash Drive', 'Anti-Virus')    |     10    |
|      ('Anti-Virus', 'Lab Top')      |     10    |
|    ('Anti-Virus', 'Lab Top Case')   |     12    |
|     ('Lab Top', 'Lab Top Case')     |     10    |
+-------------------------------------+-----------+

========== List of all frequent item sets and support levels ============
+-------------------------------------+----------------+
|          Frequent Item sets         | Support in (%) |
+-------------------------------------+----------------+
|      ('Printer', 'Flash Drive')     |       50       |
| ('Flash Drive', 'Microsoft Office') |       55       |
|    ('Flash Drive', 'Anti-Virus')    |       50       |
|      ('Anti-Virus', 'Lab Top')      |       50       |
|    ('Anti-Virus', 'Lab Top Case')   |       60       |
|     ('Lab Top', 'Lab Top Case')     |       50       |
+-------------------------------------+----------------+
```

```
========== Association and Confidence levels ==============
+-------------------------------------+----------------------+----------------------+---------------+------------------+
|           Selected sets             |      Predecessor     |        Result        | Support in (%) | Confidence in (%) |
+-------------------------------------+----------------------+----------------------+---------------+------------------+
|      ('Printer', 'Flash Drive')     |      ('Printer',)    |    ('Flash Drive',)  |       50      |        100       |
|      ('Flash Drive', 'Printer')     |    ('Flash Drive',)  |      ('Printer',)    |       50      |         77       |
| ('Flash Drive', 'Microsoft Office') |    ('Flash Drive',)  | ('Microsoft Office',) |       55      |         85       |
| ('Microsoft Office', 'Flash Drive') | ('Microsoft Office',) |    ('Flash Drive',)  |       55      |        100       |
|    ('Flash Drive', 'Anti-Virus')    |    ('Flash Drive',)  |    ('Anti-Virus',)   |       50      |         77       |
|    ('Anti-Virus', 'Flash Drive')    |    ('Anti-Virus',)   |    ('Flash Drive',)  |       50      |         71       |
|      ('Anti-Virus', 'Lab Top')      |    ('Anti-Virus',)   |      ('Lab Top',)    |       50      |         71       |
|      ('Lab Top', 'Anti-Virus')      |      ('Lab Top',)    |    ('Anti-Virus',)   |       50      |         83       |
|    ('Anti-Virus', 'Lab Top Case')   |    ('Anti-Virus',)   |   ('Lab Top Case',)  |       60      |         86       |
|    ('Lab Top Case', 'Anti-Virus')   |   ('Lab Top Case',)  |    ('Anti-Virus',)   |       60      |         86       |
|     ('Lab Top', 'Lab Top Case')     |      ('Lab Top',)    |   ('Lab Top Case',)  |       50      |         83       |
|     ('Lab Top Case', 'Lab Top')     |   ('Lab Top Case',)  |      ('Lab Top',)    |       50      |         71       |
+-------------------------------------+----------------------+----------------------+---------------+------------------+

C:\Users\sowmy\OneDrive\Desktop\NJIT\DataMining\Project\AprioriAlgorithm>
```

Input: For dataset Nike

```
C:\Windows\System32\cmd.exe

C:\Users\sowmy\OneDrive\Desktop\NJIT\DataMining\Project\AprioriAlgorithm>python3 Apriori.py nike.csv 60 80
file name  nike.csv
Support 60
Confidence 80
========= Input Transactions ==========

Running Shoe, Socks, Sweatshirts, Modern Pants
Running Shoe, Socks, Sweatshirts
Running Shoe, Socks, Sweatshirts, Modern Pants
Running Shoe, Sweatshirts, Modern Pants
Running Shoe, Socks, Sweatshirts, Modern Pants, Soccer Shoe
Running Shoe, Socks, Sweatshirts
Running Shoe, Socks, Sweatshirts, Modern Pants, Tech Pants, Rash Guard, Hoodies
Swimming Shirt, Socks, Sweatshirts
Swimming Shirt, Rash Guard, Dry Fit V-Nick, Hoodies, Tech Pants
Swimming Shirt, Rash Guard, Dry,
Swimming Shirt, Rash Guard, Dry Fit V-Nick
Running Shoe, Swimming Shirt, Socks, Sweatshirts, Modern Pants, Soccer Shoe, Rash Guard, Hoodies, Tech Pants, Dry Fit V-Nick
Running Shoe, Swimming Shirt, Socks, Sweatshirts, Modern Pants, Soccer Shoe, Rash Guard, Tech Pants, Dry Fit V-Nick, Hoodies
Running Shoe, Swimming Shirt, Rash Guard, Tech Pants, Hoodies, Dry Fit V-Nick
Running Shoe, Swimming Shirt, Socks, Sweatshirts, Modern Pants, Dry Fit V-Nick, Rash Guard, Tech Pants
Swimming Shirt, Soccer Shoe, Hoodies, Dry Fit V-Nick, Tech Pants, Rash Guard
Running Shoe, Socks
Socks, Sweatshirts, Modern Pants, Soccer Shoe, Hoodies, Rash Guard, Tech Pants, Dry Fit V-Nick
Running Shoe, Swimming Shirt, Rash Guard
Running Shoe, Swimming Shirt, Socks, Sweatshirts, Modern Pants, Soccer Shoe, Hoodies, Tech Pants, Rash Guard, Dry Fit V-Nick
--------------------------------------

Total number of transactions:  20
```

Output: For dataset Nike

```
C:\Windows\System32\cmd.exe
--------------------------------------

Total number of transactions:  20

Frequent sets for k= 1
+--------------+-----------+
|  Item sets   | Frequency |
+--------------+-----------+
| Running Shoe |     14    |
|    Socks     |     13    |
| Sweatshirts  |     13    |
|  Rash Guard  |     12    |
+--------------+-----------+

Frequent set for k= 2
+--------------------------+-----------+
|        Item sets         | Frequency |
+--------------------------+-----------+
| ('Socks', 'Sweatshirts') |     12    |
+--------------------------+-----------+

========== List of all frequent item sets and support levels ============
+--------------------------+----------------+
|    Frequent Item sets    | Support in (%) |
+--------------------------+----------------+
| ('Socks', 'Sweatshirts') |       60       |
+--------------------------+----------------+

========== Association and Confidence levels ==============
+--------------------------+------------------+------------------+----------------+------------------+
|      Selected sets       |   Predecessor    |      Result      | Support in (%) | Confidence in (%) |
+--------------------------+------------------+------------------+----------------+------------------+
| ('Socks', 'Sweatshirts') |    ('Socks',)    | ('Sweatshirts',) |       60       |        92        |
| ('Sweatshirts', 'Socks') | ('Sweatshirts',) |    ('Socks',)    |       60       |        92        |
+--------------------------+------------------+------------------+----------------+------------------+

C:\Users\sowmy\OneDrive\Desktop\NJIT\DataMining\Project\AprioriAlgorithm>
```

Input: For dataset Kmart

```
C:\Windows\System32\cmd.exe

C:\Users\sowmy\OneDrive\Desktop\NJIT\DataMining\Project\AprioriAlgorithm>python3 Apriori.py kmart.csv 45 75
file name  kmart.csv
Support 45
Confidence 75
========= Input Transactions ==========

Decorative Pillows, Quilts, Embroidered Bedspread
Embroidered Bedspread, Shams, Kids Bedding, Bedding Collections, Bed Skirts, Bedspreads, Sheets
Decorative Pillows, Quilts, Embroidered Bedspread, Shams, Kids Bedding, Bedding Collections
Kids Bedding, Bedding Collections, Sheets, Bedspreads, Bed Skirts
Decorative Pillows, Kids Bedding, Bedding Collections, Sheets, Bed Skirts, Bedspreads
Bedding Collections, Bedspreads, Bed Skirts, Sheets, Shams, Kids Bedding
Decorative Pillows, Quilts
Decorative Pillows, Quilts, Embroidered Bedspread
Bedspreads, Bed Skirts, Shams, Kids Bedding, Sheets
Quilts, Embroidered Bedspread, Bedding Collections
Bedding Collections, Bedspreads, Bed Skirts, Kids Bedding, Shams, Sheets
Decorative Pillows, Quilts
Embroidered Bedspread, Shams
Sheets, Shams, Bed Skirts, Kids Bedding
Decorative Pillows, Quilts
Decorative Pillows, Kids Bedding, Bed Skirts, Shams
Decorative Pillows, Shams, Bed Skirts
Quilts, Sheets, Kids Bedding
Shams, Bed Skirts, Kids Bedding, Sheets
Decorative Pillows, Bedspreads, Shams, Sheets, Bed Skirts, Kids Bedding
-------------------------------------

Total number of transactions:  20
```

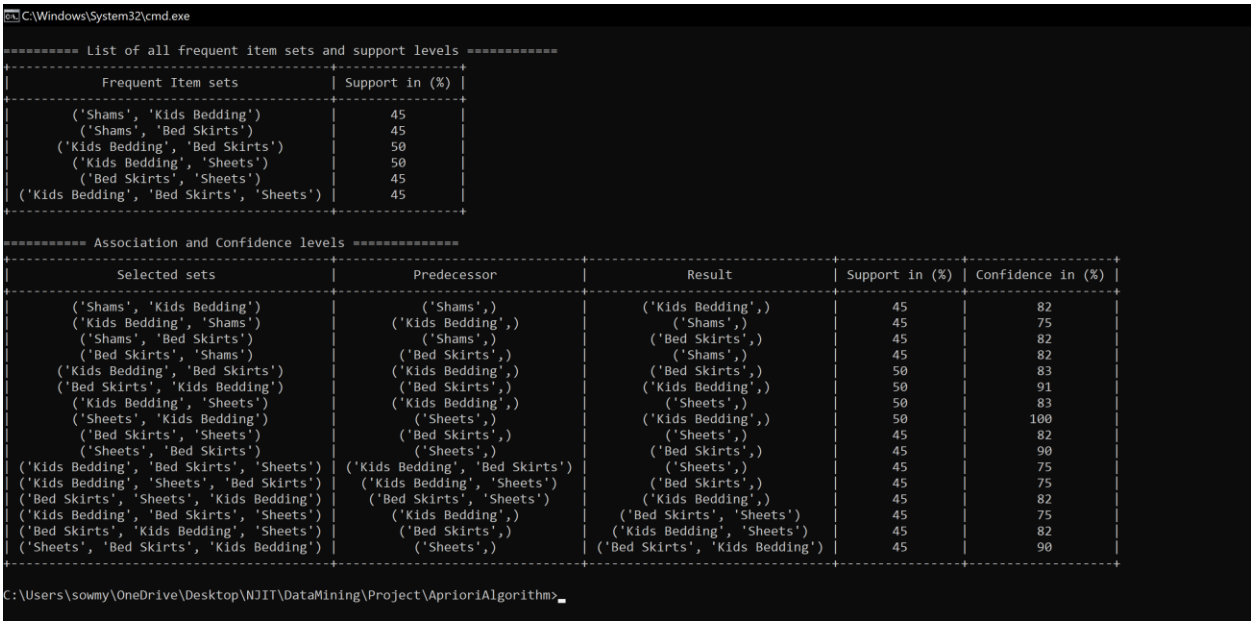
Output: For dataset Kmart

```
C:\Windows\System32\cmd.exe

Frequent sets for k= 1
+--------------------+-----------+
|     Item sets      | Frequency |
+--------------------+-----------+
| Decorative Pillows |     10    |
|       Shams        |     11    |
|    Kids Bedding    |     12    |
|     Bed Skirts     |     11    |
|       Sheets       |     10    |
+--------------------+-----------+

Frequent set for k= 2
+-------------------------------+-----------+
|           Item sets           | Frequency |
+-------------------------------+-----------+
|   ('Shams', 'Kids Bedding')   |     9     |
|    ('Shams', 'Bed Skirts')    |     9     |
| ('Kids Bedding', 'Bed Skirts') |     10    |
|   ('Kids Bedding', 'Sheets')  |     10    |
|    ('Bed Skirts', 'Sheets')   |     9     |
+-------------------------------+-----------+

Frequent set for k= 3
+----------------------------------------+-----------+
|               Item sets                | Frequency |
+----------------------------------------+-----------+
| ('Kids Bedding', 'Bed Skirts', 'Sheets') |     9     |
+----------------------------------------+-----------+
```



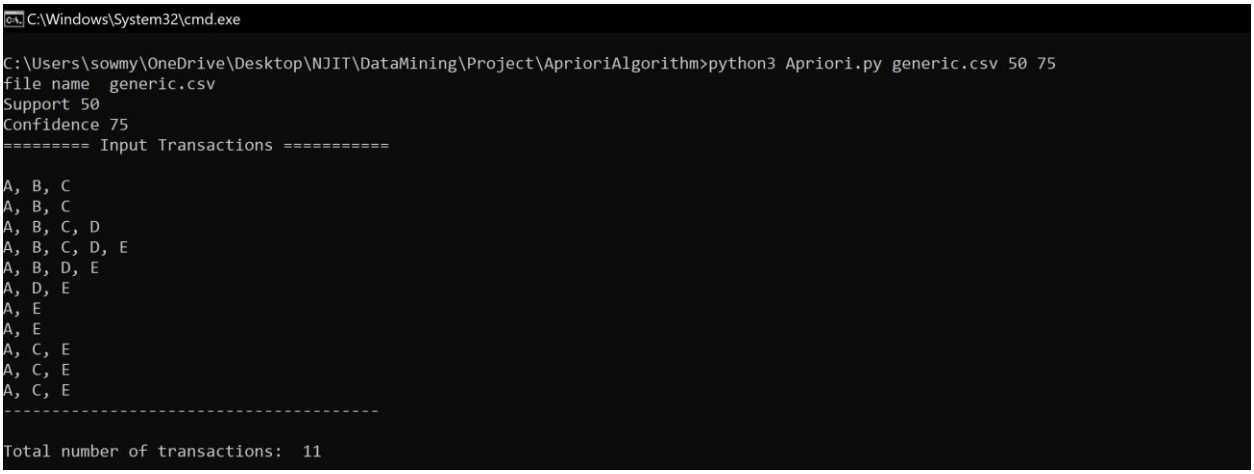
```
C:\Windows\System32\cmd.exe

========== List of all frequent item sets and support levels ============
+----------------------------------------+---------------+
|           Frequent Item sets           | Support in (%) |
+----------------------------------------+---------------+
|       ('Shams', 'Kids Bedding')        |       45      |
|        ('Shams', 'Bed Skirts')         |       45      |
|     ('Kids Bedding', 'Bed Skirts')     |       50      |
|       ('Kids Bedding', 'Sheets')       |       50      |
|        ('Bed Skirts', 'Sheets')        |       45      |
| ('Kids Bedding', 'Bed Skirts', 'Sheets') |       45      |
+----------------------------------------+---------------+

========== Association and Confidence levels =============
+----------------------------------------+-------------------------------+-------------------------------+---------------+------------------+
|             Selected sets              |          Predecessor          |             Result            | Support in (%) | Confidence in (%) |
+----------------------------------------+-------------------------------+-------------------------------+---------------+------------------+
|       ('Shams', 'Kids Bedding')        |           ('Shams',)          |       ('Kids Bedding',)       |       45      |         82       |
|       ('Kids Bedding', 'Shams')        |       ('Kids Bedding',)       |          ('Shams',)           |       45      |         75       |
|        ('Shams', 'Bed Skirts')         |           ('Shams',)          |       ('Bed Skirts',)         |       45      |         82       |
|        ('Bed Skirts', 'Shams')         |       ('Bed Skirts',)         |          ('Shams',)           |       45      |         82       |
|     ('Kids Bedding', 'Bed Skirts')     |       ('Kids Bedding',)       |       ('Bed Skirts',)         |       50      |         83       |
|     ('Bed Skirts', 'Kids Bedding')     |       ('Bed Skirts',)         |       ('Kids Bedding',)       |       50      |         91       |
|       ('Kids Bedding', 'Sheets')       |       ('Kids Bedding',)       |          ('Sheets',)          |       50      |         83       |
|       ('Sheets', 'Kids Bedding')       |          ('Sheets',)          |       ('Kids Bedding',)       |       50      |         100      |
|        ('Bed Skirts', 'Sheets')        |       ('Bed Skirts',)         |          ('Sheets',)          |       45      |         82       |
|        ('Sheets', 'Bed Skirts')        |          ('Sheets',)          |       ('Bed Skirts',)         |       45      |         90       |
| ('Kids Bedding', 'Bed Skirts', 'Sheets') | ('Kids Bedding', 'Bed Skirts') |          ('Sheets',)          |       45      |         75       |
| ('Kids Bedding', 'Sheets', 'Bed Skirts') |  ('Kids Bedding', 'Sheets')   |       ('Bed Skirts',)         |       45      |         75       |
| ('Bed Skirts', 'Sheets', 'Kids Bedding') |   ('Bed Skirts', 'Sheets')    |       ('Kids Bedding',)       |       45      |         82       |
| ('Kids Bedding', 'Bed Skirts', 'Sheets') |       ('Kids Bedding',)       |   ('Bed Skirts', 'Sheets')    |       45      |         75       |
| ('Bed Skirts', 'Kids Bedding', 'Sheets') |       ('Bed Skirts',)         |  ('Kids Bedding', 'Sheets')   |       45      |         82       |
| ('Sheets', 'Bed Skirts', 'Kids Bedding') |          ('Sheets',)          | ('Bed Skirts', 'Kids Bedding') |       45      |         90       |
+----------------------------------------+-------------------------------+-------------------------------+---------------+------------------+

C:\Users\sowmy\OneDrive\Desktop\NJIT\DataMining\Project\AprioriAlgorithm>
```

Input: For dataset Generic



```
C:\Windows\System32\cmd.exe

C:\Users\sowmy\OneDrive\Desktop\NJIT\DataMining\Project\AprioriAlgorithm>python3 Apriori.py generic.csv 50 75
file name  generic.csv
Support 50
Confidence 75
========= Input Transactions ===========

A, B, C
A, B, C
A, B, C, D
A, B, C, D, E
A, B, D, E
A, D, E
A, E
A, E
A, C, E
A, C, E
A, C, E
---------------------------------------

Total number of transactions:  11
```

Output: For dataset Generic

```
C:\Windows\System32\cmd.exe

Frequent sets for k= 1
+------------+-----------+
| Item sets  | Frequency |
+------------+-----------+
|     A      |     11    |
|     C      |     7     |
|     E      |     8     |
+------------+-----------+

Frequent set for k= 2
+------------+-----------+
| Item sets  | Frequency |
+------------+-----------+
| ('A', 'C') |     7     |
| ('A', 'E') |     8     |
+------------+-----------+

========== List of all frequent item sets and support levels ============
+--------------------+----------------+
| Frequent Item sets | Support in (%) |
+--------------------+----------------+
|     ('A', 'C')     |       64       |
|     ('A', 'E')     |       73       |
+--------------------+----------------+

=========== Association and Confidence levels ==============
+---------------+-------------+--------+----------------+--------------------+
| Selected sets | Predecessor | Result | Support in (%) | Confidence in (%) |
+---------------+-------------+--------+----------------+--------------------+
|   ('C', 'A')  |    ('C',)   | ('A',) |       64       |        100         |
|   ('E', 'A')  |    ('E',)   | ('A',) |       73       |        100         |
+---------------+-------------+--------+----------------+--------------------+

C:\Users\sowmy\OneDrive\Desktop\NJIT\DataMining\Project\AprioriAlgorithm>
```

Input: For dataset Amazon

```
C:\Windows\System32\cmd.exe

C:\Users\sowmy\OneDrive\Desktop\NJIT\DataMining\Project\AprioriAlgorithm>python3 Apriori.py amazon.csv 50 85
file name  amazon.csv
Support 50
Confidence 85
========= Input Transactions ==========

A Beginner's Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch
A Beginner's Guide, Java: The Complete Reference, Java For Dummies
A Beginner's Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition
Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition , Beginning Programming with Java,
Android Programming: The Big Nerd Ranch, Beginning Programming with Java, Java 8 Pocket Guide
A Beginner's Guide, Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition
A Beginner's Guide, Head First Java 2nd Edition , Beginning Programming with Java
Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch,
Java For Dummies, Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition , Beginning Programming with Java,
Beginning Programming with Java, Java 8 Pocket Guide, C++ Programming in Easy Steps
A Beginner's Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch
A Beginner's Guide, Java: The Complete Reference, Java For Dummies, HTML and CSS: Design and Build Websites
A Beginner's Guide, Java: The Complete Reference, Java For Dummies, Java 8 Pocket Guide, HTML and CSS: Design and Build Websites
Java For Dummies, Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition
Java For Dummies, Android Programming: The Big Nerd Ranch
A Beginner's Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch
A Beginner's Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch
Head First Java 2nd Edition , Beginning Programming with Java, Java 8 Pocket Guide
Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition
A Beginner's Guide, Java: The Complete Reference, Java For Dummies
--------------------------------------

Total number of transactions:  20
```

Output: For dataset Amazon

Source Code:

Apriori.py

```
import sys
from itertools import combinations
from typing import Dict
from prettytable import PrettyTable

filename = sys.argv[1]
fileobject = open(filename, "r")
lines = fileobject.readlines()
total_no_trans = 0
support_of_all_item_set = {}
min_supp = int(sys.argv[2])
min_conf = int(sys.argv[3])

all_trans_support = {}  # type: Dict[str, int]
all_selected_trans_support = {}  # type: Dict[str, int]
all_selected_trans = []

print("file name ", filename)
```

```
print("Support", min_supp)
print("Confidence", min_conf)
print("========= Input Transactions ===========")
print()


def get_frequent_set(selected_set, rejected_set, all_trans, n):
    global all_selected_trans
    comb = combinations(selected_set, n)
    c2 = dict()
    for i in comb:
        set_i = set(i)
        for j in all_trans:
            if set_i.issubset(j):
                if i in c2:
                    c2[i] += 1
                else:
                    c2[i] = 1
    c2 = {key: value for key, value in c2.items()}
    all_trans_support.update(c2)
    selected_set = []
    rejected_set = []

    for key, value in c2.items():
        if (value / total_no_trans) * 100 >= min_supp:
            selected_set.append(key)
        # t.add_row([key, value])
        else:
            rejected_set.append(key)
    if selected_set:
        t = PrettyTable(['Item sets', 'Frequency'])
        for i in selected_set:
            t.add_row([i, c2.get(i)])
        print()
        print("Frequent set for k=", n)

        all_selected_trans = all_selected_trans + selected_set
        #  print("reject list for k= ", n, rejected_set)
        print(t)
    return selected_set, rejected_set


c1 = dict()
all_trans = []
for line in lines:
    line = line.replace("\n", "")
    print(line)
    words = line.split(", ")
    total_no_trans += 1
    seen = set()
    for word in words:
        if word in c1:
            c1[word] += 1
        else:
            c1[word] = 1
        seen.add(word)
    all_trans.append(seen)
```

```
c1 = {key: value for key, value in c1.items()}
all_trans_support.update(c1)
selected_set = []
rejected_set = []
print("---------------------------------------")
print()
print("total number of transactions: ", total_no_trans)

t = PrettyTable(['Item sets ', 'Frequency'])
for key, value in c1.items():
    # print(key, ' :: ', value)
    if (value / total_no_trans) * 100 >= min_supp:
        selected_set.append(key)
        t.add_row([key, value])
    else:
        rejected_set.append(key)
# print("Frequent set for k=1", selected_set)
# print("reject list k=1 ", rejected_set)
print()
print("Frequent sets for k= 1")
print(t)

size = 2
while True:
    selected_set_1, rejected_set_1 = get_frequent_set(selected_set,
rejected_set, all_trans, size)
    if not selected_set_1:
        break
    rejected_set = selected_set_1
    frequent_set = rejected_set_1
    size += 1
print()
print("========== List of all frequent item sets and support levels
============ ")
t = PrettyTable(['Frequent Item sets', 'Support in (%)'])
for support in all_selected_trans:
    Support_itemset = round(all_trans_support.get(support) / total_no_trans *
100)
    t.add_row([support, Support_itemset])
print(t)

print()
print("========== Association and Confidence levels ==============")
t = PrettyTable(['Selected sets', 'Predecessor', 'Result', 'Support in (%)',
'Confidence in (%)'])

for x in all_selected_trans:
    size_of_item_set = len(x)
    itemset = set(x)

    while size_of_item_set - 1 > 0:
        comb = combinations(x, size_of_item_set - 1)
        for i in comb:
            left_side_items = i
            right_side_items = tuple(itemset - set(i))

            item_conf = round(round(all_trans_support.get(x) / total_no_trans
```

```
* 100) * 100 / round(
                all_trans_support.get(left_side_items[0]) / total_no_trans *
100), 2)
            if item_conf >= min_conf:
                # print(left_side_items, "=>", right_side_items,
round(all_trans_support.get(x) / total_no_trans * 100), round(item_conf),
"Selected")
                t.add_row([left_side_items + right_side_items,
left_side_items, right_side_items,
                           round(all_trans_support.get(x) / total_no_trans *
100), round(item_conf)])

            # else:
            # print(left_side_items, "=>", right_side_items,
round(all_trans_support.get(x) / total_no_trans * 100), round(item_conf),
"Rejected")

        size_of_item_set -= 1
print(t)
```