Alex Shung

504049665

**Part A)**

1. A subset of a frequent itemset will also be frequent. A superset of an infrequent pattern will also be infrequent.

Bold will be frequent candidates, all candidates will be listed per scan

Scan 1

|  |  |
| --- | --- |
| Item | Support |
| **A** | 4 |
| **B** | 4 |
| **C** | 5 |
| **P** | 2 |
| **R** | 2 |
| S | 1 |
| T | 1 |

Self join on frequent items produce: ab, ac, ap, ar, bc, bp, br, cp, cr, pr. No pruning possible

Scan 2

|  |  |
| --- | --- |
| **Ab** | 2 |
| **Ac** | 3 |
| **Bc** | 3 |
| **Cr** | 2 |
| Ap | 1 |
| Ar | 1 |
| Bp | 1 |
| Br | 1 |
| Cp | 1 |

Self join on frequent items produce: abc, acr, bcr. Can prune acr since ar is not frequent, and bcr since br is infrequent.

Scan 3

|  |  |
| --- | --- |
| **Abc** | 2 |

No more frequent item sets can be produced, so the algorithm terminates

**Part B)**

S = scan #, T = transaction #. Example S1 T3 = Frequent after scan 1 transaction 3 is read.

|  |  |
| --- | --- |
| C | S1 T3 |
| R | S1 T3 |
| A | S1 T4 |
| P | S1 T4 |
| B | S1 T3 |
| CR | S2 T3 |
| BC | S1 T6 |
| AB | S1 T6 |
| AC | S1 T6 |
| ABC | S2 T6 |

DIC uses two scans, Apriori uses three scans

Final list: a,b,c,p,r,ab,ac,bc,cr,abc

**Part C)**

Root

P: 1

B: 1

R: 1

B: 2

P: 1

R: 1

A: 4

B: 1

C: 5

|  |
| --- |
| C : 5 |
| A : 4 |
| B : 4 |
| P : 2 |
| R : 2 |

Ordered List: C, A, B, P, R

**Part 2 of C)**

Root

P: 1

B: 1

R: 1

B: 2

A: 2

B: 1

C: 3

All frequent subsets containing b: B, CB, CAB, AB.

Traveled the prefix of each B node (technically attached to the header table, represented as the ordered list above, all part of a single linked list). Union of the prefix with all suffixes that exceeded the threshold, however no suffix was past the support threshold. Therefore only B, CB, AB, and CAB (the left most branch) was kept.

Created trees:

B:2 -> A:2 -> C:3 -> Root

B:1 -> C:3 -> Root

B:1 -> Root

Projected Database:

B, AB, CAB, CB

B, BR, CBR, CB

BP, B

**Part D)**

<ac> support = 3 (T1, T2, T4)

<(a,c)> support = 1 (T4)

Prefix: Projected Databases

|  |  |  |
| --- | --- | --- |
| <c> : 5 | <d> : 4 | <a> : 3 |
| <(\_d)> | <(c,d)> | <(\_d)(c,d)> |
| <a> | <ca> | <dca> |
| <(c)> | <(a,c)(c)> | <(\_d)(a,c)(c)> |
|  | <c> |  |

Sequential Patterns

|  |  |  |
| --- | --- | --- |
| <c> | <d> | <a> |
|  | <dc> | <ad> |
|  | <da> | <ac> |
|  | <d(c)> | <aa> |
|  |  | <(ad)> |

Next group of projected databases (red = pruned)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| <dc> | <da> | <d(c)> | <ad> | <ac> | <aa> | <(ad)> |
|  | <(\_c)(c)> | <(\_d)> |  | <(\_d)> | <(\_c)(c,e)> | <(c,d)> |
| <a> |  | <(c)> | <ca> | <a> |  | <(a,c)(c)> |
| <(c)> |  |  |  | <(c)> |  |  |

Additional Sequential patterns

|  |  |
| --- | --- |
| <ad> | <(ad)> |
|  | <(ad)(c)> |

Project databases

|  |
| --- |
| <(ad)(c)> |
| <(\_d)> |
| <(c)> |
|  |

Final Sequential Patterns

|  |  |  |
| --- | --- | --- |
| <a> | <d> | <c> |
| <ad> | <dc> |  |
| <ac> | <da> |  |
| <aa> |  |  |
| <(ad)> |  |  |
| <(ad)(c)> |  |  |