HW2 - Frequent Itemsets

Solution

The solution is based upon two classes, described below

Apriori

The apriori class implements the apriori algorithm which finds all frequent itemsets in a set of transactions. The algorithm is based upon the fact that if a set of items is frequent, all its subsets also are. This allows us to solve the problem by finding all frequent itemsets of progressively bigger size, using the found frequent itemsets from previous iterations in order to minimize the load the algorithm has on memory.

The code does this by first looping through the entire dataset and counting each item, producing a list of singletons. After that, it moves on to finding all doubletons by iterating through all transactions and seeing if it contains any combination of two frequent singletons, and in that case, counts it. This way, we can minimize the data saved to memory by discarding all potential itemsets that have an infrequent subset.

To find tripletons, as well as all larger subsets, the same algorithm is applied: For each frequent itemset one item smaller than the current, go through all transactions, checking if either of them has a subset that is a combination of a one-item-smaller frequent itemset combined with a frequent singleton, and in that case counting it.

This is then repeated until we run out of frequent subsets for an iteration, or when the potential frequent subset is larger than any transaction.

RuleGeneration

The RuleGeneration class implements an algorithm to find all rules of specified confidence, meaning all implications of an itemset being found given that another is present that is large enough.

This is done by going through each frequent itemset and splitting it into all possible combinations (meaning trying each subset as the left side with the rest as the right side), and keeping the implications that have enough confidence.

The confidence is calculated by dividing the support for the union of the right and left sides by the support of the left side, which translates to the percentage of itemsets of the left side that also contain the right side, meaning the right side is implied with confidence c if we see the left side.

How to run

The code is run by simply running the main file the regular way you run Python scripts. If you want to change hyperparameters, this is also done in the main file.

Results

The parenthesis indicates how many of each itemset was found

Run 1

```
# Hyperparameters
s_precentage = 0.01
c = 0.5
```

Itemsets of size 1:
25 (1395), 52 (1983), 240 (1399), … (See appendix)

Itemsets of size 2:
('39', '825') (1187)
('704', '825') (1102)
('39', '704') (1107)
('227', '390') (1049)
('789', '829') (1194)
('368', '829') (1194)
('217', '346') (1336)
('368', '682') (1193)
('390', '722') (1042)

Itemsets of size 3:
('39', '704', '825') (1035)

['704'] -> ['825'] (confidence 0.61)
['704'] -> ['39'] (confidence 0.62)
['227'] -> ['390'] (confidence 0.58)

['704'] -> ['39', '825'] (confidence 0.58)
['39', '704'] -> ['825'] (confidence 0.93)
['39', '825'] -> ['704'] (confidence 0.87)
['704', '825'] -> ['39'] (confidence 0.94)
7 rules found

Apriori took 24.2 seconds
Rule generation took 0.0 seconds
Total time: 24.2 seconds

Run 2

```
# Hyperparameters
s_precentage = 0.005
c = 0.5
```

Itemsets of size 1:
25 (1395), 52 (1983), 240 (1399), … (See appendix)

Itemsets of size 2:
('39', '825') (1187)
('704', '825') (1102)
('39', '704') (1107)
('529', '782') (862)
… (See appendix)

Itemsets of size 3:
('39', '704', '825') (1035)
('283', '33', '346') (802)
('217', '283', '346') (827)
('283', '346', '515') (806)
('217', '33', '346') (802)
('217', '346', '515') (809)
('227', '390', '722') (907)

['969'] -> ['888'] (confidence 0.95)
['33', '346'] -> ['283'] (confidence 0.95)
['283', '515'] -> ['346'] (confidence 0.97)
['33', '346'] -> ['217'] (confidence 0.95)
['217', '515'] -> ['346'] (confidence 0.96)

['346', '515'] -> ['217'] (confidence 0.95)
6 rules found

Apriori took 33.58 seconds
Rule generation took 0.01 seconds
Total time: 33.59 seconds

Appendix

Run 1 as a whole

Itemsets of size 1:
25 (1395), 52 (1983), 240 (1399), 274 (2628), 368 (7828), 448 (1370), 538 (3982), 561 (2783), 630 (1523), 687 (1762), 775 (3771), 825 (3085), 834 (1373), 39 (4258), 120 (4973), 205 (3605), 401 (3667), 581 (2943), 704 (1794), 814 (1672), 35 (1984), 674 (2527), 733 (1141), 854 (2847), 950 (1463), 422 (1255), 449 (1890), 857 (1588), 895 (3385), 937 (4681), 964 (1518), 229 (2281), 283 (4082), 294 (1445), 381 (2959), 708 (1090), 738 (2129), 766 (6265), 853 (1804), 883 (4902), 966 (3921), 978 (1141), 104 (1158), 143 (1417), 569 (2835), 620 (2100), 798 (3103), 185 (1529), 214 (1893), 350 (3069), 529 (7057), 658 (1881), 682 (4132), 782 (2767), 809 (2163), 947 (3690), 970 (2086), 227 (1818), 390 (2685), 71 (3507), 192 (2004), 208 (1483), 279 (3014), 280 (2108), 496 (1428), 530 (1263), 597 (2883), 618 (1337), 675 (2976), 720 (3864), 914 (4037), 932 (1786), 183 (3883), 217 (5375), 276 (2479), 653 (2634), 706 (1923), 878 (2047), 161 (2320), 175 (2791), 177 (4629), 424 (1448), 490 (1066), 571 (2902), 623 (1845), 795 (3361), 910 (1695), 960 (2732), 125 (1287), 130 (1711), 392 (2420), 461 (1498), 862 (3649), 27 (2165), 78 (2471), 900 (1165), 921 (2425), 147 (1383), 411 (2047), 572 (1589), 579 (2164), 778 (2514), 803 (2237), 266 (1022), 290 (1793), 458 (1124), 523 (2244), 614 (3134), 888 (3686), 944 (2794), 43 (1721), 70 (2411), 204 (2174), 334 (2146), 480 (2309), 513 (1287), 874 (2237), 151 (2611), 504 (1296), 890 (1437), 73 (2179), 310 (1390), 419 (5057), 469 (1502), 722 (5845), 810 (1267), 844 (2814), 846 (1480), 918 (3012), 967 (1695), 326 (1488), 403 (1722), 526 (2793), 774 (2046), 788 (2386), 789 (4309), 975 (1764), 116 (2193), 198 (1461), 201 (1029), 171 (1097), 541 (3735), 701 (1283), 805 (1789), 946 (1350), 471 (2894), 487 (3135), 631 (2793), 638 (2288), 678 (1329), 735 (1689), 780 (2306), 935 (1742), 17 (1683), 242 (2325), 758 (2860), 763 (1862), 956 (3626), 145 (4559), 385 (1676), 676 (2717), 790 (1094), 792 (1306), 885 (3043), 522 (2725), 617 (2614), 859 (1242), 12 (3415), 296 (2210), 354 (5835), 548 (2843), 684 (5408), 740 (1632), 841 (1927), 210 (2009), 346 (3470), 477 (2462), 605 (1652), 829 (6810), 884 (1645), 234 (1416), 460 (4438), 649 (1292), 746 (1982), 600 (1192), 28 (1454), 157 (1140), 5 (1094), 115 (1775),

517 (1201), 736 (1470), 744 (2177), 919 (3710), 196 (2096), 489 (3420), 494 (5102), 641 (1494), 673 (1635), 362 (4388), 591 (1241), 31 (1666), 58 (1330), 181 (1235), 472 (2125), 573 (1229), 628 (1102), 651 (1288), 111 (1171), 154 (1447), 168 (1538), 580 (1667), 632 (1070), 832 (2062), 871 (2810), 988 (1164), 72 (2852), 981 (1542), 10 (1351), 132 (2641), 21 (2666), 32 (4248), 54 (2595), 239 (2742), 348 (1226), 100 (1749), 500 (1444), 48 (2472), 126 (1075), 319 (1371), 639 (1572), 765 (1705), 521 (1582), 112 (2680), 140 (2687), 285 (2600), 387 (2089), 511 (1015), 594 (1516), 93 (2777), 583 (1389), 606 (2668), 236 (2618), 952 (1574), 90 (1875), 593 (2601), 941 (1126), 122 (1081), 718 (1238), 1 (1535), 423 (1412), 516 (1544), 6 (2149), 69 (2370), 797 (2684), 913 (1939), 577 (1695), 110 (1801), 509 (3044), 611 (1444), 995 (1521), 343 (1599), 527 (1185), 33 (1460), 336 (1071), 989 (1289), 97 (1466), 574 (1297), 793 (3063), 598 (3219), 427 (1856), 470 (4137), 37 (1249), 992 (1116), 55 (1959), 897 (1935), 275 (1692), 51 (1612), 259 (1522), 45 (1728), 162 (1450), 378 (1149), 534 (1531), 906 (1444), 912 (1009), 576 (1337), 373 (2007), 716 (1199), 546 (1050), 665 (1297), 963 (1327), 349 (2041), 8 (3090), 197 (1230), 413 (2637), 749 (1330), 823 (1031), 94 (1201), 982 (1640), 984 (1756), 515 (1166), 692 (4993), 694 (2847), 567 (1102), 57 (2743), 800 (1916), 812 (1518), 41 (1353), 414 (1160), 923 (1753), 377 (1149), 752 (2578), 991 (1268), 998 (2713), 899 (1252), 710 (1044), 867 (1530), 170 (1203), 438 (4511), 563 (1065), 357 (1142), 332 (1861), 361 (1104), 322 (1154), 928 (1034), 75 (3151), 486 (1547), 440 (1943), 38 (2402), 784 (1257), 265 (1359), 686 (1495), 540 (1293), 468 (1089), 663 (2354), 819 (1257), 886 (3053), 429 (1037), 843 (1222), 129 (1547), 578 (1290), 510 (3281), 68 (1601), 860 (1255), 4 (1394), 887 (1671), 309 (1262), 804 (1315), 325 (1022), 826 (2022), 394 (1145), 707 (1354), 105 (1100), 815 (1358), 948 (1149), 308 (1402), 661 (2693), 634 (2492), 351 (1641), 405 (1525), 688 (1132), 949 (1414), 163 (1256), 893 (1947), 335 (1345), 173 (1080), 258 (1036), 85 (1555), 450 (2082), 428 (1021), 550 (1203), 769 (1622), 554 (1114), 366 (1031), 820 (1473), 207 (1214)

Itemsets of size 2:
('39', '825') (1187)
('704', '825') (1102)
('39', '704') (1107)
('227', '390') (1049)
('789', '829') (1194)
('368', '829') (1194)
('217', '346') (1336)
('368', '682') (1193)
('390', '722') (1042)

Itemsets of size 3:
('39', '704', '825') (1035)

['704'] -> ['825'] (confidence 0.61)
['704'] -> ['39'] (confidence 0.62)
['227'] -> ['390'] (confidence 0.58)
['704'] -> ['39', '825'] (confidence 0.58)
['39', '704'] -> ['825'] (confidence 0.93)
['39', '825'] -> ['704'] (confidence 0.87)
['704', '825'] -> ['39'] (confidence 0.94)
7 rules found

Apriori took 24.2 seconds
Rule generation took 0.0 seconds
Total time: 24.2 seconds

Run 2 as a whole

Itemsets of size 1:
25 (1395), 52 (1983), 240 (1399), 274 (2628), 368 (7828), 448 (1370), 538 (3982), 561 (2783), 630 (1523), 687 (1762), 775 (3771), 825 (3085), 834 (1373), 39 (4258), 120 (4973), 205 (3605), 401 (3667), 581 (2943), 704 (1794), 814 (1672), 35 (1984), 674 (2527), 712 (845), 733 (1141), 854 (2847), 950 (1463), 422 (1255), 449 (1890), 857 (1588), 895 (3385), 937 (4681), 964 (1518), 229 (2281), 283 (4082), 294 (1445), 352 (902), 381 (2959), 708 (1090), 738 (2129), 766 (6265), 853 (1804), 883 (4902), 966 (3921), 978 (1141), 104 (1158), 143 (1417), 569 (2835), 620 (2100), 798 (3103), 7 (997), 185 (1529), 214 (1893), 350 (3069), 529 (7057), 658 (1881), 682 (4132), 782 (2767), 809 (2163), 947 (3690), 970 (2086), 227 (1818), 390 (2685), 71 (3507), 192 (2004), 208 (1483), 279 (3014), 280 (2108), 496 (1428), 530 (1263), 597 (2883), 618 (1337), 675 (2976), 720 (3864), 855 (939), 914 (4037), 932 (1786), 183 (3883), 193 (925), 217 (5375), 276 (2479), 277 (982), 474 (815), 626 (874), 653 (2634), 706 (1923), 878 (2047), 161 (2320), 175 (2791), 177 (4629), 424 (1448), 490 (1066), 571 (2902), 623 (1845), 795 (3361), 910 (1695), 960 (2732), 125 (1287), 130 (1711), 839 (854), 392 (2420), 461 (1498), 801 (835), 862 (3649), 27 (2165), 78 (2471), 900 (1165), 921 (2425), 147 (1383), 411 (2047), 572 (1589), 579 (2164), 778 (2514), 803 (2237), 266 (1022), 290 (1793), 458 (1124), 523 (2244), 614 (3134), 888 (3686), 944 (2794), 969 (849), 43 (1721), 70 (2411), 204 (2174), 334 (2146), 480 (2309), 513 (1287), 874 (2237), 151 (2611), 432 (985), 504 (1296), 830 (841), 890 (1437), 73 (2179), 118 (916), 310 (1390), 388 (938), 419 (5057), 469 (1502), 484 (971), 722 (5845), 810 (1267), 844 (2814), 846 (1480), 918 (3012), 967 (1695), 326 (1488), 403 (1722), 526 (2793), 774 (2046), 788 (2386), 789 (4309), 975 (1764), 116 (2193), 198

(1461), 201 (1029), 395 (990), 171 (1097), 541 (3735), 701 (1283), 805 (1789), 946 (1350), 471 (2894), 487 (3135), 631 (2793), 638 (2288), 640 (932), 678 (1329), 735 (1689), 780 (2306), 935 (1742), 17 (1683), 242 (2325), 758 (2860), 763 (1862), 956 (3626), 145 (4559), 385 (1676), 676 (2717), 790 (1094), 792 (1306), 885 (3043), 522 (2725), 617 (2614), 859 (1242), 12 (3415), 296 (2210), 354 (5835), 548 (2843), 684 (5408), 740 (1632), 841 (1927), 210 (2009), 346 (3470), 477 (2462), 605 (1652), 829 (6810), 884 (1645), 234 (1416), 355 (958), 460 (4438), 649 (1292), 746 (1982), 600 (1192), 28 (1454), 157 (1140), 742 (953), 5 (1094), 115 (1775), 517 (1201), 736 (1470), 744 (2177), 919 (3710), 196 (2096), 489 (3420), 494 (5102), 641 (1494), 673 (1635), 723 (829), 362 (4388), 591 (1241), 622 (826), 31 (1666), 58 (1330), 181 (1235), 329 (964), 417 (971), 472 (2125), 573 (1229), 628 (1102), 651 (1288), 111 (1171), 154 (1447), 168 (1538), 580 (1667), 632 (1070), 832 (2062), 871 (2810), 988 (1164), 72 (2852), 585 (856), 981 (1542), 10 (1351), 132 (2641), 464 (848), 21 (2666), 32 (4248), 54 (2595), 239 (2742), 348 (1226), 100 (1749), 500 (1444), 48 (2472), 126 (1075), 319 (1371), 639 (1572), 765 (1705), 521 (1582), 112 (2680), 140 (2687), 285 (2600), 387 (2089), 511 (1015), 594 (1516), 93 (2777), 583 (1389), 606 (2668), 236 (2618), 952 (1574), 90 (1875), 593 (2601), 941 (1126), 122 (1081), 718 (1238), 1 (1535), 423 (1412), 516 (1544), 6 (2149), 69 (2370), 415 (877), 797 (2684), 913 (1939), 980 (899), 577 (1695), 110 (1801), 509 (3044), 611 (1444), 995 (1521), 343 (1599), 447 (863), 527 (1185), 33 (1460), 158 (884), 336 (1071), 989 (1289), 97 (1466), 574 (1297), 793 (3063), 598 (3219), 427 (1856), 470 (4137), 37 (1249), 858 (866), 992 (1116), 55 (1959), 95 (841), 347 (885), 481 (888), 897 (1935), 224 (919), 275 (1692), 51 (1612), 259 (1522), 45 (1728), 162 (1450), 378 (1149), 534 (1531), 906 (1444), 912 (1009), 44 (903), 96 (975), 576 (1337), 642 (830), 879 (865), 18 (813), 373 (2007), 716 (1199), 546 (1050), 568 (956), 665 (1297), 785 (947), 963 (1327), 349 (2041), 8 (3090), 197 (1230), 413 (2637), 749 (1330), 823 (1031), 94 (1201), 982 (1640), 984 (1756), 515 (1166), 692 (4993), 694 (2847), 567 (1102), 57 (2743), 800 (1916), 812 (1518), 41 (1353), 414 (1160), 590 (814), 923 (1753), 377 (1149), 160 (987), 456 (804), 752 (2578), 991 (1268), 998 (2713), 899 (1252), 114 (816), 710 (1044), 867 (1530), 170 (1203), 438 (4511), 563 (1065), 705 (888), 357 (1142), 659 (835), 332 (1861), 361 (1104), 322 (1154), 928 (1034), 75 (3151), 108 (940), 486 (1547), 440 (1943), 38 (2402), 784 (1257), 265 (1359), 624 (915), 686 (1495), 943 (821), 540 (1293), 468 (1089), 663 (2354), 819 (1257), 886 (3053), 429 (1037), 843 (1222), 129 (1547), 578 (1290), 510 (3281), 68 (1601), 860 (1255), 318 (812), 4 (1394), 304 (805), 887 (1671), 309 (1262), 804 (1315), 325 (1022), 745 (909), 268 (885), 826 (2022), 394 (1145), 707 (1354), 838 (953), 105 (1100), 815 (1358), 948 (1149), 345 (801), 308 (1402), 661 (2693), 634 (2492), 351 (1641), 405 (1525), 688 (1132), 949 (1414), 163 (1256), 893 (1947), 335 (1345), 922 (990), 173

(1080), 203 (861), 258 (1036), 629 (893), 66 (888), 314 (846), 85 (1555), 450 (2082), 428 (1021), 550 (1203), 769 (1622), 80 (826), 558 (957), 608 (999), 507 (950), 554 (1114), 366 (1031), 689 (817), 820 (1473), 831 (852), 207 (1214),

Itemsets of size 2:
('39', '825') (1187)
('704', '825') (1102)
('39', '704') (1107)
('529', '782') (862)
('227', '390') (1049)
('795', '853') (806)
('571', '795') (838)
('623', '795') (805)
('392', '862') (881)
('411', '803') (826)
('208', '290') (803)
('208', '888') (829)
('208', '969') (806)
('290', '888') (826)
('888', '969') (810)
('471', '678') (810)
('789', '829') (1194)
('392', '489') (866)
('368', '829') (1194)
('541', '72') (846)
('529', '598') (943)
('598', '782') (800)
('283', '33') (845)
('217', '283') (926)
('283', '346') (910)
('283', '515') (835)
('217', '33') (852)
('217', '346') (1336)
('217', '515') (843)
('33', '346') (844)
('346', '515') (849)
('33', '515') (824)
('923', '947') (859)
('438', '684') (825)

('684', '70') (860)
('684', '765') (812)
('684', '819') (800)
('368', '682') (1193)
('368', '494') (860)
('368', '692') (928)
('227', '722') (995)
('390', '722') (1042)
('675', '886') (823)
('471', '960') (935)

Itemsets of size 3:
('39', '704', '825') (1035)
('283', '33', '346') (802)
('217', '283', '346') (827)
('283', '346', '515') (806)
('217', '33', '346') (802)
('217', '346', '515') (809)
('227', '390', '722') (907)

['704'] -> ['825'] (confidence 0.61)
['704'] -> ['39'] (confidence 0.62)
['227'] -> ['390'] (confidence 0.58)
['208'] -> ['290'] (confidence 0.54)
['208'] -> ['888'] (confidence 0.56)
['208'] -> ['969'] (confidence 0.54)
['969'] -> ['208'] (confidence 0.95)
['969'] -> ['888'] (confidence 0.95)
['678'] -> ['471'] (confidence 0.61)
['33'] -> ['283'] (confidence 0.58)
['515'] -> ['283'] (confidence 0.72)
['33'] -> ['217'] (confidence 0.58)
['515'] -> ['217'] (confidence 0.72)
['33'] -> ['346'] (confidence 0.58)
['515'] -> ['346'] (confidence 0.73)
['33'] -> ['515'] (confidence 0.56)
['515'] -> ['33'] (confidence 0.71)
['819'] -> ['684'] (confidence 0.64)
['227'] -> ['722'] (confidence 0.55)
['704'] -> ['825', '39'] (confidence 0.58)

['39', '704'] -> ['825'] (confidence 0.93)
['39', '825'] -> ['704'] (confidence 0.87)
['704', '825'] -> ['39'] (confidence 0.94)
['33'] -> ['346', '283'] (confidence 0.55)
['283', '33'] -> ['346'] (confidence 0.95)
['283', '346'] -> ['33'] (confidence 0.88)
['33', '346'] -> ['283'] (confidence 0.95)
['217', '283'] -> ['346'] (confidence 0.89)
['217', '346'] -> ['283'] (confidence 0.62)
['283', '346'] -> ['217'] (confidence 0.91)
['515'] -> ['346', '283'] (confidence 0.69)
['283', '346'] -> ['515'] (confidence 0.89)
['283', '515'] -> ['346'] (confidence 0.97)
['346', '515'] -> ['283'] (confidence 0.95)
['33'] -> ['217', '346'] (confidence 0.55)
['217', '33'] -> ['346'] (confidence 0.94)
['217', '346'] -> ['33'] (confidence 0.6)
['33', '346'] -> ['217'] (confidence 0.95)
['515'] -> ['217', '346'] (confidence 0.69)
['217', '346'] -> ['515'] (confidence 0.61)
['217', '515'] -> ['346'] (confidence 0.96)
['346', '515'] -> ['217'] (confidence 0.95)
['227', '390'] -> ['722'] (confidence 0.86)
['227', '722'] -> ['390'] (confidence 0.91)
['390', '722'] -> ['227'] (confidence 0.87)
45 rules found

Apriori took 33.58 seconds
Rule generation took 0.01 seconds
Total time: 33.59 seconds