

Homework #2: Toivonen**Due: October 7, Sunday****100 points**

In this homework, you are asked to implement the Toivonen algorithm in Python. You are provided with a dataset which contains a set of baskets, one per line. Each basket contains a set of items represented using item numbers. You are also provided with a threshold on support ratio, denoted as t . Assume that there are n baskets in the data set. So an itemset is frequent if it appears in $n \cdot t$ baskets (taking its ceiling if $n \cdot t$ is not an integer).

We assume that the sample is 10% of the baskets. So if there are $n = 1000$ baskets, then the sample size s is 100 (again round up to its closest integer if $10\%n$ is not an integer). We assume that the support ratio threshold $t = 4/15$. Then an itemset will be frequent in the sample (of size 100) if it appears in at least 20 baskets of the sample.

To generate a sample, use Python random package. For the first sample, set the seed to be 1; second, 2; and so on. So that our present iteration times should be same as the present seed value. So to get the first sample, set `random.seed(1)`; then use `random.randint(1, n)` to obtain s random numbers in the range of 1 to n .

Sample execution:

```
python Toivonen.py dataset.txt
```

For each iteration of the algorithm, output the content of the sample created, frequent itemsets discovered from the sample, itemsets in the negative border, and any false negatives.

Finally, you need to output the total running time of your program.

Submission notes:

1. You need to print false negatives for every iteration and print the total running time, like this:

```
/Users/ruotianjiang/PycharmProjects/inf/spark/bin/python -m Toivonen /Users/ruotianjiang/Downloads/cp_for_inf553/hw/solution2/dataset/itemsets.txt
False Negatives: [47, 590, 595, (110, 356), (356, 589), (480, 589)]
False Negatives: [32]
False Negatives: [32, 47, 50, 344, 858, (1, 356), (260, 356), (296, 480), (296, 527), (356, 457), (356, 480), (356, 527), (356, 588), (356, 589), (480, 589), (480, 593)]
False Negatives: [32, 47, 50, 344, 858, (1, 356), (260, 356), (296, 480), (296, 527), (356, 457), (356, 480), (356, 527), (356, 588), (356, 589), (480, 589), (480, 593)]
False Negatives: [344, 364, (318, 356), (318, 593), (356, 457), (356, 589), (527, 593)]
False Negatives: [590, (480, 593)]
False Negatives: [1, 32, 50, 608, 858, (296, 318), (296, 356), (296, 480), (296, 527), (296, 593), (356, 593)]
False Negatives: [32, 47, 50, 595, 780, 858, (1, 356), (260, 356), (296, 480), (296, 527), (356, 457), (356, 480), (356, 589), (480, 593)]
False Negatives: [457, (296, 318), (296, 527), (296, 593), (318, 356), (318, 593), (527, 593)]
False Negatives: [318, 592, (1, 356), (110, 356), (260, 356), (296, 527), (356, 457), (356, 527), (356, 588), (356, 589), (356, 593), (527, 593)]
False Negatives: [457, (296, 318), (318, 356), (318, 593)]
False Negatives: [457, 595, (296, 318), (318, 356), (318, 593), (356, 588)]
False Negatives: [32, 47, 50, 344, 858, (1, 356), (260, 356), (296, 318), (296, 356), (296, 480), (296, 527), (356, 480), (356, 589), (480, 593)]
False Negatives: [47, 344, 590, 595, (110, 356), (356, 588), (356, 589), (480, 589)]
False Negatives: []
--- 15 times iteration ---
--- 14.6034650803 seconds ---
```

2. You need to write output and store them in a folder named “output”, where the output of each iteration is stored as a separate file named as OutputForIteration_(iteration#), like this:

▼ output	Today at 12:32 PM	--	Folder
OutputForIteration_1.txt	Today at 12:31 PM	16 KB	Plair
OutputForIteration_2.txt	Today at 12:31 PM	27 KB	Plair
OutputForIteration_3.txt	Today at 12:31 PM	14 KB	Plair
OutputForIteration_4.txt	Today at 12:31 PM	14 KB	Plair
OutputForIteration_5.txt	Today at 12:31 PM	16 KB	Plair
OutputForIteration_6.txt	Today at 12:31 PM	28 KB	Plair
OutputForIteration_7.txt	Today at 12:31 PM	18 KB	Plair
OutputForIteration_8.txt	Today at 12:31 PM	13 KB	Plair
OutputForIteration_9.txt	Today at 12:31 PM	20 KB	Plair
OutputForIteration_10.txt	Today at 12:31 PM	13 KB	Plair
OutputForIteration_11.txt	Today at 12:31 PM	22 KB	Plair
OutputForIteration_12.txt	Today at 12:31 PM	20 KB	Plair
OutputForIteration_13.txt	Today at 12:31 PM	14 KB	Plair
OutputForIteration_14.txt	Today at 12:31 PM	19 KB	Plair
OutputForIteration_15.txt	Today at 12:31 PM	30 KB	Plair

3. Every file should report the following (and in the order shown). Follow the format as shown in the sample outputs.

- output the content of the sample created,
- frequent itemsets discovered from the sample
- itemsets in the negative border

Sample Created:

```
[262, 542, 595, 724, 934], [23, 50, 318, 541, 750, 858, 913, 922, 934], [1, 2, 10, 16, 34, 50, 60, 111, 1
595, 608, 616, 743, 858, 912, 919, 924], [1, 3, 17, 25, 32, 41, 62, 65, 76, 104, 376, 494, 640, 648, 719,
318, 344, 356, 380, 431, 441, 501, 520, 541, 551, 588, 592, 608, 678, 743, 745, 750, 778, 849, 858, 903,
43, 45, 47, 48, 50, 57, 61, 62, 71, 74, 79, 83, 89, 95, 100, 102, 105, 110, 122, 135, 140, 150, 153, 155,
225, 235, 236, 237, 246, 252, 256, 260, 261, 262, 266, 270, 276, 277, 280, 281, 282, 289, 292, 293, 296,
364, 368, 371, 372, 376, 377, 378, 379, 380, 384, 408, 412, 414, 415, 416, 420, 421, 422, 424, 434, 440,
507, 508, 510, 514, 515, 516, 517, 524, 527, 529, 531, 532, 534, 539, 541, 551, 552, 553, 569, 586, 587,
650, 653, 674, 691, 694, 707, 708, 720, 733, 736, 750, 762, 780, 786, 788, 798, 800, 802, 805, 809, 828,
910, 911, 912, 913, 914, 915, 916, 917, 918, 919, 920, 921, 922, 923, 924, 926, 928, 930, 932, 933, 934,
969, 971, 972, 973, 976, 982, 992], [2, 3, 19, 45, 160, 231, 318, 342, 466, 480, 520, 543, 608, 784], [22
351, 356, 370, 377, 441, 480, 514, 532, 608, 663, 750, 778, 799, 904, 908, 923], [1, 7, 17, 32, 36, 47, 5
593, 597, 608, 720, 745, 778, 780, 923], [19, 32, 44, 95, 110, 150, 151, 158, 160, 165, 172, 173, 196, 20
420, 432, 434, 442, 454, 457, 480, 500, 527, 551, 587, 589, 590, 592, 597, 736, 780], [50, 145, 163, 296,
318, 344, 356, 364, 367, 380, 457, 480, 497, 500, 527, 588, 590, 592, 593, 595, 597, 608, 671, 720, 736,
319, 338, 339, 358, 377, 378, 388, 422, 442, 477, 494, 539, 550, 588, 592, 597, 627, 628, 708, 719, 780,
250, 260, 292, 293, 296, 318, 333, 337, 342, 356, 377, 441, 442, 456, 457, 480, 527, 529, 541, 586, 593,
474, 527, 541, 549, 608, 778, 858, 913], [260, 541, 589, 593], [1, 150, 153, 165, 260, 318, 344, 356, 364
32, 110, 126, 173, 185, 260, 353, 356, 367, 380, 405, 442, 474, 480, 493, 541, 589, 592, 651, 673, 733, 7
434, 454, 457, 474, 480, 500, 586, 590, 592, 593], [1, 6, 86, 111, 260, 265, 376, 480, 509, 527, 541, 608
17, 22, 25, 30, 32, 34, 36, 47, 50, 59, 70, 81, 95, 111, 112, 117, 123, 149, 153, 164, 176, 190, 198, 229
490, 494, 509, 527, 532, 541, 549, 589, 592, 593, 599, 608, 627, 632, 696, 707, 750, 757, 778, 787, 800,
260, 356, 357, 480, 527, 593, 858, 912, 953], [6, 47, 293, 296, 318, 356, 364, 858], [24, 32, 73, 111, 17
150, 153, 165, 185, 225, 253, 266, 282, 288, 292, 296, 315, 316, 329, 337, 344, 349, 356, 364, 377, 380,
158, 165, 223, 260, 329, 349, 353, 356, 364, 367, 380, 480, 527, 588, 589, 592, 596, 736, 780, 783, 785,
593, 595, 736, 778, 866], [1, 2, 6, 10, 11, 17, 19, 21, 25, 34, 36, 39, 48, 60, 62, 105, 110, 111, 141, 1
235, 236, 237, 246, 253, 256, 261, 265, 266, 272, 276, 277, 282, 288, 293, 296, 300, 315, 317, 318, 329,
555, 588, 590, 592, 593, 595, 608]
```

frequent itemsets:




```
(1), (32), (50), (110), (150), (260), (296), (318), (344), (356), (364), (380), (457), (480), (527), (588
(1, 356), (260, 356), (296, 318), (296, 356), (296, 480), (296, 527), (296, 593), (318, 356), (318, 593),
```

negative border:

```
(2), (3), (4), (5), (6), (7), (8), (9), (11), (12), (13), (14), (15), (16), (17), (18), (19), (20), (21),
(41), (42), (43), (44), (45), (46), (47), (48), (49), (52), (54), (55), (57), (58), (59), (60), (61), (62
(81), (82), (83), (85), (86), (87), (88), (89), (92), (93), (94), (95), (96), (97), (98), (99), (100), (1
(122), (123), (125), (126), (130), (133), (135), (137), (140), (143), (145), (147)
```

INF 553 – Fall 2018

4. Submit a zip file that includes a output folder and your Toivonen.py. Name the zip file as Lastname_Firstname_hw2.zip. For example, Smith_John_hw2.zip.

▼		Ruotian_Jiang_hw2	Today at
▶		output	Today at
		Toivonen.py	Sep 18, 2