Sara Di Bartolomeo – 1494990 – dibartolomeo.sara@gmail.com

Homework 3

Problem 1

To avoid changing the input by hand every time I changed test files, I used two different versions of the code to be used on retail.dat and webdocs.dat.

Therefore, the python files that contain the different versions of apriori are:

aprioriretail.py, aprioriirandomizedretail.py, aprioriwebdocs.py and aprioriirandomizedwebdocs.py.

You will also find two new text files, auxfile1.dat and auxfile2.dat. This is because the program rewrites the input file on one of the two auxfiles, removing all the lines that do not contain any frequent itemset and all the items in each line that do not appear in any frequent itemset, so that in the next iteration of apriori the lines to read (and the items in each line) will be a smaller number. Now, this isn't very noticeable in retail.dat, but it made me spare a few minutes on webdocs.dat.

How it works:
1) Reads the file line by line and stores in a dictionary all the items it finds, using the item id as key and the number of occurrences as value. Increments the occurrences every time a new occurrence of the item is found.
2) Discards all the single items that do not occur at least t times (t = threshold). Rewrites the input file on an auxfile removing all the non frequent items and the lines that do not contain any frequent item.
3) Reads the auxfile line by line. For each line, computes all the combinations of k=2 elements of the elements in the line. Stores each pair found in a dictionary using the same logic as before: key=itemset, value=number of occurrences.
4) Discards all the pairs that aren't frequent enough. Adds all the frequent pairs to the “frequent itemsets” array to be returned in the end. Discards all the elements in the document that do not appear in any frequent itemsets, items that aren't in any frequent itemset in a single line, lines that do not contain any frequent itemsets and lines that contain only k=2 items (or less) since they will not be useful in the next iterations.
5) Repeats steps 3 and 4 incrementing k every time until no more frequent itemsets can be found.

In the following pages you will find an analysis of the algorithm's output from the two given files: retail.dat and webdocs.dat.

Sara Di Bartolomeo – 1494990 – dibartolomeo.sara@gmail.com

About retail.dat:

retail.dat is made of 88162 lines, and there are 16471 single elements. Of this single items, only 186 of them appear at least 500 times.

Apriori typical output on retail.dat:
Threshold = 500

finding single items...
elapsed time: 0.714933156967
number of single items found: 16471
removing items under threshold...
186 single items occurring at least 500 times found

finding k = 2 sized frequent itemsets...
elapsed time: 2.30170202255
191 frequent itemsets of size k=2 found

finding k = 3 sized frequent itemsets...
elapsed time: 20.2984440327
79 frequent itemsets of size k=3 found

finding k = 4 sized frequent itemsets...
elapsed time: 4.25893211365
13 frequent itemsets of size k=4 found

finding k = 5 sized frequent itemsets...
elapsed time: 0.351223945618
0 frequent itemsets of size k=5 found

total time: 27.9563510418
total number of frequent itemsets: 283

('48', '89') occurrences: 2798
('38', '41') occurrences: 3897
('3270', '39') occurrences: 576
('32', '65') occurrences: 774
('48', '604') occurrences: 687
('41', '65') occurrences: 995
('39', '677') occurrences: 635
('389', '39') occurrences: 544
('170', '39') occurrences: 2059
('39', '589') occurrences: 708
('310', '41') occurrences: 719
('264', '48') occurrences: 534
('338', '39') occurrences: 742
('16217', '39') occurrences: 728
('39', '548') occurrences: 663
('270', '271') occurrences: 680
('41', '475') occurrences: 570
('36', '38') occurrences: 2790
('13041', '39') occurrences: 688
('270', '39') occurrences: 1194
('48', '65') occurrences: 2529
('371', '38') occurrences: 767
('117', '48') occurrences: 601
('161', '39') occurrences: 622
('38', '89') occurrences: 764
('48', '549') occurrences: 513
('147', '48') occurrences: 1036
('237', '39') occurrences: 1929
('225', '48') occurrences: 1736
('301', '48') occurrences: 658
('156', '39') occurrences: 582
('229', '39') occurrences: 503
('334', '39') occurrences: 506
('413', '48') occurrences: 1135
('1146', '41') occurrences: 501
('39', '79') occurrences: 1111
('39', '570') occurrences: 558
('38', '56') occurrences: 514
('264', '39') occurrences: 599
('271', '39') occurrences: 1434
('39', '740') occurrences: 759
('1578', '39') occurrences: 576
('48', '9') occurrences: 790
('16010', '16011') occurrences: 651
('270', '48') occurrences: 957
('14098', '48') occurrences: 698
('201', '48') occurrences: 674
('32', '39') occurrences: 8455
('258', '39') occurrences: 628
('39', '45') occurrences: 559
('185', '39') occurrences: 828
('225', '39') occurrences: 2351
('48', '570') occurrences: 514
('39', '522') occurrences: 648
('229', '48') occurrences: 545
('475', '48') occurrences: 1428
('1393', '48') occurrences: 669
('179', '39') occurrences: 653
('39', '89') occurrences: 2749
('270', '41') occurrences: 508
('39', '533') occurrences: 922
('1004', '48') occurrences: 614
('36', '48') occurrences: 1416
('39', '544') occurrences: 524
('14098', '39') occurrences: 801
('38', '790') occurrences: 508
('39', '649') occurrences: 531
('48', '589') occurrences: 628
('12925', '48') occurrences: 817
('39', '783') occurrences: 640
('37', '39') occurrences: 707
('123', '48') occurrences: 797
('48', '533') occurrences: 861
('310', '32') occurrences: 541
('39', '604') occurrences: 775
('175', '39') occurrences: 631
('1393', '39') occurrences: 789
('48', '824') occurrences: 701
('258', '48') occurrences: 643
('39', '592') occurrences: 723
('39', '824') occurrences: 703
('3270', '48') occurrences: 592
('110', '48') occurrences: 1380
('105', '38') occurrences: 643
('36', '39') occurrences: 2037
('1146', '48') occurrences: 811
('15832', '39') occurrences: 718
('242', '39') occurrences: 564
('110', '38') occurrences: 2725
('13041', '48') occurrences: 602
('16217', '48') occurrences: 647
('2958', '48') occurrences: 779
('438', '48') occurrences: 1025
('39', '9') occurrences: 832
('32', '89') occurrences: 713
('2238', '39') occurrences: 1287
('10515', '48') occurrences: 524
('48', '49') occurrences: 843
('117', '39') occurrences: 554
('18', '39') occurrences: 540
('237', '48') occurrences: 1682
('38', '65') occurrences: 643
('16010', '39') occurrences: 829
('48', '544') occurrences: 504
('1146', '39') occurrences: 983
('101', '39') occurrences: 1400
('39', '479') occurrences: 580
('31', '39') occurrences: 538
('15832', '48') occurrences: 585
('19', '39') occurrences: 592
('110', '41') occurrences: 677
('271', '48') occurrences: 1090
('147', '39') occurrences: 1137
('39', '405') occurrences: 500
('36', '41') occurrences: 700
('225', '38') occurrences: 681
('170', '41') occurrences: 805
('405', '48') occurrences: 525
('2238', '48') occurrences: 955
('225', '32') occurrences: 655
('10515', '39') occurrences: 528
('32', '48') occurrences: 8034
('48', '60') occurrences: 815
('12925', '39') occurrences: 938
('48', '783') occurrences: 583
('170', '32') occurrences: 540
('41', '48') occurrences: 9018
('101', '48') occurrences: 1311
('48', '79') occurrences: 893
('179', '48') occurrences: 624
('41', '89') occurrences: 732
('37', '38') occurrences: 1046
('39', '49') occurrences: 768
('286', '48') occurrences: 591
('170', '48') occurrences: 1557
('103', '39') occurrences: 512
('286', '38') occurrences: 1116
('249', '48') occurrences: 675
('32', '41') occurrences: 3196
('338', '48') occurrences: 779
('237', '41') occurrences: 597
('310', '39') occurrences: 1852
('110', '39') occurrences: 1759
('48', '740') occurrences: 520
('1135', '48') occurrences: 541
('39', '475') occurrences: 1500
('39', '78') occurrences: 774
('48', '522') occurrences: 582
('170', '38') occurrences: 3031
('255', '39') occurrences: 1057
('237', '32') occurrences: 568
('39', '76') occurrences: 529
('271', '41') occurrences: 584
('161', '48') occurrences: 517
('38', '39') occurrences: 10345
('301', '39') occurrences: 715
('2958', '39') occurrences: 655
('479', '48') occurrences: 513
('2238', '41') occurrences: 570
('249', '39') occurrences: 752
('39', '48') occurrences: 29142
('371', '39') occurrences: 532
('48', '548') occurrences: 658
('1327', '39') occurrences: 1156
('237', '38') occurrences: 683
('310', '48') occurrences: 1692
('39', '60') occurrences: 983
('16010', '48') occurrences: 730
('48', '677') occurrences: 666
('38', '55') occurrences: 657
('23', '39') occurrences: 512
('38', '48') occurrences: 7944
('185', '48') occurrences: 816
('1327', '48') occurrences: 968
('225', '41') occurrences: 877
('39', '65') occurrences: 2787
('32', '38') occurrences: 2833
('48', '78') occurrences: 824
('48', '592') occurrences: 743
('39', '956') occurrences: 600
('39', '438') occurrences: 1260
('19', '48') occurrences: 591
('1004', '39') occurrences: 583
('201', '39') occurrences: 669
('255', '48') occurrences: 1057
('37', '48') occurrences: 565
('39', '413') occurrences: 1130
('123', '39') occurrences: 726
('286', '39') occurrences: 750
('39', '41') occurrences: 11414
('48', '956') occurrences: 540
('338', '39', '48') occurrences: 516
('1327', '39', '48') occurrences: 720
('170', '38', '39') occurrences: 2019
('2238', '39', '48') occurrences: 788
('101', '39', '48') occurrences: 946
('39', '48', '533') occurrences: 632
('110', '38', '39') occurrences: 1740
('39', '48', '78') occurrences: 610
('1393', '39', '48') occurrences: 507
('39', '48', '79') occurrences: 699
('32', '48', '89') occurrences: 573
('271', '39', '48') occurrences: 827
('38', '39', '48') occurrences: 6102
('110', '38', '41') occurrences: 666
('2958', '39', '48') occurrences: 570
('32', '39', '41') occurrences: 2359
('286', '38', '48') occurrences: 581
('225', '39', '48') occurrences: 1400
('39', '48', '65') occurrences: 1797
('32', '38', '48') occurrences: 1646
('371', '38', '39') occurrences: 526
('39', '48', '9') occurrences: 546
('147', '39', '48') occurrences: 742
('310', '39', '48') occurrences: 1347
('170', '38', '48') occurrences: 1538
('36', '39', '41') occurrences: 572
('225', '38', '39') occurrences: 535
('36', '38', '48') occurrences: 1360
('39', '48', '89') occurrences: 2125
('237', '38', '39') occurrences: 512
('110', '39', '48') occurrences: 1037
('170', '39', '41') occurrences: 624
('249', '39', '48') occurrences: 510
('39', '475', '48') occurrences: 1092
('225', '39', '41') occurrences: 726
('38', '39', '89') occurrences: 589
('16010', '39', '48') occurrences: 529
('270', '39', '48') occurrences: 733
('39', '48', '592') occurrences: 515
('310', '39', '41') occurrences: 625
('1146', '39', '48') occurrences: 623
('39', '48', '49') occurrences: 617
('39', '48', '60') occurrences: 609
('32', '39', '65') occurrences: 512
('37', '38', '39') occurrences: 684
('310', '41', '48') occurrences: 547
('39', '41', '48') occurrences: 7366
('170', '32', '38') occurrences: 532
('32', '39', '89') occurrences: 532
('39', '48', '604') occurrences: 520
('39', '41', '89') occurrences: 619
('36', '38', '39') occurrences: 1945
('41', '48', '65') occurrences: 663
('39', '41', '65') occurrences: 792
('237', '39', '48') occurrences: 1244
('39', '438', '48') occurrences: 780
('170', '38', '41') occurrences: 794
('38', '39', '41') occurrences: 3051
('39', '413', '48') occurrences: 781
('32', '39', '48') occurrences: 5402
('110', '38', '48') occurrences: 1361
('32', '38', '41') occurrences: 805
('286', '38', '39') occurrences: 728
('225', '41', '48') occurrences: 537
('38', '48', '89') occurrences: 578
('37', '38', '48') occurrences: 557
('12925', '39', '48') occurrences: 588
('32', '41', '48') occurrences: 2063
('36', '38', '41') occurrences: 671
('110', '39', '41') occurrences: 515
('14098', '39', '48') occurrences: 540
('255', '39', '48') occurrences: 813
('123', '39', '48') occurrences: 509
('170', '39', '48') occurrences: 1206
('36', '39', '48') occurrences: 1116
('41', '48', '89') occurrences: 606
('185', '39', '48') occurrences: 555
('32', '38', '39') occurrences: 1840
('38', '41', '48') occurrences: 2374
('38', '39', '41', '48') occurrences: 1991
('39', '41', '48', '89') occurrences: 522
('36', '38', '39', '48') occurrences: 1080
('170', '38', '39', '48') occurrences: 1193
('36', '38', '39', '41') occurrences: 553
('39', '41', '48', '65') occurrences: 547
('32', '38', '41', '48') occurrences: 540
('32', '38', '39', '48') occurrences: 1236
('110', '38', '39', '41') occurrences: 511
('110', '38', '39', '48') occurrences: 1031
('32', '38', '39', '41') occurrences: 622
('170', '38', '39', '41') occurrences: 615
('32', '39', '41', '48') occurrences: 1646

Sara Di Bartolomeo – 1494990 – dibartolomeo.sara@gmail.com

How randomized apriori works:
In the beginning the program doesn't read all the lines, but samples a small fraction of them and only uses the sampled ones to estimate how many (and which) frequent itemsets there are. The randomized implementation will obviously generate some false positives and some false negatives. It then runs a second part in which it reads the entire file checking if the previous part got any false positives: it is a simplified version of apriori in which only the previously found frequent itemsets are considered.

The first part will use a much lesser threshold because, obviously, lesser lines mean lesser occurrences.

This is what happens if we keep the threshold at 50:

> Randomized Apriori typical output on retail.dat:
> Threshold = 50 (sampling probability * original threshold)
> Sampling probability = 0.1
>
> sampling file...
>
> elapsed time: 0.0203039646149
> finding single items...
> elapsed time: 0.0348508358002
> number of single items found: 10282
> removing items under threshold...
> 198 single items occurring at least 50.0 times found
>
> finding k = 2 sized frequent itemsets...
> elapsed time: 0.117894172668
> 212 frequent itemsets of size k=2 found
>
> finding k = 3 sized frequent itemsets...
> elapsed time: 0.879712104797
> 92 frequent itemsets of size k=3 found
>
> finding k = 4 sized frequent itemsets...
> elapsed time: 0.233855009079
> 16 frequent itemsets of size k=4 found
>
> finding k = 5 sized frequent itemsets...
> elapsed time: 0.0220558643341
> 0 frequent itemsets of size k=5 found
>
> **number of frequent itemsets: 320**
>
> **verifying randomized results...**
> **true frequent itemsets found: 265**
> **elapsed time: 11.1654407978**

So, this randomized implementation has only produced 72 correct frequent itemsets. At first, it found 281, but when the second part of the algorithm – the part that checks if the frequent itemsets found by randomized apriori really are frequent – was executed, it only found 72 of them to be truly frequent.

Since it would be very expensive to find the false negatives, we prefer twitching the threshold to include in the output of the randomized part a lot more false positives and less false negatives.

> Randomized Apriori typical output on retail.dat:
> Threshold = 45 (sampling probability * original threshold * twitching)
> Sampling probability = 0.1
>
> ...
> **number of frequent itemsets: 359**
>
> **verifying randomized results...**
> **true frequent itemsets found: 278**
> **elapsed time: 9.06151580811**

By twitching the threshold, we obtain more and more false positives in the output of randomized apriori. This means that it will:
- Require more time for randomized apriori and checking part.
- Obtain more accurate results.

Let's see now how it works on a much bigger file so that the time difference will be much more noticeable:

About webdocs.dat:

Webdocs.dat has 1692082 lines and 5693364 single items.

Apriori typical output on webdocs.dat:
Threshold = 500000

finding single items...
elapsed time: 64.7349300385
number of single items found: 5693364
removing items under threshold...
22 single items occurring at least 500000 times found

finding k = 2 sized frequent itemsets...
elapsed time: 92.0613751411
65 frequent itemsets of size k=2 found

finding k = 3 sized frequent itemsets...
elapsed time: 284.212777138
64 frequent itemsets of size k=3 found

finding k = 4 sized frequent itemsets...
elapsed time: 477.428351879
29 frequent itemsets of size k=4 found

finding k = 5 sized frequent itemsets...
elapsed time: 118.96787715
6 frequent itemsets of size k=5 found

finding k = 6 sized frequent itemsets...
elapsed time: 22.7897229195
0 frequent itemsets of size k=6 found

total time: 1061.24191999
total number of frequent itemsets: 164

('122', '146') occurrences: 640230
('122', '51') occurrences: 762448
('122', '8') occurrences: 1227876
('122', '308') occurrences: 561212
('49', '51') occurrences: 697664
('51', '8') occurrences: 743087
('121', '124') occurrences: 563805
('122', '49') occurrences: 1115102
('122', '379') occurrences: 547301
('49', '84') occurrences: 509170
('308', '49') occurrences: 520417
('122', '516') occurrences: 836156
('49', '514') occurrences: 508823
('1', '122') occurrences: 528270
('379', '49') occurrences: 513813
('150', '8') occurrences: 559739
('60', '8') occurrences: 612275
('122', '84') occurrences: 553657
('121', '49') occurrences: 656200
('8', '878') occurrences: 621832
('22', '8') occurrences: 530503
('146', '49') occurrences: 562603
('121', '516') occurrences: 582876
('122', '149') occurrences: 572890
('8', '81') occurrences: 511182
('49', '60') occurrences: 576949
('121', '8') occurrences: 698749
('516', '8') occurrences: 819487
('379', '8') occurrences: 542115
('171', '516') occurrences: 550190
('171', '8') occurrences: 723175
('122', '22') occurrences: 541990
('158', '8') occurrences: 569056
('124', '60') occurrences: 501058
('124', '51') occurrences: 572463
('8', '84') occurrences: 539506
('516', '60') occurrences: 512727
('49', '8') occurrences: 1067926
('122', '524') occurrences: 509634
('122', '514') occurrences: 544288
('49', '516') occurrences: 756934
('121', '51') occurrences: 524036
('150', '49') occurrences: 515964
('514', '8') occurrences: 534123
('122', '171') occurrences: 752142
('171', '49') occurrences: 672542
('122', '150') occurrences: 605971
('158', '49') occurrences: 524911
('121', '122') occurrences: 720291
('122', '60') occurrences: 628989
('124', '8') occurrences: 801613
('149', '8') occurrences: 539193
('146', '8') occurrences: 610431
('124', '516') occurrences: 614423
('122', '81') occurrences: 514754
('124', '171') occurrences: 530808
('122', '124') occurrences: 850323
('308', '8') occurrences: 541685
('124', '49') occurrences: 753266
('122', '878') occurrences: 639377
('149', '49') occurrences: 500449
('122', '158') occurrences: 607604
('1', '8') occurrences: 512104
('49', '878') occurrences: 586871
('51', '516') occurrences: 582962
('121', '124', '8') occurrences: 550911
('122', '22', '8') occurrences: 522661
('122', '49', '8') occurrences: 1035314
('49', '51', '516') occurrences: 547231
('49', '514', '8') occurrences: 500487
('122', '60', '8') occurrences: 605625
('122', '516', '60') occurrences: 508912
('122', '150', '8') occurrences: 534602
('124', '51', '8') occurrences: 553285
('171', '516', '8') occurrences: 539655
('121', '122', '516') occurrences: 581244
('121', '122', '8') occurrences: 695061
('122', '49', '60') occurrences: 570152
('122', '51', '516') occurrences: 578952
('171', '49', '8') occurrences: 640616
('49', '51', '8') occurrences: 674483
('124', '171', '8') occurrences: 513169
('146', '49', '8') occurrences: 531900
('122', '124', '49') occurrences: 734441
('122', '124', '51') occurrences: 563888
('122', '124', '516') occurrences: 607487
('121', '51', '8') occurrences: 515834
('124', '49', '516') occurrences: 567412
('122', '379', '8') occurrences: 533005
('122', '124', '8') occurrences: 786584
('122', '308', '8') occurrences: 520840
('122', '8', '84') occurrences: 531560
('121', '49', '516') occurrences: 548834
('122', '514', '8') occurrences: 530469
('122', '8', '878') occurrences: 600225
('122', '124', '171') occurrences: 522681
('124', '49', '51') occurrences: 536581
('121', '122', '124') occurrences: 561777
('122', '149', '8') occurrences: 515359
('122', '49', '84') occurrences: 504284
('122', '171', '49') occurrences: 658247
('122', '158', '8') occurrences: 542054
('122', '171', '8') occurrences: 700249
('516', '60', '8') occurrences: 505149
('122', '49', '51') occurrences: 685528
('171', '49', '516') occurrences: 511772
('122', '146', '49') occurrences: 545624
('379', '49', '8') occurrences: 503662
('122', '49', '878') occurrences: 568782
('121', '124', '49') occurrences: 525979
('121', '516', '8') occurrences: 574996
('122', '158', '49') occurrences: 500996
('122', '51', '8') occurrences: 731321
('49', '8', '878') occurrences: 563360
('122', '146', '8') occurrences: 588417
('124', '49', '8') occurrences: 713925
('122', '171', '516') occurrences: 543052
('122', '49', '514') occurrences: 506153
('122', '379', '49') occurrences: 507061
('49', '516', '8') occurrences: 738535
('121', '49', '8') occurrences: 642484
('122', '516', '8') occurrences: 805092
('122', '49', '516') occurrences: 746940
('124', '516', '8') occurrences: 599234
('122', '308', '49') occurrences: 500185
('51', '516', '8') occurrences: 573784
('121', '122', '49') occurrences: 653171
('121', '122', '51') occurrences: 522795
('49', '60', '8') occurrences: 561618
('121', '122', '124', '8') occurrences: 549587
('122', '516', '60', '8') occurrences: 502356
('121', '122', '49', '516') occurrences: 547869
('122', '124', '51', '8') occurrences: 551180
('122', '49', '51', '516') occurrences: 545103
('122', '124', '516', '8') occurrences: 594475
('122', '171', '49', '8') occurrences: 632561
('122', '49', '60', '8') occurrences: 558628
('122', '124', '49', '516') occurrences: 564239
('171', '49', '516', '8') occurrences: 505953
('122', '171', '49', '516') occurrences: 508727
('122', '51', '516', '8') occurrences: 571220
('121', '122', '516', '8') occurrences: 573750
('122', '124', '49', '51') occurrences: 529476
('122', '49', '51', '8') occurrences: 669483
('122', '49', '516', '8') occurrences: 731576
('122', '146', '49', '8') occurrences: 520432
('122', '171', '516', '8') occurrences: 533716
('121', '122', '51', '8') occurrences: 514974
('121', '122', '49', '8') occurrences: 640419
('122', '124', '171', '8') occurrences: 507619
('122', '124', '49', '8') occurrences: 707037
('121', '49', '516', '8') occurrences: 544367
('122', '49', '8', '878') occurrences: 550369
('124', '49', '51', '8') occurrences: 523160
('121', '124', '49', '8') occurrences: 519149
('49', '51', '516', '8') occurrences: 541826
('121', '122', '124', '49') occurrences: 524899
('124', '49', '516', '8') occurrences: 559199
('121', '122', '124', '49', '8') occurrences: 518347
('122', '124', '49', '51', '8') occurrences: 521876
('122', '124', '49', '516', '8') occurrences: 557056
('122', '171', '49', '516', '8') occurrences: 503451
('122', '49', '51', '516', '8') occurrences: 540435
('121', '122', '49', '516', '8') occurrences: 543574

all the itemsets found by apriori on webdocs.dat

Let's see now the output of the randomized version:

Randomized Apriori typical output on webdocs.dat:
Threshold = 50 (original threshold * sampling probability)
Sampling probability = 0.0001

sampling file...
elapsed time: 0.713606119156
finding single items...
elapsed time: 0.00580811500549
number of single items found: 9950
removing items under threshold...
26 single items occurring at least 50.0 times found

finding k = 2 sized frequent itemsets...
elapsed time: 0.00900816917419
88 frequent itemsets of size k=2 found

finding k = 3 sized frequent itemsets...
elapsed time: 0.0316860675812
120 frequent itemsets of size k=3 found

finding k = 4 sized frequent itemsets...
elapsed time: 0.0746419429779
76 frequent itemsets of size k=4 found

finding k = 5 sized frequent itemsets...
elapsed time: 0.0917589664459
21 frequent itemsets of size k=5 found

finding k = 6 sized frequent itemsets...
elapsed time: 0.00591993331909
2 frequent itemsets of size k=6 found

finding k = 7 sized frequent itemsets...
elapsed time: 0.000824928283691
0 frequent itemsets of size k=7 found

number of frequent itemsets: 307

verifying randomized results...
true frequent itemsets found: 159
elapsed time: 309.098135948

Twitching the threshold (0.9 * original threshold):

Randomized Apriori typical output on webdocs.dat:
Threshold = 45 (original threshold * sampling probability * twitching)
Sampling probability = 0.0001

...
number of frequent itemsets: 493

verifying randomized results...
true frequent itemsets found: 164
elapsed time: 471.825613022

Sara Di Bartolomeo – 1494990 – dibartolomeo.sara@gmail.com

Problem 2

The python simulation for problem 2 can be found in

simulation.py

If a value appears in a basket with probability p and there are m baskets, the expected value of occurrences of an item is m*p. Therefore, if m = 10^5 and p=0.005, a single value is expected to appear circa 500 times.

If we, instead, consider a pair, we will need the probability for an item to appear in a basket multiplied by the probability for the other item in the same pair to appear in the same basket, all this repeated by the number of basket. Therefore, the expected number of occurrences of a particular pair is m*p^2. If m=10^5 and p=0.005, a pair is expected to appear circa 2,5 times.

Simulation.py implements a simulation of the experiment using m=10^5, p=0.005 and n=2000.
This is the result of 10 simulations:

Simulation 0:
there are 25 items that occur at least ten percent more times than the expected value
there are 3 pairs that occur at least five times more than the expected value
the most frequent pair appears 13 times

Simulation 1:
there are 19 items that occur at least ten percent more times than the expected value
there are 3 pairs that occur at least five times more than the expected value
the most frequent pair appears 14 times

Simulation 2:
there are 27 items that occur at least ten percent more times than the expected value
there are 0 pairs that occur at least five times more than the expected value

Simulation 3:
there are 31 items that occur at least ten percent more times than the expected value
there are 2 pairs that occur at least five times more than the expected value
the most frequent pair appears 13 times

Simulation 4:
there are 33 items that occur at least ten percent more times than the expected value
there are 2 pairs that occur at least five times more than the expected value
the most frequent pair appears 13 times

Simulation 5:
there are 31 items that occur at least ten percent more times than the expected value
there are 0 pairs that occur at least five times more than the expected value

Simulation 6:
there are 25 items that occur at least ten percent more times than the expected value

there are 3 pairs that occur at least five times more than the expected value
the most frequent pair appears 13 times

Simulation 7:
there are 20 items that occur at least ten percent more times than the expected value
there are 1 pairs that occur at least five times more than the expected value
the most frequent pair appears 13 times

Simulation 8:
there are 31 items that occur at least ten percent more times than the expected value
there are 2 pairs that occur at least five times more than the expected value
the most frequent pair appears 13 times

Simulation 9:
there are 23 items that occur at least ten percent more times than the expected value
there are 3 pairs that occur at least five times more than the expected value
the most frequent pair appears 13 times

The obviously noticeable thing is that pairs exceed their expected value much more than single items.

A naif explanation:
Pairs are a lot more than single items. Infact, the number of possible combinations of 2000 items is:

$$\binom{2000}{2} = 1999000$$

Since they are random, there are a lot more possibilities for a pair to appear more than the expected value. Even more, the expected value of a pair is much lower than the one for a single item, giving more space for the randomness to make a pair appear much more than its expected value.

Deeper explanation:
We can consider this distribution a binomial distribution. A binomial distribution is made of a series of independent Bernoulli experiments, so experiments that can only come up as positive or negative.
Let's consider just one pair. Assume that each basket represents an experiment: the experiment is successful if the pair is in the basket, unsuccessful if the pair is not in the basket. Each pair has a probability of p^2 = 0,000025 to be in a basket.
The variance in a binomial distribution is n*p*(1-p). In this case, n is the number of experiments, i.e. the number of baskets. Therefore, the variance for this particular case is 2,4999375.

If we calculate the variance for the distribution of the single items using similar premises (n as the number of baskets, p=0.005 as the probabilty for an item to appear in a basket) we obtain 497,5.

This numbers may induce us to think that the number of occurences of the pairs should be much less distant from the expected value as they actually are. Indeed, with binomial distribution we can calculate the probability of having a pair occur over twelve times (expected value*5):

$$1 - \sum_{k=0}^{12} \binom{100\,000}{k} (0.005 \times 0.005)^k \, (1 - 0.005 \times 0.005)^{100\,000-k}$$

= 2.38*10^-6

Yet, such a small number is not negligible since we are dealing with 1999000 elements. Therefore the expected number of pairs exceeding their expected value by five times is approximatively 4,75762.

A way to deal with this problem:
In any experiment we do, we may not know if the result we obtain really are significant or are just due to a particularly unlikely chance. Moreover, dealing with this huge amount of data makes even small probabilities become relevant. Therefore we must always check that our results are significant i.e. the probability of achieving the same result by random chance is vanishing. We can do that by estimating the probability of the event happening by chance and setting thresholds in our algorithm which discard occurrences that may be due to randomness. In the above case, for example, we could establish that a pair is considered frequent if it appears above six times its expected value. (the probability of this happening is 10^-8 and so having 1999000 elements the expected number of pairs appearing so many times is way below 1 i.e. 0.02). In this case we can be reasonably sure that it's not due to random chance.