Yogiraj Awati

1. **The dataset included with this problem is publications.txt. Parse this data, and comment on how it differs from the previous file (AP Train.txt), in terms of number of publications, authors, venues, references, and years of publication.**

   Following are the metrics of publications.txt

   Count of total publications: 2146341

   Count of total authors: 1232494

   Count of total venues: 8707

   Count of total references: 528263

   Count of total years: 80

   There was also difference in the way the data was represented:
   a) Each publication record started with Index in AP_Train.txt while in publications.txt, record started with Conference. Sequencing of information per publication was altered
   b) In previous data set author was separated by ";" while in current it was separated by ","

2. **Coauthor discovery: Please use FP-Growth to analyze coauthor relationships, treating each paper as a basket of authors.**

   A) What happens when you successively decrease the support threshold using the values {1e-4, 1e-5, 0.5e-5, 1e-6}

   Frequent Items:
   Support Threshold to 1e-4: 645
   Support Threshold to 1e-5: 58956
   Support Threshold to 0.5e-5: 168270
   Support Threshold to 1e-6: NA

   Observation:

   As support decreases, the size of Frequent Items increases. This is due to the fact that, when the support decreases more items satisfy the minimum support count. This also lead to more noise in the result and I believe the size of the FP Tree is also large. I also found the program took longer time for small thresholds.

B) Keep threshold = 0.5e-5 and report the top 5 co-authors for these researchers: Rakesh Agrawal, Jiawei Han, Zoubin Ghahramani and Christos Faloutsos according to frequency.

('Printing top :', 5, ' co-authors for ', 'Rakesh Agrawal')
('Co-Author: ', (u'Ramakrishnan Srikant', 33))
('Co-Author: ', (u'Jerry Kiernan', 16))
('Co-Author: ', (u'H. V. Jagadish', 15))
('Co-Author: ', (u'Michael J. Carey', 11))
('Co-Author: ', (u'Roberto J. Bayardo Jr.', 11))

('Printing top :', 5, ' co-authors for ', 'Jiawei Han')
('Co-Author: ', (u'Xifeng Yan', 108))
('Co-Author: ', (u'Philip S. Yu', 77))
('Co-Author: ', (u'Deng Cai', 52))
('Co-Author: ', (u'Xiaofei He', 51))
('Co-Author: ', (u'Hong Cheng', 43))

('Printing top :', 5, ' co-authors for ', 'Zoubin Ghahramani')
('Co-Author: ', (u'David L. Wild', 15))
('Co-Author: ', (u'Katherine A. Heller', 13))
('Co-Author: ', (u'Michael I. Jordan', 11))

('Printing top :', 5, ' co-authors for ', 'Christos Faloutsos')
('Co-Author: ', (u'Agma J. M. Traina', 46))
('Co-Author: ', (u'Caetano Traina Jr.', 46))
('Co-Author: ', (u'Hanghang Tong', 27))
('Co-Author: ', (u'Spiros Papadimitriou', 26))
('Co-Author: ', (u'Jimeng Sun', 24))

3. Academic community discovery

a. We will now use FP growth to analyze academic communities. To do so, represent each author as a basket in which the items are the venues in which the author has at least one publication. What happens as you decrease the support threshold using values {1e-3, 0.4e-3, 1e-4}?

Frequent Itemsets for Support Threshold to 1e-3: 981
Frequent Itemsets for Support Threshold to 0.4e-3: 4355
Frequent Itemsets for Support Threshold to 1e-4: 93561

Observation:

As support decreases, the size of Frequent Items increases. This is due to the fact that, when the support decreases more items satisfy the minimum support count. This also lead to more noise in the result and I believe the size of the FP Tree is also large. I also found the program took longer time for small thresholds.

b. Keep the threshold=0.4e-3 and report results. For each area, based on seed conferences please rank the top 10 venues that authors also publish in.

('Printing top :', 10, ' data for ', 'NIPS')
('Value: ', (u'CoRR', 4235))
('Value: ', (u'ICML', 2078))
('Value: ', (u'Journal of Machine Learning Research - Proceedings Track', 1285))
('Value: ', (u'Journal of Machine Learning Research', 1247))
('Value: ', (u'UAI', 1029))
('Value: ', (u'Neural Computation', 921))
('Value: ', (u'IEEE Trans. Pattern Anal. Mach. Intell.', 714))
('Value: ', (u'CVPR', 693))
('Value: ', (u'Neurocomputing', 692))
('Value: ', (u'Neural Networks', 596))

('Printing top :', 10, ' data for ', 'KDD')
('Value: ', (u'CoRR', 940))
('Value: ', (u'ICDM', 862))
('Value: ', (u'CIKM', 715))
('Value: ', (u'IEEE Trans. Knowl. Data Eng.', 598))
('Value: ', (u'SDM', 597))
('Value: ', (u'ICML', 534))
('Value: ', (u'WWW', 503))

('Printing top :', 10, ' data for ', 'INFOCOM')
('Value: ', (u'GLOBECOM', 14179))
('Value: ', (u'ICC', 13334))
('Value: ', (u'IEEE Journal on Selected Areas in Communications', 7051))
('Value: ', (u'IEEE/ACM Trans. Netw.', 6754))
('Value: ', (u'Computer Networks', 6569))
('Value: ', (u'CoRR', 6546))
('Value: ', (u'Computer Communications', 5251))
('Value: ', (u'WCNC', 2789))
('Value: ', (u'IEEE Transactions on Wireless Communications', 2627))
('Value: ', (u'IEEE Communications Letters', 1763))

('Printing top :', 10, ' data for ', 'VLDB')
('Value: ', (u'ICDE', 6754))
('Value: ', (u'SIGMOD Conference', 5069))
('Value: ', (u'IEEE Trans. Knowl. Data Eng.', 1902))
('Value: ', (u'IEEE Data Eng. Bull.', 1809))
('Value: ', (u'SIGMOD Record', 1783))
('Value: ', (u'EDBT', 1782))
('Value: ', (u'PVLDB', 1617))
('Value: ', (u'CoRR', 1292))
('Value: ', (u'CIKM', 1234))
('Value: ', (u'VLDB J.', 1199))

('Printing top :', 10, ' data for ', 'ACL')
('Value: ', (u'COLING', 1013))
('Value: ', (u'LREC', 752))
('Value: ', (u'CoRR', 748))
('Value: ', (u'EMNLP', 587))
('Value: ', (u'HLT-NAACL', 571))
('Value: ', (u'INTERSPEECH', 527))

COLLABRATION:
Discussed strategy for problems and installation with Ashish Kalbhor and Rushikesh Badami