University of Ioannina,
Department of Computer Science & Engineering

Data Mining – CSE 012/CS 059
Spring 2017

Assignment for September Exam

The deadline for the fourth assignment is September 24, before the end of the day. Turn in the code, with instructions on how to run it. The report should include detailed observations on the results. For late submissions the late policy on the web page will be applied. Details for turn-in will be posted on the web page of the class. There will be an oral examination of the Assignment in the week following the deadline.

Question 1

Assume that you are given as input a table with *n* rows and *m* columns, with 0/1 values. You want to find all *(r,c)-tiles* of 1's, that is, sets of *r* rows and *c* columns such that the corresponding submatrix has all 1's. Note that tiles may be overlapping. Describe an efficient algorithm for solving this problem that makes use of the APriori idea.

Question 2

Prove that for an undirected graph the stationary distribution of a random walk is proportional to the degree of the nodes. If P is the transition matrix of the random walk, and π is the stationary distribution for which $\pi = \pi P$, show that for node i the probability π_i is proportional to d_i where d_i is the number of edges incident on node i.

Question 3

In Assignments 1 and 2 you were asked to implement frequent-itemset and recommendations algorithms and apply them on a set of tweets. The dataset contained only users that were Trump followers. In Assignment 3 you were given an additional dataset that contains both Trump and Clinton followers. Apply your code for both problems on this new dataset, and report on the results, exactly as in the first two Assignments. Report also any differences with the previous experiments on the dataset that contained only Trump followers. Do you see itemsets in the results that are characteristic of each group? Does the recommendation accuracy improve?

You can correct or improve your code from assignments 1 and 2 if necessary. Report the changes, and submit the updated code.

Question 4

In this question you will consider the problem of predicting which candidate a user follows, using the retweet graph.

Use the retweet graph you created for Assignment 4. Use a 80/20 split of the nodes. Consider the nodes in the 80% as labeled and the rest as unlabeled. Then do label propagation on the graph to find the label of the unlabeled nodes. For this, assign a value -1 to the Trump followers, and +1 to the Clinton followers. Do the label propagation as we described in Lecture 11: the labeled nodes are absorbing, and for the non-labeled nodes with compute a value with the iterative process we described in Lecture 11. The process continues until there is essentially no change in the values. Assign label +1 to the nodes with positive value and -1 to the nodes with negative value. Compute the accuracy of the prediction and produce a confusion table. Compare with the results of the classifier that you created in Assignment 4, which is trained on the 80% of the labeled users.

In the second part of the question you will create a new retweet graph with who tweets whom, but this time you will also keep users that are not in the file with the class labels. Apply the same iterative filtering process to keep nodes with degree greater than 10, and keep the largest connected component (the resulting graph should have about 37K nodes). Apply the same label propagation process as before on this new graph, using the labels of the nodes that are in the class file, and compute a value for the nodes that are not in the filw. Use the values to categorize nodes as Trump followers, Clinton followers, or Neither followers (for example you may label the nodes as Trump followers if they have value less than -0.5, Clinton followers if they have value greater than 0.5, and Neither if it is in between – choose the thresholds appropriately depending on the values that you have in the data).Then, sample 20 random nodes from each class and manually examine them to obtain the ground-truth labeling of the nodes. For the evaluation, study profile and the tweets of the user, and justify your labeling of the nodes. Compute the precision of the algorithm for each class on the sample.

Hand-in your code and your report. In your report describe at a high level your implementation, and in detail your results. You should also give details about the labeling of the nodes, and how you reached the conclusion for the ground truth label.