University of Ioannina,
Department of Computer Science & Engineering

Data Mining – CSE 012/CS 059
Spring 2017

Assignment 1

This is the first assignment. The deadline for the assignment is November 17, 11:59 pm. Turn in the code and your results, and submit the remaining questions either electronically, or on paper. The report should be written electronically For late submissions the late policy on the page of the course will be applied. Details for the turn-in, and how to write reports are on the Assignments web page of the course.

Question 1 (Reservoir Sampling)

In this question you are required to modify the Reservoir Sampling algorithm to sample $\boldsymbol{K}$ items from a stream of N items, uniformly at random, so that each element has probability K/N to appear in the sample.

1. Describe the algorithm for sampling K items uniformly at random from a stream of N items. The algorithm should work in a single pass over the data, reading the items one by one, without prior knowledge of the size of the stream N, and using $O(K)$ of memory (assume the size of an item is fixed). **Do not** write code or pseudocode for this part; just explain the logic of the algorithm in English in a simple way.
2. Prove that your algorithm produces a uniform sample, that is, for every $i, 1 \leq i \leq N$, the i-th element has probability K/N to appear in the sample.
3. Write a program **in Python** that implements the sampling algorithm. Your program should sample K lines from a text document. It should be possible to use the program from command line. It should take as command line argument the value of K, read lines from the standard input, and output the sample in the standard output. For example the following command should print a random sample of 10 lines from the file input.txt:
   "`sample.py 10 < input.txt`".

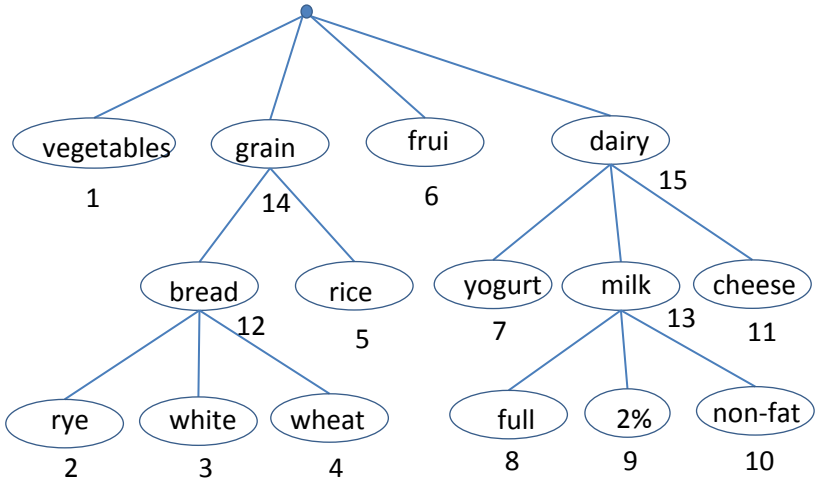
Question 2

On the Assignments page of the course there is a file "data.csv". The file contains three comma-separated columns of 1000 values, with column names A, B, and C. The values in B and C are a function of those in A. Your goal is to find the function of columns B, C, and column A. To find this function create plots of B and C against A, as described in class. Hand in an Iron Python Notebook, which should contain the code for processing the data, the plots and computations that you did, and a report with your conclusions.

Question 3

In the image below you are given a simple food taxonomy tree. Every leaf of the tree corresponds to a simple item. Every internal node corresponds to a general category, or high-level item. Every node is represented by a number that is shown in the tree. Answer the following questions.

1. Let $X = \{x_1, x_2, \dots, x_k\}$ be a frequent itemset. Suppose that we replace item x_i with one of its ancestors in the taxonomy tree (if it has any), so that we get an new itemset X'. What is the relationship between the support of X and X'?
2. Consider the database of simple items shown in the table below. Use the support threshold minsup = 7/8. Find all frequent itemsets consisting of high-level items in the taxonomy (non-leaves). Keep in mind that if a simple item appears in a transaction then we assume that all ancestors of the item also appear in the transaction.
3. Based on the methodology you used in 2, propose an algorithm for finding frequent itemsets including high level items for data organized in a taxonomy. Your algorithm should be efficient (not go through all itemsets).



tid	items
1	2 3 6 7
2	1 3 4 6 8 11
3	3 9 11
4	1 5 6 7
5	1 3 8 10 11
6	3 5 7 9 11
7	4 6 8 10 11
8	1 3 5 8 11

Question 4

For this question you will use the data from Yelp Academic Challenge dataset. You can find them here. We are not interested in the data about images. It is recommended that you use the data in JSON format, for easier processing.

For the question we will focus on the city of Toronto. Our goal is to understand the characteristics of the different neighborhoods of Toronto, using the Yelp data. From the file business.json keep the businesses in the city of Toronto for ahcih we have information about the neighborhood, and about the categories they belong to. Keep the neighborhoods with at least 100 businesses.

Using this data, propose a methodology for finding for each neighborhood the 10 categories that best describe it. The categories you find should be frequent enough in the neighborhood, and more frequent than in other neighborhoods (e.g., the category Food is not interesting since it appears in all neighborhoods with approximately the same frequency – on the other hand, we do not care about very rare categories). Your methodology should define a metric for how important is a category for a neighborhood. You can process the data in different ways, and throw out data that you believe do not offer useful information. Make your choices clear in your report.

Report for each neighborhood the categories output by your methodology (simple reporting is OK, but if you decide to do some kind of visualization it will be counted in your favor). Then for each neighborhood find out its true characterstics, and report the whether the categories that you found describe accurately the neighborhood. Use the web for this task: there are many online tourist guides that give neighborhood information, and Airbnb has summaries for the different neighborhoods.

In your report, include a description of your methodology, your results, and a commentary on the results.

Bonus: For many businesses, there is a field Ambiance with different keywords that describe the ambiance of the business, and a true/false value. Apply your methodology using these keywords instead of the categories, and comment on the description of the neighborhood you get with these keyowrds.