University of Ioannina,
Department of Computer Science & Engineering

Data Mining – CSE 012/CS 059
Fall 2019

Assignment 1

This is the first assignment. The deadline for the assignment is November 21, 11:59 pm. Turn in the code and your results, and submit the remaining questions either electronically, or on paper. Whenever a report is required, it should be written electronically. In some cases, the report may be in the notebook that is submitted, together with the computations. 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. The assignments should be done **individually**.

Question 1

A. 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. 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).

1. Describe the algorithm for sampling K items uniformly at random from a stream of N items. **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`".

B. Describe an algorithm for sampling a random node from a tree, for which you do not know in advance the number of nodes it has, and you cannot store it in memory.

Question 2

Let $x_1, x_2, \ldots, x_n$, n real values, sorted in increasing order, and assume that $n = 2k + 1$ is an odd number. Show that the real number z that minimizes the sum of distances

$$\sum_{i=1}^{n} |x_i - z|$$

is the median value x_{k+1}.

Hint: Place z on the median, and examine what happens when you move it to the left or right.

Question 3

You are given the table below with the preferences of users for movies, and the ratings of the users for each movie.

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	Movie 6
User X	5	3		1		?
User Y	4	2	1			1
User Z	3			1	3	3
User W	2	5	1	5	3	4

Our goal is to predict the value (X,6) of User X for Movie 6.

You will use the User-User Collaborative Filtering algorithm, in two versions:
In the first version you will use the preference table as is, you will compute the cosine similarity between X and the rest of the users, and you will compute the rating for the cell (X,6) as the weighted average of the ratings of the two most similar users to X.
In the second version you will first subtract the mean rating from each row, you will compute again the cosine similarity between X and the rest of the users, and you will now compute the divergence from the mean for the cell (X,6) as the weighted average of the divergence of the two most similar users. You will compute the rating by adding this estimated divergence to the mean value of X.
You can do these computations either in a program, or by hand. Report your results (similarities, estimated rating and divergence) for both versions in your report.

What do you observe? Which cells are more or less important for the computation of similarity in each version? How do the nearest neighbors change? What is the rating you compute? Explain your observations.

Question 4

From the site of the book The Data Science Manual, in the data page, download the Movie Data. The data is in a csv file, where each row is a movie, and the columns are attributes, described in Description link. There are data for 3201 movies. The goal is to analyze this data, understand their form, pose some questions, and answer them using the data.

Load the data into a Pandas Dataframe. The data, like all real data, are incomplete and they have noise, so you will need to do some data cleaning whenever necessary. There are many ways to handle these cases, make your decisions clear, and write clearly how these decisions affect the data (e.g., if you remove some data, clearly state

the criteria you used, and how much data you removed; if you replace missing values, state clearly the replacement value and how it affects your results). It would also be a good idea to change the attribute names so that they are easier to work with.

You will consider the following questions:

1. Create histograms for the attributes “Worldwide Gross” (the gross revenue of the movie worldwide), “Rotten Tomatoes Rating”, “IMDB Rating”, “IMDB votes” (be careful with “Worldwide Gross” which is stored as a string). What do you observe? How do the value distributions look like?
   For the attribute “Major Genre” compute for each genre the number of movies with this genre. Create a table and a bar plot that shows this distribution. What do you observe for the popularity of the genres?
   For the above you will use the Pandas library.
   Then, for the attributes “Worldwide Gross” and “IMDB Votes” you will create your own histogram, and you will plot it in log-log scale. For your histogram, you will create bins that will grow exponentially. For the plot on the x axis you will have the mean value of the bin interval (revenue or votes), and on the y axis the number of movies that fall within this bin. Create a scatter plot with the values, with logarithmic scale on both axes. Κάνετε ένα scatter plot με τις τιμές. What do you observe? What kind of distribution do the revenue and the IMDB votes follow?
2. After the histogram analysis we will study correlations between attributes. First, we will consider the numerical attributes “Worldwide Gross”, “Rotten Tomatoes Rating”, “IMDB Rating”, “IMDB votes”. The first represents the popularity of the movie in the cinemas, while the last the popularity of the movie online. The other two represent the quality of the movie as it is being judged by the experts and the public. You will consider the correlation of all pairs of attributes. Produce a scatter plot for each pair (Pandas allow to put all figures in a single plot if you wish), and compute the Pearson and Spearman correlation coefficients, as well as their p-values. What correlations do you observe and how strong are they? What do you observe in the scatter plots with “World Gross”? How can you make these plots visually clearer?
   Then you will study if there is a correlation between the genre and the movie revenue. You can ignore very rare genres, but make your choices clear in your report. Create a bar plot (with confidence intervals) with the mean worldwide gross for each genre. Use the t-test to examine which of the differences you observe are statistically significant. What is your conclusion?
3. It is often said “The no longer make good movies like they used to”. Use the attributes Release date, Rotten Tomatoes Rating and IMDB Rating to study if this is true. Create a plot with the mean quality (as it is measured by the Rotten Tomatoes Rating and the IMDB Rating) and plot how it changes over different decades (you can ignore decades with too little data). What do you observe? What are some possible explanations for these plots?
4. Pose a question of your own and explore it with data. State clearly the question you will study, the measurements you will do to answer it, present plots and numbers that show the analysis you did and support your answer (which may be negative, but even in this case there should data to support it), and discuss your findings.

Turn in a notebook with all the computations and plots and the report text. If you want you can have the report in a separate text, but whenever plots are needed to make a point they should be in the text.

Question 5

Google trends is a service by Google that allows to study the query frequency in the search engine over time (and over space for some coutries). For example, in this blog post, they use Google trends to study the time and day that people decide to break up (also, the blog post has information about the library Pytrends that allows to pull data from Google trends). Pose a question of your own and explore it with Google trends data. The question can be about anything (e.g., politics, sports, TV, everyday life, etc.). In your report state clearly the question you will study, the measurements you will do to answer it, present plots and numbers that show the analysis you did and support your answer (which may be negative, but even in this case there should data to support it), and discuss your findings.

In your analysis you can use a library like Pytrends or do queries manually to collect the data. Describe what you did in your report. Turn in your report and your code and data.