Large Scale Data Management Periklis Andritsos

Deadline for submission 12.06.2024 @ 11:59pm

Introduction

The paper aims to analyze big data and present your results. The assignment will consist of 2 parts, the first part will be writing a 6 page report and the second part will be creating code to run your experiments. The submission date will be June 12, 2024.

Written report

The report should include the following:

- Be up to 6 pages, using 11 pt font, single space. • Include the following subsections: o Introduction o Problem definition and motivation (it is good to include an example of the use of its results, e.g. who are they useful for?)

  o Brief description of the data set you used o Description of the data analysis method (specific emphasis on this) o Experimental Results (special emphasis on this) o Discussion/Critical evaluation of results o Conclusions

The data analysis subsection should describe the techniques you used and an explanation why! It is very important to try to convince the reader that a particular technique being used is the one that suits the problem. Be clear and concise.

The experimental results section should include all the experiments you used. Discuss the parameters and especially the execution times, as well as any evaluation measures used. Include tables/figures as necessary (most data analysis documents have them). Note that you are not graded on the 'beauty' of your charts, but on the message they convey and how clearly they are described.

Code

Your experiments should be done in the Python language. You can use any implementation platform, e.g. Jupyter-lab, Google Colab, IDLE, etc. The main thing is that they can be reproduced. Pay close attention to the use of comments in your code.

Note: if you use external sources, you should mention them in comments within the code.

Data sources

To prepare the work you can use data from the following indicative sources:

Google Data set search: https://
datasetsearch.research.google.com/

KDnuggets Datasets: https://
www.kdnuggets.com/datasets/index.html

kaggle Datasets: https://
www.kaggle.com/datasets

144 libraries of datasets:
https://data.world/datasets/library

What will you submit?

1. The document of the final report of the work in pdf format 2.
The Python code either in a . ipynb to either a .py file (or both)

Final score

Reminder that the final grade will be as follows o
60% from the final exam o 40%
from the assignment o
To calculate the final grade as well as the theory exam
work should have a grade above 50%.