Trier University
Department IV – Computer Sciences
Prof. Dr.-Ing. Ralf Schenkel (H 539, schenkel@uni-trier.de)

Trier, April 14, 2021

Assignment 1 of the course Big Data Analytics

Summer 2021

Deadline 10:15h on Wednesday, April 21, 2021

Task 1: Publications in dblp: Part I 15 points

In the download folder for this exercise you find the file `publications.zip`. The archive contains a CSV file that provides the number of publications of some authors in dblp. The structure of a line in the file is

`PERSON_ID, NUMBER_PUBLICATIONS`

where

- `PERSON_ID` is a unique identifier for a scientist.
- `NUMBER_PUBLICATIONS`, is the number of publiations by that scientist that are known to dblp.

In a first step:

- Read the data with a statistic tool of your choice.
- Determine basic statistical features of the number of publications, such as minimum, maximum, lower quartile, upper quartile, median and mean.
- Draw a box plot of the publication number distribution. Briefly describe the box plot in your own words. In particular, what do you learn about your data from the box plot and what are potential issues

Hand in: The statistic results, a rendering of the box plot, your description and the program (e.g., R script) used to determine the results, renderings.

Task 2: Publications in dblp: Part II 30 points

With the data from Task I. Answer the following research question:

What is the average number of publications per scientist in this example?

To this goal:

(a) Draw a histogram of the publication number distribution. Choose suitable bins for the x-axis. Briefly explain your choice of bins.

(b) Determine at least two classes of publication numbers that need to be filtered out. A class is a set of lines from the CSV file that is discarded because of a common property. For each class:

- Motivate your choice in a short description. In particular, describe how this class manifests in statistical data/visualizations-

- Name the size of the class (in number of data points) and an example.
- Describe the influence of the discarded classes on the overall result. (How do statistical properties change. In particular, how does discarding data affect the answer to the research question.

Hand in: Renderings of the histogram (png or pdf) and the program used to create it. Descriptions of the discarded classes and changes to the result.

These tasks will be discussed in the tutorial on April 28, 2021.

General information:

[redacted]