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

Trier, October 31, 2022

Assignment 1 in the course Digital Libraries and Foundations of Information Retrieval

Winter Semester 2022

Deadline 12:15h on Monday, November 7, 2022

Task 1: Digital Preservation in Libraries 8 + 7 Points

One of the tasks of „normal“ and digital libraries is preservation of objects and documents.

(a) Discuss at least two problems that non-digital libraries have with the perservation of their contents. Think for example of books or movies. How are these problems solved? Mention one approach to solve each problem.

(b) In digital libraries, two major problem areas with preservation are *storage media* and *data formats*. Describe in a few sentences which typical problems arise and how one could solve them.

Task 2: Digitization of existing documents 5 + 5 + 5 Points

One of the possible tasks of digital libraries is the digitization of non-digital artefacts.

(a) A typical example for digitization is scanning paper-based documents. But there are also other kinds of artefacts that can be digitized, i.e., converted into a digital representation. Describe at least two kinds of artefacts that are not paper-based documents, but for which digitization could make sense.

(b) Discuss at least three advantages of digitizing artefacts (including paper-based documents).

(c) Digitized documents are not always equivalent replacements for the corresponding originals. Describe at least two analyses that cannot be done with digital copies of paper-based documents or other artefacts.

Task 3: Analysis of publication titles 5+5+3+2 Points

It is an interesting question if the style of English publication titles has changed over the last decades. One possible research question here could be if the frequency of titles consisting of two parts, separated by a colon, has increased. An example for such a title is ‘Making SENSE: socially enhanced search and exploration’. In this task, you are supposed to check if this is true or not, using data on publications from dblp provided in Moodle (file `titles.zip`). In this file, each line corresponds to one publication. Each line consists of three components, separated by a tabulator character (`\t` in Java). The first component represents the type of the publication (`article` or `phdthesis`), the second the publication year, the third the title of the publication. The data is a 10% sample of all articles and PhD theses in dblp as of September 1, 2022.

(a) Write a small program (in any programming language, for example Java or Python) that computes, for each year, the fraction of publication titles from that year consisting of two parts, separated by a colon. Create a chart for that data showing how it develops over time, using a tool of your choice (for example Excel, Calc, or gnuplot). Is it true that such titles appear more frequently now than in the past?

(b) Compute, for each such title, the length of its first part in words, and generate a chart that shows the frequency of the different lengths.

(c) Which anomalies did you observe in the data that may result in fractions or frequencies that are not fully correct?

(d) Are there other patterns of titles that would be worth analyzing?

These tasks will be discussed in the meeting on November 15, 2022.

General remarks:

- The tutorial group takes place on Mondays in the regular meeting in F55 at 12:15.
- The first meeting of the tutorial group will be on November 15, 2022.
- To be admitted to the final oral exam, you need to acquire at least 50% of the points in the assignments. In addition, you need to present at least one solution of a task in a convincing way during the tutorial.
- It is preferred to submit in groups of size up to two (but not larger); in that case, only one submission is sufficient for the whole group. Write the names of all group members on your solutions.
- Solutions must be handed in before the deadline
  - in Moodle (`https://moodle.uni-trier.de/`, course `DL-IR-2022`) as as a PDF or, if submitting multiple files, as an archive (.zip or comparable).

  Submissions that arrive after the deadline will not be considered. The name of at least one group member should occur in the file name of your submission. If you want to modify a previously uploaded solution, just re-upload your solution.
- Graded versions of your submissions will be returned in Moodle until the following tutorial.
- Announcements regarding the lecture **and** the tutorial group will be done in the StudIP course for the lecture.