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

Trier, November 7, 2022

Assignment 2 in the course Digital Libraries and Foundations of Information Retrieval

Winter Semester 2022

Deadline 12:15h on Monday, November 14, 2022

Task 1: Author disambiguation in DBLP 15 Points

The following DBLP profiles contain (with high probability) multiple authors with the same name:

- `https://dblp.uni-trier.de/pid/00/969.html` (Andreas Becker)
- `https://dblp.uni-trier.de/pid/72/3357.html` (Egon Müller)
- `https://dblp.uni-trier.de/pid/43/4520.html` (Arthur Schmidt)
- `https://dblp.uni-trier.de/pid/39/3460.html` (Paul Schmidt)
- `https://dblp.uni-trier.de/pid/18/6479.html` (Ralf Schmidt)
- `https://dblp.uni-trier.de/pid/69/4068.html` (Peter Schmitt)

You can find all profiles also in Moodle; note that the profiles on the DBLP Web page may be different, use the profiles in Moodle for this task. Select **one of these profiles** and disambiguate it, i.e., determine which different persons are contained in this profile and which publications belong to them. The coauthor graph, which can be found at the end of each profile, can help you here. In addition, you can access the online version of the texts and determine the affiliations from them, if they are available. Use the DBLP key or a textual description to represent the publications.

Example: We disambiguate the profile of Thomas Schmitt (see Moodle; in DBLP, the profile has already been mostly disambiguated).

- Thomas Schmitt, Weinheim:
  `journals/it/SchmidS79`, `journals/it/SchmidS78`
- Thomas Schmitt, Technische Universität Dresden: `phd/dnb/Schmitt95`, `conf/miip/SchmittFOAF98`, `conf/parelec/KortkeSM00`
- Thomas Schmitt, Winterthur Insurance: `conf/wirtschaftsinformatik/AdomeitBS01`
- Thomas Schmitt, FH Trier: `conf/dagstuhl/OechsleS01`
- Thomas Schmitt, Stockholm Bioinformatics Centre: `journals/bmcbi/KlammerMSS09`, `journals/nar/OstlundSFKMRFS10`, `journals/bib/SchmittMSS11`, `journals/nar/AlexeyenkoSTGFS12`, `journals/nar/SchmittOS14`
- Thomas Schmitt, Institute of Automotive Engineering, Technische Universität Darmstadt: `conf/ivs/RodemerkHWWS12`
- Thomas Schmitt, INRIA: `conf/gcai/SchmittCS16`, `conf/ictai/SchmittGCS17`

Notice that the coauthor graph indicates only six different persons.

Task 2: Hirsch Index 3+4+4+4 Points

We discussed the Hirsch index (often called h-index) in the lecture. It is often used to assess the performance of scientists, but it is also often criticized.

(a) Assume that a scientist has a h-index of 11. What can you say about the number of their publications and citations? Create an exemplary list of publications and their citation counts which leads to a h-index of 11.

(b) In an appointment procedure at a university, two scientists are compared with a h-index of 20 each (computed by Google Scholar). However, they work in different areas and have different ages. Explain why the h-index is not a good benchmark in this situation; give at least two reasons.

(c) Assume that you should compare the performance of two scientists without taking their publication or citation count into account. This could be the case, for example, if you are the member of an appointment committee. Give at least three other options. Which of them can be quantified, i.e., can be explained with numbers?

(d) There are a number of alternative author-level metrics for measuring the bibliometric impact of authors. From `https://en.wikipedia.org/wiki/Author-level_metrics`, pick two metrics and discuss why they may be better suited than the plain Hirsch index.

Task 3: Classification vs. full text 5+4+6 Points

A major task of libraries is the content-based indexing of publications, for example by assigning by assigning them to classes of a classification system.

(a) Choose, from the ACM Computing Classification System 2012 (`https://dl.acm.org/ccs`), at least five classes that can be used for publications from the area of digital libraries.

(b) As an alternative to classification, it is possible to use arbitrary keywords that are not part of a classification system. Discuss at least one advantage and one disadvantage of this approach when compared to classification systems.

(c) Compare searching based on classification of publications (where the query is a class from the classification system) to searching based on arbitrary keywords for publications, based on the title of publications, and based on the full text of publications (where the query consists of arbitrary text in all three cases) with respect to the expected number and quality of results. Where do you expect most results, where the best results? Explain your answer.

Task 4: **Bonus task**: Computation of the Hirsch Index 15 Points

Write a short program in a programming language of your choice that reads the publications of an author and their citation counts from a file and computes the corresponding Hirsch index. An example input file is given in Moodle. Each line of the file contains the DBLP key of the publication and its citation count, separated by a tabulator character (`\t`). The author from the example has a Hirsch index of 7.

These tasks will be discussed on November 21, 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 14, 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.