Data Mining (CS524)

Indian Institute of Technology Ropar

Lab Assignment 2 **Release Date: 10/02/2019**
Maximum Marks: 50 **Due Date: 20/02/2019**

DBLP (`https://dblp.org/`) is an online database that contains the records of all research articles published in almost all important computer science conferences and journals. Each individual record consists of an ordered list of author names (in case of multiple authors), title of the research article, the venue and year of publication, and other associated details. The set of authors collaboratively publishing an article are called *co-authors*. Download a small DBLP dataset[1] in XML format from here (`http://tarique.in/cs524-2019/dblp50000.xml`) to perform the following tasks.

T 1. Implement the following three frequent pattern mining algorithms in C++/Java/Python.

- Apriori algorithm: *Rakesh Agrawal and Ramakrishnan Srikant, Fast algorithms for mining association rules. VLDB, pages 487-499, Chile, 1994.*
- FP-Growth algorithm: *Jiawei Han, Jian Pei, and Yiwen Yin, Mining Frequent Patterns without Candidate Generation, ACM SIGMOD, Dallas, 2000.*
- ECLAT algorithm: *Mohammad Javeed Zaki, Scalable algorithms for association mining, IEEE TKDE 12 (3): 372390, 2000.*

[*10+10+10 marks*] ■

T 2. Find all the sets of 2 or more authors who have frequently co-authored articles, using all the three implementations. Document the results, and report the running times of all the versions. Also mention the machine configuration in which the experiments are performed. [*5 marks*] ■

T 3. Generate all the strong association rules ($A \Rightarrow B$) from the frequent co-authorships obtained in task **T 2**, such that $|A| \geq 2$ and $|B| \geq 2$. Document the results.

[*5 marks*] ■

T 4. Are all the strong association rules obtained in task **T 3** interesting? If not, use a good null-invariant correlation measure to identify the interesting association rules. Document the results. [*5 marks*] ■

T 5. Find all the pairs (A, B) of *strongly negatively correlated* frequent sets of co-authors, such that such that $|A| \geq 2$ and $|B| \geq 2$. Document the results.

[*5 marks*] ■

[1] Thanks to Felix Naumann for the dataset.

Assume reasonable thresholds and mention in the ReadMe document, wherever required.

End of the Assignment

Submission Guidelines
Create a folder and name it as `entrynumber_A2`, where `entrynumber` is your own entry number. This folder should include the following.

(*i*) Properly named and organised implementation source code files (with minimum self-explanatory comments) for the tasks. Use of existing libraries and built-in functions is allowed for reading the input files, writing into the output files, performing mathematical operations, and computing basic statistical measures, like mean, median, etc. For all other computations, detailed codes are to be written (without using any libraries or built-in functions).

(*ii*) A properly named electronic file in pdf format containing documentations of all the tasks. It can be prepared using any text editor or latex, but needs to be finally converted into pdf for submission.

(*iii*) A `ReadMe` text file containing the instructions for executing the source code to obtain the desired results. Any additional information or note may also be included in this file.

Submit the `entrynumber_A2` folder in a zipped file through Moodle.

Note:

(*i*) *This is an individual assignment.*

(*ii*) *Marks distribution within each task is as follows. a) 10% for minimum comments in the source code files, b) 20% for logic and ideas of implementation, c) 60% for implementation, and d) 10% for proper documentation and organisation of the submission.*

(*iii*) *Late submissions will face a penalty of 10% (of the full marks) for each day of delay.*

(*iv*) *Presenting some other person's work as your own without proper citation of the source is an act of plagiarism. It is a serious offence and will be treated strictly.*

(*v*) *Queries, if any, can be directed to our TA Aroof Aimen (`2018csz0001@iitrpr.ac.in`) through email.*