Assignment #4

Date: 27th March, 2023

Due date: 4nd April 2023 11:59PM Total Marks: 100

Late Submission and Plagiarism Policy

- Late submissions are not encouraged. Late submission within 3-4 hours of due datewill get 30% Marks deduction. After 4 hours all submissions will be marked zero
- You need to submit the code files named after Question#, a pdf report file and a demo video of running code.
- Viva may be taken for any assignment(s) with suspected plagiarism.

In this assignment we will use DBLP dataset. DBLP is a computer science bibliography website. Starting in 1993 at the University of Trier, Germany, it grew from a small collection of HTML files and became an organization hosting a database and logic programming bibliography site. The database contains more than 6 million bibliographic entries, including articles, conference papers, and proceedings, from over 3,000 conferences and workshops, and over 1,600 journals. You need to implement this assignment using **Java** only.
DPLB: https://dblp.uni-trier.de/xml/ (Downloadable link)
The dataset can be downloaded as a single xml file. For this assignment you need to use the only the articles data from DBLP dataset. Consisting of the following attributes.

Title - The title of the article.
Authors - The name(s) of the author(s) who wrote the article.
Year - The year in which the article was published.
Pages - The starting and ending page numbers of the article in the publication.
Volume - The volume number of the journal or conference proceedings in which the article was published.
Journal/Booktitle - The name of the journal or book in which the article was published.
Url - A URL to the online version of the article (if available).
EE - An electronic edition of the article (if available).
Crossref - A cross-reference to the proceedings or book in which the article was published.
Abstract - A brief summary of the article's content.
Bibsource - The source from which the article's bibliographic information was obtained.
Biburl - A URL to a bibliographic entry for the article (if available).

Question#1 [40 Marks]

Design a MapReduce Algorithm (Java Code) to find the number of articles published in each journal per year from the DBLP articles dataset.
You should explain how the input is mapped into (key, value) pairs by the map stage, i.e., specify what is the key and what is the associated value in each pair, and, if needed, how the key(s) and value(s) are computed. Then you should explain how the output (key, value) pairs of the map stage are processed by the reduce stage to get the final answer(s). (Word count: 200 words)

Question#2 [60 Marks]

Design a MapReduce Algorithm (Pseudo-code and Java Code) to find the co-authorship graph from the DBLP dataset. In our graph each author is represented as a node. We wish to find the pair of authors who have published article(s) together. The edge weight is equal to the number of publications. You should explain how the input is mapped into (key, value) pairs by the map stage, i.e., specify what is the key and what is the associated value in each pair, and, if needed, how the key(s) and value(s) are computed. Then you should explain how the output (key, value) pairs of the map stage are processed by the reduce stage to get the final answer(s). (Word count: 200 words)

What to submit:

- You need to submit a pdf file containing the report. Attach Screenshot of the results in the report.
- Java file of each program. Follow the naming conventions I20-xxx_QNumber
- A demo video of the code