MINING OF MASSIVE DATA
SPRING 2024
ASSIGNMENT 1- MAP REDUCE

DUE DATE: 12 Feb 2024
SUBMISSION: *Upload the Source code and the output file on Google Classroom in a zip file with your roll number.*

INPUT FILE

You are given an input text file named citation.txt. It contains information regarding the research papers published in various journals. The complete file Citation-network V1 can be found at https://cn.aminer.org/citation. The format of the file is as follows:

#* --- paperTitle
#@ --- Authors
#t ---- Year
#c --- publication venue
#index 00---- index id of this paper

QUESTION: Write an efficient MapReduce program for the following problems. To make your algorithm efficient, you should use combiners or in-mapper aggregation techniques that use arrays.

1. Process the citation.txt input file and output the number of papers published in each decade: 1970s, 1980s, 1990s, 2000s, 2010s, and 2020s.

2. Create an inverted index of the citation file. Your inverted index will output the year followed by the comma-separated list of the titles of the papers published in that year.

> Sample Output format :
> Year1 -> PaperTitle, Paper Title
> Year2 -> Paper Title

3. Produce a list of co-authors of each author in the given input file.

> Sample Output (Author -> List of Co -authors)
> David Jones -> Sam Nick, Ali Javed , Daniel Brown
> Sam Nick -> David Jones, Zan Jao, Ali Javed
> Ali Javed -> David Jones ,Sam Nick
> Zan Jao -> Sam Nick
> Daniel Brown -> David Jones

4. Find the average number of papers published each year.
5. List the names of authors who have written the maximum number of papers.
6. Find the names of authors who have written at most one paper in a year.
7. Find the title of papers such that their venue is not mentioned in the input file.