Fundamentals of Big Data Analytics

Assignment-02

MAP-REDUCE

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 title of papers such that their venue is not mentioned in the input file.

Question 2: Write a MapReduce program using MRJob to find the distribution of word lengths. Specifically, for each word length (1-letter words, 2-letter words, etc.), calculate how many words of that length exist in the dataset.

Question3: Write a MapReduce program using MRJob to find the top 10 most frequent words, excluding common stopwords (like "the", "is", "and", etc.).

Note:- You can let the input for question 2 and 3 be that of a paragraph text, not a csv of words/phrases