BIG DATA ANALYTICS–Spring 2021

ASSIGNMENT 6 and Assignment 7

Due Date: 11th June 2021. Upload on Google Classroom with your roll number.

Submission Details:
For Assignment 6: Upload properly intended and commented Source code of Map Reduce program.
For Assignment 7: Upload the Word Document containing the Association rules and their explanation.

You can work alone or in a group of two. Plagiarism from other students or the internet will result in -5% absolute.

Assignment 6: Implement Distributed Apriori Algorithm in PySpark.

a) Implement Distributed Apriori Algorithm for frequent pattern mining in PySpark to find frequent patterns that fulfill given min_support criteria.

b) Write PySpark code to generate RULES using the Apriori algorithm. Your algorithm will take the output of the part a that is frequent patterns as input and generate rules with given confidence and lift.

Assignment 7: Run your algorithm on the given dataset.

a) **Consider the dataset given with the assignment. Preprocess and understand the dataset.** After preprocessing the dataset, find patterns in the data using your PYSpark association rule mining algorithm implemented in Assignment6.

b) Experiment with different parameters so that you get at least 20-30 strong rules (e.g., rules with high lift and confidence which at the same time have relatively good support).

c) Select the best 10 most "interesting" rules and for each specify the following:

- an explanation of the pattern and why you believe it is interesting and how can it be helpful.
- any recommendations based on the discovered rule that might help the user.

Note: The top 5 most interesting rules are most likely not the top 5 in the result set of the Apriori algorithm. They are rules that, in addition to having high support, lift, and confidence, also gives some non-trivial, useful information based on the underlying business objectives.

Data set Description

Attributes	Type Of data	Details
age	Continuous data converted to Categorical	Age of the people for 17 to 85 is converted to Young, Middle, and old aged
work class	Categorical	Private,Goverment,Self-Employed,Without-pay
education	Categorical	Incomplete-Education,HS-Graduates, Associates,Bachelors,Masters,Doctorate
education-num	Continuous	Number of Years studied
marital-status	Categorical	Never-married,Married,NO-Spouse,Divorced where No-Spouse indicates widow, separated
occupation	Categorical	Technical-Working-Class, Lower-Working-Class, Other-Service,Exec-managerial,Armed-Forces
relationship	Categorical	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race	Categorical	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex	Categorical	Female, Male.
hours-per-week	Continuous	Number of Hours Worked Per Week
native-country	Categorical	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.