Big Data Mining: HW#1

By J. H. Wang

Oct. 4, 2019

Programming Exercise: the First Analysis Program

- Goal: Getting familiar with your big data mining environment and writing your first analysis program
  - MapReduce on multi-node Spark (for CS students)
  - or Python in Jupyter Notebook

- Input: Numeric data (to be detailed later)

- Output: Results of simple statistics (to be detailed later)

Tasks and Data

- Tasks
  - Performing simple statistics on numeric data (as detailed in the following slide s)
- Data: an open dataset from UCI Machine Learning Repository
- You have to submit the generated output
- You also have to output the efficiency (running time) of each task

Input Data

- Data:
  - [**Individual household electric power consumption dataset**] from UCI Machine Le arning Repository
    - About 2 million instances, 20MB (compressed) in size
  - Available at: https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption

- Format:
  - One text file consisting of lines of records
  - Each record contains 9 attributes separated by semicolons:
    Date, time, global_active_power, global_reactive_power, voltage, global_intensity, sub_metering_1, sub_metering_2, sub_metering_3

Detailed Information about Data Attributes

- 1.date: Date in format dd/mm/yyyy
- 2.time: time in format hh:mm:ss
- 3.**global_active_power**: household global minute-averaged active power (in kilowatt)
- 4.**global_reactive_power**: household global minute-averaged reactive power (in kilowatt)
- 5.**voltage**: minute-averaged voltage (in volt)
- 6.**global_intensity**: household global minute-averaged current intensity (in ampere)
- 7.sub_metering_1: energy sub-metering No. 1 (in watt-hour of active energy)
  - It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered)
- 8.sub_metering_2: energy sub-metering No. 2 (in watt-hour of active energy)
  - It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.
- 9.sub_metering_3: energy sub-metering No. 3 (in watt-hour of active energy)
  - It corresponds to an electric water-heater and an air-conditioner.

Tasks in this Homework

- 3 subtasks:
  - (**30pt**) (1) Output the minimum, maximum, and count of the columns: ‘global active power’, ‘global reactive power’, ‘voltage’, and ‘global intensity’
  - (**30pt**) (2) Output the mean and standard deviation of these columns
  - (**40pt**) (3) Perform min-max normalization on the columns to generate normal ized output

Output Format

- (1) 3 values: min, max, count
- (2) 2 values: mean, standard deviation
- (3) 1 file:
  - Each line: <normalized global active power>, <normalized global reactive pow er>, <normalized voltage>, and <normalized global intensity>

Implementation Issues

- Missing values
- Conversion of data types

References

- UCI ML repository:
  - Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Note on Programming Exercises

- Programming exercises can be done as a team
  - at most two persons per team

- Programming language
  - Java on Hadoop (for CS students)
  - Java, Scala, Python, or R on Spark (for CS students)
  - Or Python in Jupyter Notebook

Homework Submission

- For implementation projects, please submit a compressed file containi ng:
  - Your environment setup
    - How many PCs, what spec, network setup, …
  - Your source codes
  - The generated output
  - Documentation on how to compile, install, or configure the environment, and also the detailed responsibility of each member

- Due: 2 weeks (Oct. 18, 2019)

Homework Submission Site

- Programs or projects in electronic files must be submitted directly to the TA online at Open Cyber Classrooms
  - http://mslin.ee.ntut.edu.tw
- Please follow the instructions before your first login
  - **Account**: Use your *student ID* as the account and password at your first login. Please change the password *as soon as possible* for better security.
  - *Note*: Even if you already have accounts for other courses, you are still *required* to do it at your first login for this course.
  - **Filename**: Compress your source code and related files into one compressed file. Please nam e it according to your ID and each homework. For example, [id]_HW1.zip, [id]_Quiz.tar.gz.
- If you cannot successfully submit your work, please contact with the TA or the instr uctor

Evaluation of Results

- In completion of each of the tasks, you get part of the scores
  - Correctness of Output
  - Efficiency
- Please specify the environment setup of your (virtual) machines
- You might need to demo if your program was unable to run

Questions or Comments?