Lab 03: Spark

Nguyen Bao Long

04/04/2023

Abstract

In this assignment, you will learn about Spark installation, running some Machine learning examples in Spark environment on a single computer.

1 General information

1.1 Grading

- There are 2 requirements. The first one takes 60% and the second takes 40%.

1.2 Submission guidelines

- Construct the submitted folder as follows:

```
team/ (e.g. MeowMeow)
|-- src/
|   |-- notebook.ipynb
|   `-- word_count/
|-- doc/
|   `-- word_count.pdf
`-- output/
    `-- word_count/
```

  - `./src` is the folder for your source code.
  - `./doc` is the folder for your report.

- Compress `./team` folder and submit on Moodle.

2 Requirement 1

- Access this link and do all the exercises.
- Follow these instructions:
  - Read the requirements of the problem and think about a solution.

- Try to solve the problem by yourself first.
- If you can solve the problem, implement your solution using `MLlib` in `PySpark`. Include comments in your code to explain your thought process and any assumptions you made.
- You will implement 3 classification algorithms (`Decision Tree`, `Naive Bayes`, and `Random Forest`) on `Jupyter Notebook` (`./src/notebook.ipynb`). **`Restart Kernel and Run all cells` before submitting on Moodle.**.
- In `Word_count` problem, your code will be put in a separated folder named `./src/word_count`. The output will be stored in `./output/word_count`. You also need to create a report (`./doc/word_count.pdf`) on this problem.
- If you are stuck or need help, reference external sources to get inspiration and guidance. However, you should always correctly cite your sources and not copy code or solutions without permission.

3 Requirement 2

- Find 2 more datasets.
- Remember to attach the link to your chosen datasets as well as describe them in detail.
- Implement 2 more machine learning algorithms using `MLlib` on these datasets in the same notebook file (`notebook.ipynb`) with Requirement 1.

4 References

- Le Ngoc Thanh. Lab 3 – Spark. HCMUS, 2023.
- CS5590