criteria

Harito ID

2025-09-25

Evaluation Criteria

Code runs correctly and without errors (50%)

- ☐ Read the C4 dataset into a Spark DataFrame.
- ☐ Implement a Spark ML Pipeline.
- ☐ Use RegexTokenizer or Tokenizer for tokenization.
- ☐ Use StopWordsRemover to remove stop words.
- ☐ Use HashingTF and IDF for vectorization.
- ☐ Fit the pipeline and transform the data.
- ☐ Save the results to a file.
- ☐ Log the process.

Eg: 0 / 8

Report Writing (50%)

- ☐ Clearly state the implementation steps.
- ☐ How to run the code and log the results.
- ☐ Explain the obtained results.
- ☐ Clearly state the difficulties encountered and how to solve them.
- ☐ If referencing external sources, clearly state the references.
- ☐ If using pre-trained models, clearly state which model, from where, and the prompt.

Eg: 0 / 6

Total Score

The score is calculated as follows:

$$Score = \left(0.5 \times \frac{\text{Code}}{\text{Total Code}} + 0.5 \times \frac{\text{Report}}{\text{Total Report}}\right) \times 10$$

Where: - `Code` is the number of completed code tasks. - `Total Code` is the total number of code tasks. - `Report` is the number of completed report tasks. - `Total Report` is the total number of report tasks.

For example, if you complete 3 code tasks and 2 report tasks:

$$Score = \left(0.5 \times \frac{3}{8} + 0.5 \times \frac{2}{6}\right) \times 10 \approx 3.5$$

The score is rounded to the nearest 0.25.

Notes:

- The above criteria are mandatory.
- You can change the weight of the criteria, but the total weight must be 1.
- You can submit in groups, but you must specify the group members.
- Late submissions will be penalized.
- You must cite any external sources, pre-trained models, libraries, or tools used.