COMP 5434 Big Data Computing

Coding Assignment 1

Lecturer: LI Zecheng

Assignment - MapReduce Word Count

Given a dataset of the Harry Potter series *hp.txt*, your task is to write a MapReduce-style program to count the number of words in it. Typically, a MapReduce program consists of three phases:

1. In the **map** phase, data from the input text file is tokenized into words to form key value pairs. The key is the word from the input file and the value is '1'. For instance, if you consider the sentence "the best experts were at a loss to explain", the map phase will split the string into individual words. The whole sentence is split into 9 tokens (one for each word) with a value of 1, as shown below:
   ('the', 1)
   ('best', 1)
   …
   ('explain', 1)

2. In the **shuffle** phase, key-value pairs generated in the map phase are taken as input and then are sorted alphabetically. The output of the shuffle phase will look like this:
   ('best', 1)
   …
   ('the', 1)
   ('the', 1)

3. In the **reduce** phase, the outputs of the shuffle phase are taken as input. The values for the same keys are added up to calculate the number of occurrences of a particular word. The output of the reduce phase look like this:
   ('harry', 8032)
   ('potter', 7211)
   …
   ('voldemort', 1202)

You should count the occurrences of all words in the given dataset, sort them by the number of occurrences, and output the top 20 frequently occurring words.

Implementation Instructions:

You can use one of the following methods to complete this assignment.

1. You could simply write three functions, map(), shuffle(), and reduce() to complete it.
2. You could use the Python built-in function map() and library functools's function reduce()to complete it.
   For map(), you could refer: https://docs.python.org/3/library/functions.html#map
   For reduce(), you could refer: https://docs.python.org/3/library/functools.html
3. You could use our provided MapReduce class *mapreduce.py* to write a program. You could inherit it to create a new class that is specialized for the word count job.
4. You could use the concurrency libraries provided by Python, such as threading, multiprocessing, and asyncio, to complete it.

Criteria for Evaluation:

1. **Correctness of Results (30 points)**
   The MapReduce program accurately counts the number of words in the input data. The output of top 20 frequently occurring words matches the expected result.
2. **Map Function (30 points)**
   The map function correctly extracts words from the input file, and emits intermediate key-value pairs for each word.
3. **Reduce Function (30 points)**
   The reduce function correctly aggregates the word counts for each word.
4. **Shuffle (10 points)**
   The shuffle phase is implemented correctly and efficiently.

Grading Standards:

- 90-100 points: The program meets all criteria for evaluation with exceptional quality and demonstrates an advanced understanding of MapReduce.
- 80-89 points: The program meets most criteria for evaluation with good quality and demonstrates a solid understanding of MapReduce.
- 70-79 points: The program meets some criteria for evaluation with acceptable quality and demonstrates a basic understanding of MapReduce.
- 60-69 points: The program meets few criteria for evaluation with poor quality and demonstrates a limited understanding of MapReduce.
- 0-59 points: The program fails to meet the criteria for evaluation.

The assignment accounts for 15% of the total grade, so the final score will be multiplied by the appropriate factor into the total grade.