**IDS 561 Homework 1**

**Due: 02/16/2022 Wednesday 6:00pm (before class)**

In this homework you will mimic the process of the MapReduce task. Specifically, you will write your own map and reduce functions to mimic the process of mapper and reducer.

The task is to find the maximum temperature for each year. This program, see slide 26 in Lecture 2, is a typical program for beginners to learn the MapReduce programming framework.

While doing this homework you will learn:

1) how to prepare the data.

2) how to write map and reduce functions.

3) get a better understanding of how mapper and reducer work.

**Dataset**

The input of this homework is a text document (1000 lines) which includes 2000 pairs of Year/month and average temperature of that month. It is raw data. You need to do some data cleaning work to prepare it for the next step.

**Task**

You are supposed to build several functions to mimic each step of MapReduce. They are:

|  |  |  |  |
| --- | --- | --- | --- |
| Function | Description | Input of this function | Output of this function |
| Data cleaning function | Some data cleaning jobs, such as removing punctuations and special symbols, converting string to integer (you need to do that to calculate the max.), etc. | Raw text data | Clean date, temperature pairs in proper data structure. |
| Data split function | Split the dataset into two parts:  Part1 includes the first 500 lines of the  Raw data, Part2 includes the rest 500 lines. | Output of data cleaning function | Two separated subsets: Part1 and Part2. Save them as objects in Python |
| Mapper function | Two mapper functions that produce a set of key-value pairs for Part1 and Part2 subsets respectively. | Output of data split function | Key-value pairs of Part1 and Part2. |
| Sort function | Sort by key of Part1 and Part2 together,  with an ascending sort order | Output of mapper function | Sorted Key-value pairs for the whole dataset |
| Partition function | All the months in year 2010 to 2015 are sent to Reducer1,  and the others (2016 to 2020) are sent to Reducer2. | Output of sort function | Two ascending  ordered partitions. |
| Reducer function | Collect all values belonging to the key and find the maximum temperature for the two ordered partitions. | Output of partition function | Maximum temperature of the ordered partitions. |
| Main function | Wrap all the steps together and combine the output of the two partitions together. | Output of reducer function | Final result of max temperature. |

The figure below shows the basic workflow of this word count task.

Diagram

Description automatically generated

Note:

1. Using multi-thread is highly encouraged but not required. Here are some tutorials of Python multithreading:

<https://www.toptal.com/python/beginners-guide-to-concurrency-and-parallelism-in-python>

<https://realpython.com/intro-to-python-threading/>

<https://www.geeksforgeeks.org/multithreading-python-set-1/>

1. Same code for the two mapper functions or the two reducer functions is ok.
2. You can use any data types provided in Python, e.g., list, dictionary and so on.

**What to submit (individual submission)**

You need to submit two files: a Python file and a CSV file.

**Python file**: Your python code. Please add comments to make it readable.

**CSV file**: Your final word count output. The format should look like this:

|  |  |
| --- | --- |
| **year** | **Max temp** |
| 2010 | 98 |
| 2012 | 105 |
| … | … |