IDS 561 Homework 1

Due: 02/23/2021 Tuesday 3:30pm (before class)

In this homework you will mimic the process of MapReduce task. Specifically, you will write your own map and reduce functions (without distributing to several machines) to mimic the process of mapper and reducer.

The task is to count the number of occurrences of each word in a text file. This program, known as Word Count, is the equivalent of the standard "Hello, world!" program you typically write when you learn a new programming language.

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 (around 13,000 lines) which includes several paragraphs. It is a raw data. You need to do some data cleaning works to prepare it for next step.

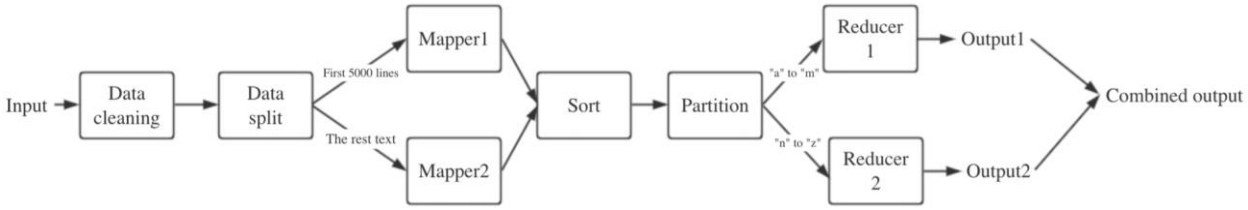
Task

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 numbers, punctuations and special symbols, uppercase to lower case.	Raw text data	Clean text data
Data split function	Split the dataset into two parts: Part1 includes the first 5000 lines of the text file, Part2 includes the rest text.	Output of data cleaning function	Two separated subsets: Part1 and Part2.
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 tokens (i.e., words) starting with letter “a” to “m” are sent to Reducer1, and the others (“n” to “z”) are sent to Reducer2.	Output of sort function	Two ascending ordered partitions.
Reducer function	Collect all values belonging to the key and count the frequency of words for the two ordered partitions.	Output of partition function	Word frequency 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 word counting.

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



Note:

① Using multi-thread is highly encouraged. 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/
② Same code for the two mapper functions or the two reducer functions is ok.
③ You can use any data types provided in Python, e.g., list, dictionary and so on.

What to submit (one submission per group)

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:

Word	Frequency
apple	123
banana	45
…	…