Assignment

This coursework requires you to write four MapReduce programs. These programs should be written using Python 3 and the Python mrjob library. Each solution should distribute computation across multiple map and/or reducer tasks.

Part 1
Given a CSV file where each line contains a set of numbers, write a MapReduce program which determines the maximum of all numbers in the file. For example, consider the following sample CSV file:
2,2,3
4,3

Given this CSV file, the maximum is 4.

Entitle the python program in question part1.py. That is, entering the following command at the terminal should result in your MapReduce program being applied to fileName.csv
pipenv run python part1.py fileName.csv

Part 2
Write a mapReduce program which takes as input a CSV file containing comma separated words and outputs for each word the lines that the word appears in. For example, consider the following file:
goat,chicken,horse
cat,horse
dog,cat,sheep
buffalo,dolphin,cat
sheep

The corresponding output will be the following:
"buffalo" ["buffalo,dolphin,cat"]
"cat" ["buffalo,dolphin,cat", "cat,horse", "dog,cat,sheep"]
"chicken" ["goat,chicken,horse"]
"dog" ["dog,cat,sheep"]
"dolphin" ["buffalo,dolphin,cat"]
"goat" ["goat,chicken,horse"]
"horse" ["cat,horse", "goat,chicken,horse"]
"sheep" ["dog,cat,sheep", "sheep"]

Entitle the python program in question part2.py. That is, entering the following command at the terminal should result in your MapReduce program being applied to fileName.csv
pipenv run python part2.py fileName.csv

Part 3
Given a file containing words separated by spaces, write a MapReduce program which counts the number of times each 4 word sequence appears in the file.

For example, consider the following file:
one two three four seven one two three four
three four seven one
seven one two three

The number of times each 4 word sequence appears in this file is:
"three four seven one" 2
"four seven one two" 1
"one two three four" 2
"seven one two three" 2
"two three four seven" 1

Entitle the python program in question part3.py. That is, entering the following command at the terminal should result in your MapReduce program being applied to fileName.txt
pipenv run python part3.py fileName.txt

Part 4
Uniform Resource Locator (URL) links describe the structure of the web. Consider a CSV file where each line contains two URLs which specify a single link. That is, the first and second values on each line specify the source and destination of the link in question. For example, consider the following sample CSV file:
url1,url2
url1,url3
url2,url3
url4,url5
url2,url4

Given such a CSV file, write a MapReduce program which finds all paths of length two in the corresponding URL links. That is, it finds the triples of URLs (u, v, w) such that there is a link from u to v and a link from v to w.

For example, the sample CSV file above contains the following paths of length two:
url2, url4, url5
url1, url2, url3
url1, url2, url4

Entitle the python program in question part4.py. That is, entering the following command at the terminal should result in your MapReduce program being applied to fileName.csv
pipenv run python part4.py fileName.csv