SI 618 Fall 2022 Lab 5 – MrJob

This lab is to familiarize you with the process of writing MapReduce code and running it locally on your laptop.

You are given a dataset containing the categorical association of products on the Product Hunt site. The dataset consists of 10 files in the tsv (tab-separated values) format. Each row of a file contains information about a product in the following 3 fields

1. release_date – the day a product was released
2. category_tags – the categories with which the product can be associated
3. upvotes – the number of upvotes received by the product

Credit: The dataset is a formatted version of one of the original datasets (76,822 Product Hunt products) used in the 'Gamer and the Nihilist' report.

Task:

In this lab, you will have to use mrjob to compute the average of the number of upvotes received by a product in each category on each day. Please round down the average to two decimal points.

The desired output for this lab can be found in the preferred_output.tsv file. Note that the order of lines in your output file may be different from the desired output, and that is OK.

To get started, download the 'si618_lab5_starter_code.py' file, and rename it as '<uniqname>_si618_lab5.py'.

Part 0. Installing mrjob

If you don't already have the mrjob Python 3 module installed, you should install it by running

```
$ pip install mrjob
```

or

```
$ pip3 install mrjob
```

Part 1. MapReduce

Add code to <uniqname>_si618 _lab5.py where specified. Then, run this script locally (on your computer) on the input files in the bills directory.

To run the code, you should enter the following:

```
$ python <uniqname>_si618_lab5.py ./data -o si618_lab5_output
```

or

```
$ python3 <uniqname>_si618_lab5.py ./data -o si618_lab5_output
```

If your script runs successfully, the script will create a directory called `si618_lab5_output` containing files with names like 'part-000000', 'part-00001', etc. (the exact number of files depends on the number of cores used to run the mrjob code). If you formatted the output correctly, each line in the output files should have the format <bill_type><tab><word><tab><frequency>.

Now, to concatenate those into a single file, run

```
$ cat si618_lab5_output/part* > <uniqname>_si618_lab5_output.tsv
```

As an additional note, you could produce this text output in a single line when running your script by adding > `si618_lab5_output_youruniquename.txt` to the command used to run the script above:

```
python <uniqname>_si618_lab5.py ./data -o si618_lab5_output > <uniqname>_si618_lab5_output.tsv
```

This would redirect the output from the terminal to the text file specified after the '>'.

What to submit:

Submit two files:

- Python source code file <uniqname>_si618_lab5.py
- Merged output file <uniqname>_si618_lab5_output.tsv