SI 618 Fall 2020 Lab 4 – 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 of bills introduced at the US congress from the 80th congress up to the 116th (partial) in the zip file. The dataset consists of 10 files in the tsv (tab-separated values) format. Each row of a file contains the following 9 fields

1. Id – unique identifier of the bill
2. Type of bill {“HR”: house bill, “s”: senate bill, “sres”: senate resolution, “hcon”: house concurrent resolution, “scon”: senate concurrent resolution, “hjres”: house joint resolution, “sjres”: senate joint resolution}
3. Title of the resolution
4. Which chamber introduced the bill (0 – House, 1 – Senate)
5. The congress in which the bill was introduced (80 to 116)
6. The year the bill was introduced
7. Did the bill pass in the House? (Boolean)
8. Did the bill pass in the Senate? (Boolean)
9. Was the bill passed into law? (Boolean)

Credit: This dataset is a formatted version derived from the source datasets at http://www.congressionalbills.org/ (Adler and Wilkerson)

In the lab, you need to compute the frequency of words that are at least 4 letters long and at most 15 letters long (numbers and punctuation must be excluded) in the titles of congressional bills **for each bill type**. Consider the following example title.

“To extend the coverage of Federal old age and survivors insurance to self employed individuals”

Then our regular expression should capture:

extend 1

coverage 1

Federal 1

survivors 1

insurance 1

self 1

employed 1

individuals 1

To get started, download the 'si618_lab4.py’ file, and rename si618_lab4.py' as ‘<uniqname>_si618_lab4.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 _lab4.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_lab4.py ./bills -o si618_lab4_output
```

or

```
$ python3 <uniqname>_si618_lab4.py ./books -o si618_lab4_output
```

If your script runs successfully, the script will create a directory called `si618_lab4_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_lab4_output/part* > <uniqname>_si618_lab4_output.tsv
```

Note that the order of lines in your word output file may be different from the desired output (see preferred output.tsv), and that is OK.

As an additional note, you could produce this text output in a single line when running your script by adding > `si618_lab4_output_youruniquename.txt` to the command used to run the script above:

```
python <uniqname>_si618_lab4.py ./bills -o si618_lab4_output >
<uniqname>_si618_lab4_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_lab4.py
- Merged output file <uniqname>_si618_lab4_output.tsv