COMP9313 23T2 Project 1 (12 marks)

Problem statement:

Detecting popular and trending topics from news articles is important for public opinion monitoring. In this project, your task is to perform text data analysis over a dataset of Australian news from ABC (Australian Broadcasting Corporation) using **MRJob**. The problem is to compute the weights of each term regarding each year in the news articles dataset and find out the most important terms in each year whose weights are larger than a given threshold.

Input files:

The dataset you are going to use contains data on news headlines published over several years. In this text file, each line is a headline of a news article, in the format of "date,term1 term2 ... ... ". The date and text are separated by a comma, and the terms are separated by a space character. A sample file is like below (note that the stop words like “to”, “the”, and “in” have already been removed from the dataset):

```
20191124,woman stabbed adelaide shopping centre
20191204,economy continue teetering edge recession
20200401,coronanomics learnt coronavirus economy
20200401,coronavirus home test kits selling chinese community
20201015,coronavirus pacific economy foriegn aid china
20201016,china builds pig apartment blocks guard swine flu
20211216,economy starts bounce unemployment
20211224,online shopping rise due coronavirus
20211229,china close encounters elon musks
```

This small sample file can be downloaded at:
https://webcms3.cse.unsw.edu.au/COMP9313/23T2/resources/88346

Term weights computation:

To compute the weight for a term regarding a year, please use the TF/IDF model. Specifically, the TF and IDF can be computed as:

- TF(term t, year y) = the frequency of t in y
- IDF(term t, dataset D) = $\log_{10}$ (the number of years in D/the number of years having t)

Finally, the term weight of term t regarding the year y is computed as:

- Weight(term t, year y, dataset D) = TF(term t, year y)* IDF(term t, dataset D)

Please import math and use math.log10() to compute the term weights.

Code format:

Please name your Python file "project1.py" and compress it in a package named "zID_proj1.zip" (e.g. z5123456_proj1.zip). The code template can be downloaded at: https://webcms3.cse.unsw.edu.au/COMP9313/23T2/resources/88347.

Command of running your code:

To reduce the difficulty of the project, you are allowed to pass the total number of years to your job. We will also use more than 1 reducer to test your code. Assuming there are 20 years, β is set to 0.5, and we use 2 reducers, we will use the command like below to run your code:

$ python3 project1.py -r hadoop hdfs_input -o hdfs_output --jobconf myjob.settings.years=20 --jobconf myjob.settings.beta=0.5 --jobconf mapreduce.job.reduces=2

- **hdfs_input:** input file in HDFS, e.g., "hdfs://localhost:9000/user/comp9313/input"
- **hdfs_output:** output folder in HDFS, e.g., "hdfs://localhost:9000/user/comp9313/output"
- You can access the total number of years and the value of β in your program like "**N = jobconf_from_env('myjob.settings.years')**", (use "**from mrjob.compat import jobconf_from_env**" in your code).

Output format:

You need to output all terms whose term weights regarding each year are larger than the given threshold value β (note that one term could appear in different years). The format of each line is: "Term\tYear,Weight". You need to sort the results first by the terms in **alphabetical** order and then by the years in **descending** order.

For example, given the above data set and β=0.4, the output can be checked at (there is no need to remove the quotation marks which are generated by MRJob): https://webcms3.cse.unsw.edu.au/COMP9313/23T2/resources/88345.

Submission:

Deadline: Monday 26th June 11:59:59 PM

If you need an extension, please apply for a special consideration via "myUNSW" first. You need to submit through Moodle. If you submit your assignment more than once, the last submission will replace the previous one. To prove successful submission, please take a screenshot as the assignment submission instructions show and keep it by yourself. If you have any problems with submissions, please email siqing.li@unsw.edu.au.

Late submission penalty

5% reduction of your marks for up to 5 days, submissions delayed for over 5 days will be rejected.

Marking Criteria:

Your source code will be inspected and marked based on readability and ease of understanding. Your source code's documentation (comments on the codes) is also important. Below is an indicative marking scheme:

Submission can be compiled on Hadoop: 4
Submission can obtain correct results on a single reducer: 1
Submission can obtain correct results on multiple reducers: 2
Submission uses the combiner or in-mapper combining: 1
Submission uses the order inversion: 1
Submission uses the secondary sort: 1
Submission uses only a single MRStep: 1
Code format and structure, Readability, and Documentation: 1

Cautions:

- Source code that can only be compiled in the local environment is not acceptable.
- Your source code must be able to run with the provided command on Hadoop. Otherwise, it will be treated as compiling unsuccessfully.
- Please provide sufficient comments for your source code so that we can easily mark your readability and documentation. Detailed comments are not mandatory, but it is suggested that your comments could clearly describe the main logic of your source code. For example, your comments should explicitly guide the tutors to find out how you implement the combiner, sorting, etc.
- In the output file, do not include any additional "space" between the fields. In each line, only **"\t"**, **","**, ";" are the valid separators.
- Using multiple MRSteps is allowed but you will lose 1 mark.

Plagiarism:

The work you submit must be your own work. Submission of work partially or completely derived from any other person or jointly written with any other person is not permitted. The penalties for such an offense may include negative marks,

automatic failure of the course, and possibly other academic discipline. Assignment submissions will be examined manually.

Relevant scholarship authorities will be informed if students holding scholarships are involved in an incident of plagiarism or other misconduct.

Do not provide or show your assignment work to any other person - apart from the teaching staff of this subject. If you knowingly provide or show your assignment work to another person for any reason, and work derived from it is submitted you may be penalized, even if the work was submitted without your knowledge or consent.