INF-617 - Big Data - 2018 - 1st semester
Prof: Edson Borin
TA: Antonio Carlos Guimarães Junior

Suppose you have two files: **measurements.txt** and **sensor-location.txt**.

sensor-location.txt is a text file that contains in each line a number that identifies a sensor and the location of the sensor separated by the TAB character (\t). The following example illustrates the contents of the file:

1	Campinas, SP
2	Florianopolis, SC
3	Rio de Janeiro, RJ
4	Campo Grande, MS
5	Paranavai, PR
6	Amparo, SP
7	Ubatuba, SP
8	Guaruja, SP
9	Paranavai, PR

Notice that sensors 5 and 9 are located in Paranavaí, PR.

The **measurements.txt** file contains readings of temperature/humidity/light sensors. Each line contains a reading, which has the following format:

- Characters 1-5: sensor id;
- Characters 6-9: year (4 digits);
- Characters 10-11: month (2 digits: 01 - 12);
- Characters 12-13: day (2 digits);
- Characters 14-15: hour (2 digits);
- Characters 16-17: minute (2 digits);
- Characters 18-20: temperature (3 digits). Celsius x 10;
- Characters 21-23: humidity (3 digits);
- Characters 24-26: lux (3 digits).

The following example illustrates the contents of the file:

```
000012014012718401670572l7
000012014010905151710424O2
000052014011809024930224l5
000092014010803283310450l1
```

The first entry on the file indicates that sensor 1 performed on January the 27th of 2014 at 18:40 the following readings: temperature=16.7, humidity=57, and lux=217.

Your task is to implement a set of MapReduce programs that:

1. Reports for each of the 12 months the location that had the highest temperature on that month.
    - Each line shall contain (month, location), where month is the name of the month (January, February, March, April, May, June, July, August, September, October, November, December) and location is the name of the location (e.g. Paranavai)
    - The result shall be store in a file named “item1.out”
2. Reports for each location the maximum and minimum temperature ever recorded.
    - Each line shall contain (location, min. temp, max. temp).
    - The result shall be store in a file named “item2.out”
3. Reports for each location the average temperature and standard deviation on January and on July.
    - Each line shall contain (location, jan. avg, jan. stdev, jul. avg, jul. stdev), where jan. avg is the average and jan. stdev is the standard deviation of all measurements for that location on January.
    - The result shall be store in a file named “item3.out”

You must submit all the MapReduce programs and a single script named “run” (without any extension, using a shebang to indicate the interpreter). This script shall execute all the necessary steps to produce the output file for each of the above-mentioned items. However, all the data manipulation must be handled by the MapReduce programs only. The use of the Hadoop Streaming module is advised, but not mandatory.

Evaluation: I'll first download all your submitted files to a single folder, which will already contain the input files. Then, I'll transfer the folder to the Bitnami Virtual Machine. I'll give execution permission for your “run” script and run it for at most 5 minutes. Finally, I'll inspect the results on files item1.out, item2.out and item3.out.