Distributed Data Infrastructures, Fall 2020, Project 2

Assignment

In this project, you are supposed to use GraphLab to analyze data sets. GraphLab is a high performance, open-source, big data framework which we have discussed in the course. We provide you with a data set and you should run GraphLab either on AWS or locally to analyze the data set. The dataset was released by Telecom Italia for their Open Big Data Challenge in 2014. It contains telecommunication records, weather, air quality, electricity consumption for city of Milan and province of Trentino in Italy in November and December 2013. You can find detailed description of the dataset from the paper: https://www.nature.com/articles/sdata201555.

Requirements

You need write a program that uses GraphLab to provide answers to the following questions.

First three questions are worth 2 points total.

1. Find the most congested communication period of the day in Milan and Trentino.
2. List top 5 Italian provinces which are most called by residents of Milan and Trentino on average.
3. List top 5 languages tweeted by distinct users in Milan. How popular is Finnish as a tweeting language in Milan?

The following questions are worth 3 points total.

4. Compare call and internet activity between 24th, 25th and 26th December to 26th, 27th, 28th November for Milan. Plot the distribution.
5. Find correlation between user communication activity and different weather conditions (e.g. rain, snow etc.) in Milan and Trentino.

The final questions are worth 2 points each.

6. Plot the heatmap of user telecommunication activity for both Milan and Trentino. Do you observe any shift in communication pattern of users between day and night? (A typical day time is between 8AM to 8PM)
7. Investigate and plot the correlation between air quality and weather (temperature, sunshine, precipitation, etc.).

Documentation

In the documentation, you should explain how your code solves the problems and how it uses GraphLab. You also need to provide the answers to the above questions.

Grading

Grading is based on the correctness of the program and the answers, quality of the program code, and associated documentation.

Guidelines

The assignment is individual work. You can of course discuss any problems you encounter with other students, but sharing code is not allowed and if found, will be considered as plagiarism.

Deliverables

Program source code with documentation. You can return the code either as a python script or an iPython notebook (.ipynb). The document should explain how you have solved the problems and provide answers to the questions from Requirements section.

Restrictions

For plotting questions, you cannot use the plot function in GraphLab canvas GUI. You must use matplotlib to plot your data.

Timeline

The assignment is due on 13.12.2020 at 23:00. No extensions will be given.

Return

Store all the files in a directory that has same name as your username. Zip this directory, name the zip-file "username_DDI20_EX2.zip", and return the zip-file via Moodle. Please indicate clearly your name and student ID in every source code file.

Set Up

Please see links provided below on how to get started with AWS or do a local install. Note that GraphLab requires Python 2.7 to work!

Data sets and sample code

You can find the data sets on Ukko2 at /wrk/group/grp-ddi-2020/project2/graphlab-data/. The datasets for Milan and Trentino are in different folders. For ease-of-use, the dataset has been unzipped such that CSV files for individual days/workload can be downloaded individually. However, the ZIP version of dataset is also available in "zip-archive" folder of each city.

More Information

You may also find the following links useful:
GraphLab user guide: https://turi.com/learn/userguide/
GraphLab documentation: https://turi.com/products/create/docs/
Local Installation instructions for GraphLab Create:
https://turi.com/download/install-graphlab-create.html
AWS EC2 Installation instructions for GraphLab Create:
https://turi.com/download/install-graphlab-create-aws-coursera.html