Big Data Course Project 2018/2019 **Deadline: Feb 10th at the latest**
Exam Period: Jan/Fab 2019

The NY taxi commission has provided the data of their taxi for free in order to understand better how their service works.
http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
They contain trips that were done by different kinds of taxi companies in the last few years.
From where they went, to where, and how much they paid, when, etc.

Our goal is to understand the different kinds of traffic that exist between the different zones.
We would like to create a graph

- Create a Spark program that reads the data, identifies categories of the trips (hint: Clustering?) and creates a profile (meaning a set of characteristics) for each category (hint: statistics). Then create a graph in which every node represents an area and an edge between two nodes means a set of trips that have been done between them. The edges will be of different type, one type for every category (hint: a solution can be use a graph db, but can be done even with Spark and DataFrames)

- Identify the top-5 drivers, i.e., drivers that have done the most trips, the longest distance, have made the most money, etc.
- Identify the best locations, i.e., the locations with the most pickups, the most drop-offs,
- Quantify the total pick-up and drop-off by time of the day (per hour) and per location, as well as overall.
- Illustrate how the fare changes over time (overall fares collected throughout the day, per hour, location, and also overall).
- Extra: Any other kind of analysis you can think about is welcome to be added in the project.

Analyze the results, meaning comment on the observations that you see.

You can save the results of your analysis as a CSV table. For your report, you can report them in the best possible way. (e.g. a latex table or a graph)
Optional but worth the try: Create an interface that illustrates your results in a better (graphical) way. This can be done through a JavaScript page, D3.js, Excel, MATLAB, or anything else you like, or you know.
Create a report in which you describe in full details the way that you have implemented the above and what you have used, and what results you have obtained and the analysis (comments) that you have performed.
For the delivery, you need to create :

- One directory called src that contains your code
- One directory called report that contains the sources of your report (latex or Word)
- One directory called data that contains the data you used
- The pdf file of your report.

Place them under one directory in google drive and share it with velgias@unitn.it. Once you do so, send an email to velgias@unitn.it with the link to that directory.
The name of your pdf report and of the name of your directory should be BD_XXXX_YYYY where the XXXX and YYYY are the matricola numbers of the group members.

The format of your report (template) cannot be arbitrary. The template will be provided on the google classroom. Make sure you follow that. No different formats are accepted.

Note for people also taking the data mining project: The project is different for those students also taking the data mining project. If you are taking the data mining project and you are not already aware of the different project, talk to the professor.

The project is for groups of up to 3 persons