Term Project

MET CS 777 - Big Data Analytics
Term Project
(20 points)

GitHub Classroom Invitation Link
`https://classroom.github.com/a/m0ENa2iP`

1 Description

The goal of this project assignment is to give students an opportunity to build a large scale data science project using a real-world data. Students should select a public data set, think about a research project for example a clustering or classification problem or any machine learning model, and implement the model using cluster computing platforms like Apache spark, and evaluate the results.

2 Data

Select a data set from the available public data sets (you can find a list of public data sets here http://www.teymourian.de/public-data-sets-for-data-analytic-projects/).

- The data set must not be very Large
- You can reuse one of the assignment data sets.

3 Define a Project (4 Points)

Define a data science research question based on the data set. You can find a list of such public data sets at the end of this guide. You can select one of the listed data sets or search the web and pick up one of the available public data sets on the Web.

You should describe the following items

1. What is your data set about?
2. What is exactly your research question? What do you want to learn from data? What is your learning model ,e.g., a Classification, Clustering, etc?
3. What do you expect after implementation of
4. How do you want to evaluate your project? How to access the correctness of your model? How well would you expect that the model will work?

4 Implementation (6 Points)

- You need to implement your project using Apache Spark and you are allowed to use the Machine Learning Library of Spark (Spark ML, Spark Mlib).
- You need to correctly implement your training model and test the model based on separating the data set to training and testing subsets.
- You code should compile and we should be able to read your README.md file to understand how to run your project. Provide in README file clear instructions on how to run your project.
- Run your implementation (on your Laptop or on Cluster if data is large) and generate the results.

5 Create Recorded Presentation Video of your Project (10 Points)

- Create a document or a presentation (PPT or other formats) to describe your project and results
- Describe your code.
- Describe the results of your project in a professional way.
- You may want to visualization diagrams and describe the results based on some diagrams - but having diagrams is not a MUST have to get the full credit.
- Describe the model and results of your project in a way that every person in this field can read, enjoy and understand.

Present your talk on your desktop and record it using a desktop recording software. You can also use your webcam and add it to your presentation.

You can use **Kaltura Capture Recorder.** `https://onlinecampus.bu.edu/bbcswebdav/courses/00cwr_odeelements/metcs/cs_Kaltura.htm`

You have the following 3 options to create a presentation video:

- **Method 1:** Use the Kaltura Capture and BU my-media system. You can find the HowTos on our Blackboard system or here `https://www.bu.edu/tech/services/teaching/instructional-video/my-media/my-media-faqs/using-kaltura-capture/`
- **Method 2:** Alternatively, you can record your presentation using ZOOM client application, start your zoom client, start a zoom meeting (you will be the only participant), click on share your desktop, turn on camera and record your presentation, end the zoom session and you will get a MP4 video file of your presentation. Upload the file to Blackboard system
- **Method 3:** Alternatively, you can use any desktop recording software, generate a MP4 file and upload MP4 to blackboard, or upload it to some other cloud storage (like google Drive, DropBox) and share with us the URL Link of your presentation (submit the link in Blackboard).

Note 1: We recommend to use Kaltura and BU my-media system.

Note 2: Your video presentation should be between 8 minutes to maximum 20 minutes long (Min 8 mi, Max 20 min)

Note 3: You should turn on your Web Camera so that we can see your face during the presentation, and be able to identify the presenter.

Grading will be based on quality of your presentation, and correct describing of algorithms or concepts.

5.1 Turnin

Please include all of your documents in your Github repository for the term project.