PROJECT DESCRIPTION

The project is an essential part of this class. It will allow you to demonstrate your Big Data (BD) and Analytics skills and create something that you are proud of. It can also be a valuable addition to your projects portfolio that you can demonstrate to prospective employers.

Project Requirements

For the project, you have to perform analysis of one or more datasets using BD techniques. Some of the requirements of the project are:

- The datasets should be chosen from a standard repository, such as Kaggle competitions, KDD cup competitions or Stanford Large Network Dataset Collection (SNAP) or any other reliable source. If you are not sure, please consult the instructor or the TA.

- The project needs to involve significant use of Big Data and Analytics technologies. Some of the specific requirements are: [You should try to include as many of these as possible in your project]
    - Use of HDFS
    - Use of Spark libraries for model creation – MLlib, GraphX, etc
    - Use of Spark libraries for model selection, parameter tuning, and evaluation
    - Use of Spark machine learning pipelines
    - If you are doing a project on recommender systems, you need to evaluate and validate your model properly.

- Your results should be strong enough in terms of accuracy and other evaluation metrics, and this will be one of the criteria for grades.

- You should create a well formatted project **report** that should cover the following sections:
    - Introduction and problem description
    - Related work
    - Dataset description (including features, attributes, etc)
    - Pre-processing techniques
    - Your proposed solution, and methods [This section should have enough details – both theoretical, and practical]
    - Experimental results and analysis [Details are expected]
    - Conclusion
    - Contribution of team members
    - References

An excellent example of what to include in such a report can be found here:
http://www.cs.utexas.edu/~mooney/cs391L/paper-template.html

Some examples of excellent reports can be found at: (Note: You cannot select these topics)
http://cs229.stanford.edu/projects2015.html
http://cs229.stanford.edu/projects2014.html
http://cs229.stanford.edu/projects2013.html

All contents of your report must be original. You cannot copy sentences, paragraphs, figures, or anything else from outside sources. As a graduate student, you are expected to work with maturity and diligence.
Again, your report will be checked for plagiarism. Any violation will carry strong penalties, including reporting the incident to university authorities.

- Team size requirements: Project can be done in teams of 1 to 4 students. More than 4 students cannot be in a team under any circumstances. You can only form team within the same class and section. You are not allowed to work or collaborate with students from other sections of this class.

- Project selections have to be approved by the instructor.

- You have to submit your responses on the Google Form available here: https://forms.gle/aakKCKfvJL3vyS9p9

- The final project report is due at midnight Friday Aug 2. Project demos and presentations will be required in front of the TA during the last week of class i.e. the week starting July 29. You can use at most 2 free days for the project and everything will close down by Sunday Aug 4. These are strict deadlines.

Project Ideas

Below are some of the project ideas. You can choose any one of them. Note that for the data science competitions, you have multiple options. You are free to choose any active competition, but you will have to follow the requirements completely. You cannot pick and choose which requirements you will satisfy.

Below are some suggested topics.
Note: Two teams can work on the same project, **but cannot collaborate with each other.**

1. Participate in the Yelp dataset challenge and submit a good entry:

http://www.yelp.com/dataset_challenge

2. Take part in an **active** Kaggle competition that involves significant amount of Big Data and Machine Learning technologies

https://www.kaggle.com/competitions

3. Take part in a previous KDD cup challenge

http://www.kdd.org/kdd-cup

You can take part in any previous year's cup.

4. Take part in an **active** Driven Data competition.

https://www.drivendata.org/

5. Machine learning based analysis of stock market investing techniques

Ideas:

- Simulation of systematic trading techniques, such as backtesting
https://en.wikipedia.org/wiki/Technical_analysis#Systematic_trading
- Simulation and analysis of backtesting using R packages such as backtest, PerformanceAnalytics, quantmod, etc

6. Take part in a competition from KDnuggets
https://www.kdnuggets.com/competitions/

7. Take part in a competition from Innocentive
https://www.innocentive.com/

8. Take part in a competition from TunedIT
http://tunedit.org/

9. Detect tight communities (groups of people that interact frequently with each other, but not with others) in a social network.

Data can be obtained from: http://snap.stanford.edu/data/

10. Detect opinion spammers (people who post fake reviews, or too many negative reviews) in an opinions dataset

Data can be obtained from: http://snap.stanford.edu/data/#reviews

11. Take part in a TopCoder Data Science challenge:
https://www.topcoder.com/community/data-science/

12. Suggest your own project having sufficient complexity and Big Data component.

Deliverables and Deadlines

Deadline	Project Phase	Deliverable
Sunday July 14 Midnight	Project Selection Team Formation	Submit your details on Google Forms https://forms.gle/aakKCKfvJL3vyS9p9 Please check for instructor's comments and approval at: https://bit.ly/2XEHhvF
Sunday July 21 Midnight	Project Status Report	Submit a report containing following on eLearning: • Dataset details, such as number of features, instances, data distribution • Techniques you plan to use • Experimental methodology • Coding language / technique to be used • Preliminary Results (if available)
Friday August 2 Midnight	Final Report	Submit final documents on eLearning: • Detailed Final Project Report • Code • README file indicating how to run your code ** Your report and code will be checked for plagiarism **