PROJECT DESCRIPTION

The project is an essential part of this class. It will allow you to demonstrate your Big Data Analysis 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

- The project has two key components to it:
  - Understanding a recent machine learning technique and associated algorithm(s)
  - Implement and apply it to a standard dataset of sufficient complexity

- A list of acceptable techniques is mentioned later in this document. You are also free to propose any additional technique to the instructor, provided it is of sufficient complexity and you understand that you need to implement the main algorithm from scratch without using any libraries.

- Your project deliverables should consist of two major parts:

  1. A report in IEEE conference format
     https://www.ieee.org/conferences/publishing/templates.html

     It should have the following sections at the minimum:
     - Abstract
     - Introduction and background work
     - Theoretical and conceptual study of the technique/algorithm you would like to implement
     - Results and analysis. Please include results in tabular or graphical formats. Be sure to analyze your results well.
     - Conclusion and future work
     - References

The report excluding the references should be 4-6 pages long. The final file should be converted to PDF format before submission.

2. Your code, link to dataset, results, and instructions for compiling and running

Below are the requirements:

- You should build your project such that it can run on Apache Spark. You can use either Scala or PySpark for coding. In case of Scala, you should build a jar file that can run on AWS cluster. In case of PySpark, you need to provide a notebook or other code that can run on AWS.

- You are allowed to use public paths on AWS S3 or UTD web account.

- Be sure to include instructions on how to compile and run your code.

- Emphasis should be on **implementing** one or more algorithm from the technique that you have studied, and **not just use a standard library for it**. For example, if you are studying recurrent neural network (RNN), you could implement i.e. code a LSTM. You are free to use any pre-processing library that you need.

- You are free to choose a dataset of your choice from sources like Kaggle, or any other source. Do not include the dataset as part of the deliverables, instead host it on your UTD web account or AWS S3. This will allow the TA to run your code without having to search the dataset or download huge files. If you do not know how to host data on UTD account, contact the TA.

- A log file of your experiments and parameters should be maintained and submitted. Example of a log file is shown below:

Experiment Number	Parameters Chosen	Results
1	Neural Net: Number of layers = 4 Neurons = (8, 8, 4, 2) Error Function = RMSE Regularization Parameter = 0.6	Train/Test Split = 80:20 Size of dataset = 10,000 Training Accuracy = 95% Test Accuracy = 88% Training RMSE = 1.67 Test RMSE = 3.08 .....
2	.....	.....
....	.....	.....

- Below are some further administrative requirements:

  - All contents of your report must be original. You have to write the report in your own words. It is acceptable to include figures from the references, provided you state the source clearly in the caption.

  - 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.

  - Only one person per team must submit the documents. Make sure the names of all team members are prominently written on the front page of all documents.

  - If you have free days remaining, you can use at most 2 free days for the final project and code submission. Free days cannot be used for the

project status report. Free days will be counted on an individual basis. If you have no more free days left, a penalty of 10% per late day will be imposed. All submissions will be closed 48 hours after the deadline.

Project Topics

Below is the list of topics that you can choose from:

- Spark streaming for novel real time applications such as novel class detection, IoT, web traffic analysis, stock market prediction, real-time anomaly detection, political actor detection, etc. See this presentation for possible topics
  http://credit.pvamu.edu/MCBDA2017/Slides/InvitedTalkDay2n2-Khan.pdf
  Note: you cannot do anything similar to what you did in homework.

- Novel graph mining applications using GraphX. See this presentation for possible topics:
  https://wikt-daz2016.fiit.stuba.sk/wp-content/uploads/2016/11/WIKT-DaZ-2016_Vaculik_keynote.pdf
  Note: Your project should involve significant work and you cannot do anything similar to what you did in homework.

- Document summarization using NLP techniques

- Recurrent Neural Networks (RNN) for machine translation

- Recurrent Neural Networks (RNN) for time series prediction (e.g. stock market, weather, hurricane intensity data)

- Image and video captioning with deep neural networks

- Autoencoders for bioinformatics or image processing

- Scene recognition with deep neural networks

- Deep Reinforcement Learning
- Genetic sequence analysis using deep neural networks
- Reinforcement learning for game playing
- Meta-Learning
- Transfer Learning
- Adversarial Machine Learning
- Statistical Relational Learning
- Human assisted Machine Learning

Deliverables and Deadlines

Deadline	Project Phase	Deliverable
Monday March 28 Midnight	Project Status Report	One person per group must submit a one-page report containing following on eLearning: • Project Topic • Team Members • Technique/Algorithm you plan to implement • Dataset details, such as number of features, instances, data distribution • Coding language / technique to be used • Preliminary Results (if available)
Monday April 25 Midnight	Final Report	• One person per team must submit via eLearning all deliverables as described in the project requirements section. ** Your report and code will be checked for plagiarism **