Introduction to Data Mining
DATS 6103
Due: *Check Syllabus*

Final Project

The objective of the final project is to apply what you have learned in this course to a real world problem – a problem in your area of interest. You can apply any data mining algorithms that we cover in the course. Below is a list of websites that contain data sets that might be suitable for your final project.

Websites with datasets

Data Sets For BI/Analytics/Visulization Projects
https://sqlbelle.com/2015/01/16/data-sets-for-bianalyticsvisualization-projects/

Kaggle competitions and data sets
https://www.kaggle.com/

Deep learning data sets, Christos Christofidis
https://github.com/ChristosChristofidis/awesome-deep-learning

Data sets for deep learning benchmarking
http://deeplearning.net/datasets/

Project page for Stanford course on Visual Recognition
http://cs231n.stanford.edu/project.html

Project page for Stanford course on Natural Language Processing
http://cs224d.stanford.edu/project.html

Open Data for Deep Learning
https://deeplearning4j.org/opendata

Top 10 Popular Publicly Available Datasets
https://analyticsindiamag.com/top-10-popular-publicly-available-datasets-deep-learning-research/

1 Group

1.1 Group Proposal

After you have selected a topic, a network, and a data set, submit a proposal of what you plan to do for the project. The proposal should be a few hundred words, and should address the following items.

- What problem did you select and why did you select it?
- What database/dataset will you use? Does it need to be cleaned?
- What data mining algorithm will you use? Will it be a standard form, or will you have to customize it?
- What softwares will you use to implement the network? Why?
- What reference materials will you use to obtain sufficient background on applying the chosen network to the specific problem that you selected?
- How will you judge the performance of your results? What metrics will you use?
- Provide a rough schedule for completing the project.

1.2 Group Presentation

You will give a 15 to 20 minute presentation of your final project.

1.3 Group Final Report

1. Introduction. An overview of the project and an outline of the report.
2. Description of the data set.
3. Description of the data mining and learning or cleaning algorithm or other algorithms that you used. Provide some background information on the development of the algorithm and include necessary equations and figures.
4. Experimental setup. Describe how you are going to use the data to clean and preprocess. Explain how you will implement the data mining technique in the chosen software and how you will judge the performance. Write a complete report with theoretical description and verify this mathematical concepts with applying it with actual data. Provide enough information about the codes tat you have written. Write your codes in sperate subroutines and call the functions if needed?. Explain each subroutine.
5. Results. Describe the results of your experiments, using figures and tables wherever possible. Include all results (including all figures and tables) in the main body of the report, not in appendices. Provide an explanation of each figure and table that you include. Your discussions in this section will be the most important part of the report.

6. Summary and conclusions. Summarize the results you obtained, explain what you have learned, and suggest improvements that could be made in the future.

7. References.

8. A separate appendix should contain documented computer listings.

1.4 Deliverables

The Group Proposal will be due by 2 weeks after the final project is posted. For Group Final Report and Group Presentation timing check the syllabus.

- Make a folder and name it Group-Proposal. Write your final group proposal in a word document and save it as a PDF file. Place the group proposal in the Group-Proposal Folder.

- Make a folder and name it Final-Group-Project-Report. Write your final group project report in a word document and save it as a PDF file. Move it to the Final-Group-Project-Report folder.

- Make a folder and name it Final-Group-Presentation. Create a powerpoint presentation for your group presentation. Save the presentation as a PDF file and move it to the Final-Group-Presentation folder.

- Make a new folder and name it Code. Save all of your codes for the final project in it. For example, if you write code for multiple parts of the project, name them properly and write a README file for it. This README file should explain what order codes need to be run in (e.g., codes to fetch data should be run first, then code to preprocess, and then modeling, etc.) and a short description of what each script does.

- Have one group member create a GitHub repository for the project and name it Final-Project-GroupX where `X` is your group number (The rest of the group members can be added in as collaborators and fork it, so you only need one repo for the project). Then, push the 4 folders that we discussed above into the repository that you created. You should have a markdown file (`README.md`) that explains the structure of the repoistory and how it works. Make it as clear as possible.

2 Individual

2.1 Individual Final Report

1. Introduction. An overview of the project and an outline of the shared work.
2. Description of your individual work. Provide some background information on the development of the algorithm and include necessary equations and figures.
3. Describe the portion of the work that you did on the project in detail. It can be figures, codes, explanation, pre-processing, training, etc.
4. Results. Describe the results of your experiments, using figures and tables wherever possible. Include all results (including all figures and tables) in the main body of the report, not in appendices. Provide an explanation of each figure and table that you include. Your discussions in this section will be the most important part of the report.
5. Summary and conclusions. Summarize the results you obtained, explain what you have learned, and suggest improvements that could be made in the future.
6. Calculate the percentage of the code that you found or copied from the internet. For example, if you used 50 lines of code from the internet and then you modified 10 of lines and added another 15 lines of your own code, the percentage will be $\frac{50-10}{50+15} \times 100$.
7. References.

2.2 Deliverables

- Write your individual report in a word document for your contributions to the final project.
- Save this word document as a **Single PDF file** and rename the PDF file as firstname-lastname-final-project. Then create a folder titled Individual-Final-Project-Report and move the PDF file of your report into it.
- Create a new folder titled Code and then put all of your codes into to it. For example, if you wrote code for a portion of the project, put the code in a script, `mywork.py`, and move it into the Code folder.
- Make a new folder and name it firstname-lastname-individual-project. Then, move the above two folders into this folder. Finally add this folder to the GitHub Repo.
- Failing to do these steps properly will result in a reduced grade.

3 Uploading

- **One member of the group have to upload the GitHub repo link (includes all the folders above) into BlackBoard. Please check that the link is working before uploading it**.