Guidelines for the Capstone Project (CKME 136) at Ryerson University

Following are the initial instructions to get you started. Note that to take this course you must have passed all the five courses in the certificate program.

1. Identify the theme on which you want to work on:
   Text Classification and Sentiment Analysis
   Classification and Regression (non-textual dataset)
   Exploratory Analysis (clustering, pattern mining, etc.)
   Recommender system
   Anomaly Detection (outlier detection)
   Data Mining and knowledge discovery
2. No private datasets are allowed after the course start date due to the delay in legal process of filing a non-disclosure agreement with the university. If you wish to use any organization's dataset, then dataset has to be publicly available on their website.
3. Here is a list of public repositories of data that you can use for your projects. You need to select one of the following dataset or any other public dataset that match one of the above themes that you selected in Step 1:
   https://www.data.gov/
   https://www.healthdata.gov/
   https://data.medicare.gov/data/hospital-compare
   http://www.dol.gov/open/data.htm
   www.toronto.ca/open
   https://www.ontario.ca/page/sharing-government-data
   https://nycopendata.socrata.com/
   http://www.gsa.gov/portal/content/181595
   http://open.canada.ca/en
   http://www.statcan.gc.ca/eng/rdc/data
   http://climate.weather.gc.ca/
   http://archive.ics.uci.edu/ml/
   http://githubarchive.org
   http://www.crowdflower.com/data-for-everyone
   http://www.kaggle.com/competitions

   MIT ICU Open Source Database:
   MIMIC: https://mimic.physionet.org/
   SQL Cookbook to extract data from MIMIC: https://github.com/MIT-LCP/mimic-cookbook

4. Here are some of the project ideas:
    a. If you are unsure of where to begin, consider working on the Titanic Competition (http://www.kaggle.com/c/titanic). There is a good amount of help available on Kaggle forums for the competition (http://www.kaggle.com/c/titanic/forums). If you have already worked on a Kaggle project in the past, then you can submit that project too.
    b. Analyze twitter data for sentiment analysis; e.g., classify tweets to medical and non-medical tweets and then identify health issues in different regions.
    c. Analyze twitter data to identify the new phrases and idioms in different regions. For this project, you need to first extract valid n-grams (http://en.wikipedia.org/wiki/N-gram) in the tweet (e.g., use bigrams or trigrams) and identify the most common phrases by a region. You can even make this project available online to general public and update the phrases in real time (if you put few advertisements there, then you might earn some profit too.
    d. Take a look at the fertilizer data: http://catalog.data.gov/dataset/fertilizer-use-and-price . Identify three different research questions and perform exploratory analysis; e.g., what will be the consumption of nutrients per crop in future years? To solve this, you can employ regression analysis.
5. In the beginning, you need to identify the theme, the dataset, and your research questions. You will need to write an abstract of 100 words or less about your project. The abstract should contain: (a) the brief context about the problem and the theme you have chosen; (b) the problem that you are solving (e.g., the research questions or the summary of research questions.); (c) the data you are using; and (d) the techniques (e.g., classification, recommender system, sentiment analysis, etc.) and the tools that you are proposing to solve it. You will submit the abstract on the Blackboard.
6. Note that instructors will not be coding for you, they will provide you technical and theoretical guidance during the project. You shall be building the product yourself.
7. You will be assigned to one of the clusters based on the theme selected and you will be supervised in small groups. It is also possible that we will ask you to use another dataset as your peers in the group if the required task is not feasible to achieve.
8. A data scientist also possesses superior communication skills. You will also need to document this project. We shall provide you the instructions for the documentation later on separately during the course. In general, you are building a data product and you will provide an overview of the overall architecture of your product and the results that you get.
9. The project code will be shared using a public Github (www.github.com) repository between you and the instructor. If you would like to use any other repository, then discuss it with the instructor. If you find Github difficult to use, then you can send the code as a zip file by an email.
10. Late submissions will be penalized with 20% loss of grade for the submitted part of the project.
11. Your project may be used as part of a research paper in future. In that case, you will be one of the co-authors of the paper.

Milestones

Date	**Objective**	**Deliverable**
January 30	Project abstract as described above[1]	A Word or PDF document on Blackboard
February 20	Literature review, data description and approach (template will be provided)[2,3]	A word or PDF document on Blackboard [3]
March 20	Initial Results and the Code [4]	Discuss by meeting with instructors
April 10	Final Results and the project report (template will be provided)[4]	A word or PDF document on the Blackboard*
April 17-20	Final presentations (10 minutes)	To be done in a classroom between 6pm to 9pm

1- Abstract: Clearly define the problem in your abstract. The abstract creation guidelines are already shown above.

2- Literature review: Read a few research papers and technical articles related to the same problem as yours

3- Don't put source code in the document, just provide the Github (or another) link

4- Data description: First clean your data and then provide summary statistics, number of attributes, correlation between attributes, interesting trends (such as outliers and possible reasons), etc.

5- Results: Think about simple replications/simulation based on your literature review for your dataset. You can also come up with new approaches.

Evaluation Criteria

Deliverables	Marks
Project abstract as described above	10%
Data Description and Approach Details (template will be provided)	20%
Initial Results and the Code	10%
Final Results and the project report (template will be provided)	40%
Final presentations (10 minutes)	20%

Github

Tutorial: https://try.github.io/

Github on Windows: https://windows.github.com/

Instructors

Amir Nabatchian - anabatchian@ryerson.ca

Ceni Babaoglu - cenibabaoglu@ryerson.ca

Naeem Irfan - naeem@ryerson.ca

Shariyar Murtaza - syed.shariyar@ryerson.ca

Tamer Abdou tamer.abdou@ryerson.ca