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. 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.
2. Here is a list of public repositories of data that you can use for your projects:

<https://www.data.gov/>   
<https://www.healthdata.gov/>   
<https://data.medicare.gov/data/hospital-compare>   
<http://www.dol.gov/open/data.htm>   
<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.kaggle.com/competitions>

Twitter data can be used too (you will need to use an API in R, Python, etc.) to get the twitter data. Any other public website can also be mined.

1. Here are some of the project ideas:
   1. 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.
   2. Analyze twitter data for sentiment analysis; e.g., classify tweets to medical and non-medical tweets and then identify health issues in different regions.
   3. 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 ☺).
   4. 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.
2. In the first week, you need to identify 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; (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, exploratory charts, etc.) and the tools that you are proposing to solve it. You will submit the abstract on the Blackboard.
3. Once you submit the abstract, one of the instructors will be assigned to you as an individual mentor for your project. 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.
4. 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.
5. The project code will be shared using a public Github ([www.github.com](http://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. There are 5 bonus marks for having code on Github.
6. 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. In case of the private dataset (provided NDA signed), then you will have to obfuscate the private information before submitting the final project documentation.

# Milestones

|  |  |  |
| --- | --- | --- |
| Date | Objective | Deliverable |
| May 10, 2015 | Project abstract as described above | A Word or PDF document on Blackboard |
| May 25, 2015 | Data Description and Approach Details (template will be provided) | A word or PDF document on Blackboard\* |
| June 1, 2015 | Initial Results and the Code | A word or PDF document on the Blackboard\* |
| June 10, 2015 | Final Results and the project report (template will be provided) | A word or PDF document on the Blackboard\* |
| June 15, 2015 | Final presentations (10 minutes) | To be done in a classroom between 6pm to 9pm |

\*Don’t put source code in the document, just provide the Github link

# Meetings with Instructors

There will be weekly meetings with the instructors. Each instructor will be available between 6pm and 9pm on Tuesday each week. Each student will have half an hour slot. The meetings will start from week 2. The exact meeting times for the students will be announced on May 11.

There will be no meeting on week 1. Please send your questions on the course forum on Blackboard. One of the instructors will respond as soon as possible.

# Github

Tutorial: <https://try.github.io/>   
Github on Windows: <https://windows.github.com/>

# Instructors

Bora Caglayan (bora.caglayan@ryerson.ca)

Shariyar Murtaza (syed.shariyar@ryerson.ca)

Parisa Lak (parisa.lak@ryerson.ca)