MET CS 777 – Big Data Analytics

Term Project (20 points)

1. Description

The goal of this project assignment is to give students an opportunity to build a large-scale data science project using real-world data. Students should select a public data set, think about a research project for example a clustering or classification problem or any machine learning model, implement the model using cluster computing platforms like Apache Spark, and evaluate the results.

This class project has two due dates:

1. Due date to submit a proposal
2. Final project submission date.

2. Project Proposal (Due date on blackboard)

Select a public data set and define a data science research question based on the data set. You can find a list of such public data sets at the end of this guide. You can select one of the listed data sets or search the web and pick up one of the available public data sets on the Web. Your data sets must not be large but should have the potential to be a large data set, for example, if I have access to 1000 newspaper articles and I can imagine that this data set can be larger than 1000 items.

You should write a proposal for your project and submit it for our approval. You submit a PDF or MS Word, or Markdown-formatted document.

You should describe the following items

1. What is your data set about?
2. What is exactly your research question? What do you want to learn from the data? What is your learning model,e.g., a Classification, Clustering, etc …?
3. What do you expect after the implementation of
4. How do you want to evaluate your project? How to access the correctness of your model? How well would you expect that the model will work?

3. Implementation

You need to use a cluster computing platform like Spark – but we do not limit it to Spark, e.g., you can also use Apache Flink or other systems.

You need to correctly implement your training model and test the model based on separating the data set into training and testing subsets.

Your code should compile and we should be able to read your README.md file to understand how to run your project. Provide in the README file clear instructions on how to run your project.

4. Presentation of Results

- Describe the results of your project in a professional way.
- You may want to visualize diagrams and describe the results based on some diagrams – but having diagrams is not a MUST have to get the full credit.
- Describe the model and results of your project in a way that every person in this field can read, enjoy and understand.
- You need to create a video presentation of your term project and upload the video to the course Media Gallery

5. Grading

There is no one correct answer for this project and there is no performance threshold based on which your grade is determined. We will grade the project based on

- the originality of the project idea – how nice and new it is
- correctness of the selected model and application scenario
- correct implementation of your projects
- presentation of study results.

6. Public Datasets

- https://www.gdeltproject.org/
- https://github.com/awesomedata/awesome-public-datasets
- https://github.com/apiad/datasets-list
- https://github.com/datasets/openml-datasets/tree/master/data the same data as this list https://www.openml.org/search?type=data

- https://archive.ics.uci.edu/ml/datasets.html
- https://www.data.gov/
- https://www.kaggle.com/datasets
- http://datamob.org/
- https://sites.google.com/a/drwren.com/wmd/details
- https://data.cityofnewyork.us/data
- http://snap.stanford.edu/data/index.html
- http://aws.amazon.com/datasets/

Event Data

- GeoLife GPS Trajectories http://research.microsoft.com/en-us/downloads/b16d359d-d164469e-9fd4-daa38f2b2e13/
- Activity Recognition in the Home Setting Using Simple and Ubiquitous Sensors (MIT Project 2004) http://courses.media.mit.edu/2004fall/mas622j/04.projects/home/
- https://sites.google.com/a/drwren.com/wmd/home **Frequent Itemset Mining Dataset Repository**
- http://fimi.ua.ac.be/data/

Airline On-time Performance

- http://www.eecs.wsu.edu/~yyao/StreamingGraphs.html
- http://openflights.org/data.html

Collection and Streaming of Graph Datasets

- http://www.eecs.wsu.edu/~yyao/StreamingGraphs.html

Data Streams

- http://www.quora.com/Where-can-I-find-public-or-free-real-time-or-streaming-data-sources
- 3 hourly weather forecast and observational data - UK locations http://data.gov.uk/dataset/metoffice_uklocs3hr_fc
- There is also an IRC chan with the live log of Wikipedia edits on the #en.wikipedia channel of the irc.wikimedia.org server. http://blog.programmableweb.com/2010/12/26/64-new-apis-realtime-facebook-twitter-streaming-and-google-shopping/

New York Taxi Datasets

- https://data.ny.gov/
- TLC Trip Record Data http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
- Another set published by Chris Whong http://chriswhong.com/open-data/foil_nyc_taxi/ this data set is used for the DEBS 2015 challenge http://www.debs2015.org/call-grand-challenge.html
- Another Description of this dataset http://publish.illinois.edu/dbwork/open-data/

OpenStreet Map

- http://wiki.openstreetmap.org/wiki/Planet.osm

Dublin Bus GPS sample data from Dublin City Council (Insight Project)

- https://data.gov.ie/dataset/dublin-bus-gps-sample-data-from-dublin-city-council-insight-project

  Further data set at https://data.gov.ie

Wikipedia Dump

- Wikipedia Dump https://dumps.wikimedia.org/

Amazon Review Data Downloader

- https://github.com/aesuli/Amazon-downloader

7. Academic Misconduct Regarding Programming

In a programming class like our class, there is sometimes a very fine line between "cheating" and acceptable and beneficial interaction between peers. Thus, it is very important that you fully understand what is and what is not allowed in terms of collaboration with your classmates. We want to be 100% precise so that there can be no confusion.

The rule on collaboration and communication with your classmates is very simple: you cannot transmit or receive code from or to anyone in the class in any way---visually (by showing someone your code), electronically (by emailing, posting, or otherwise sending someone your code), verbally (by reading code to someone) or in any other way we have not yet imagined. Any other collaboration is acceptable.

The rule on collaboration and communication with people who are not your classmates (or your TAs or instructor) is also very simple: it is not allowed in any way, period. This disallows (for example) posting any questions of any nature to programming forums such as **StackOverflow**.

As far as going to the web and using Google, we will apply the **"two-line rule"**. Go to any web page you like and do any search that you like. But you cannot take more than two lines of code from an external resource and actually include it in your assignment in any form. Note that changing variable names or otherwise transforming or obfuscating code you found on the web does not render the "two-line rule" inapplicable. It is still a violation to obtain more than two lines of code from an external resource and turn it in, whatever you do to those two lines after you first obtain them.

Furthermore, you should cite your sources. Add a comment to your code that includes the URL(s) that you consulted when constructing your solution. This turns out to be very helpful when you're looking