# MET CS 777 – Big Data Analytics

# Term Project

## 1. Description

The primary objective of the Term project is to enable students to build a large-scale data science project using real-world data. To accomplish this goal, students will need to select a public data set and formulate a research project, such as a clustering or classification problem or any other machine learning model. Once they have identified their research project, they will need to implement the machine learning model using Apache Spark and evaluate the results.

It is important to note that the Term project has two due dates. The first due date is to submit a proposal outlining your project, and the second due date is the final project submission. By adhering to these deadlines, you will be able to manage your time effectively and complete the project in a timely manner.

## 2. Project Proposal (Due date on blackboard)

For this project, you will need to select a public data set and define a data science research question based on the data set. You can choose from the list of provided public data sets or find one on the internet. It is important to note that the data set should not be too large but should have the potential to be scaled up. For instance, if you have access to 1000 newspaper articles, you should be able to imagine that this data set can be expanded to more than 1000 items.

To get started, you will need to write a proposal for your project and submit it for approval in either a PDF or MS Word. Your proposal should describe the data set you have selected and outline your research question. You should also provide information about your proposed machine learning model, such as whether you plan to use classification or clustering techniques.

Additionally, you should describe the expected outcomes of your project and how you plan to evaluate the results. It is important to explain how you will assess the correctness of your model and how well you expect it to work.

Once you have developed your proposal, please submit it for approval.

To successfully complete the project, you will need to describe the following items in your proposal:

- Data set description: Provide a detailed description of the public data set you have selected, including its source, format, and any relevant details about the data.
- Research question: Clearly define your research question and explain why it is important to study. What do you want to learn from the data?

- Machine Learning model: Specify the type of machine learning model you plan to use, such as classification or clustering, and explain why you have chosen this model.
- Expected outcomes: What do you expect to achieve after the implementation of your learning model? What do you hope to learn or discover from your analysis of the data?
- Evaluation plan: Explain how you plan to evaluate your project and assess the correctness of your model. What metrics or methods will you use to evaluate the effectiveness of your learning model? How well do you expect the model to work, and how will you measure its performance?

By providing detailed answers to these questions, you will be able to create a solid proposal that can guide your work throughout the project.

## 3. Implementation

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

- PySpark must be used, but you don't need to run it on Google Cloud. It is acceptable if the term project is done on your laptop.
- It is crucial to accurately implement your ML model and thoroughly test it by separating the data set into training and testing subsets.
- In order to ensure that your project runs smoothly, it is important that your code compiles without errors.
- To help us understand how to run your project, please provide clear and concise instructions in the README file, which must be easily accessible and easy to understand.

## 4. Grading

Your grade will not be determined based on a fixed performance threshold. Instead, we will evaluate your project based on several criteria, which include

- the originality of your project idea
- the correctness of the selected model and application scenario
- the correct implementation of your project
- the presentation of study results

## 5. Submission Guidelines

Please carefully review the following guidelines to ensure a smooth submission process:

- Code Files:
  - Include all the code files that are necessary for your project.
  - Organize your code files in a clear and structured manner, making them easy for others to navigate and understand.
- Data Files or Link:
  - If your project relies on specific data files, include them in your submission.
  - Alternatively, provide a link to the data file(s) if they are too large to be attached.
  - Ensure that the data files are well-documented and any dependencies are clearly indicated.
- Report (PDF/DOC):
  - Prepare a comprehensive report summarizing your project.
  - Include an introduction, methodology, results, discussion, and conclusion.
  - Use either a PDF or DOC format for your report.
  - Ensure that your report is well-structured and properly formatted.
  - When discussing your model and project results, it is important to do so in a way that is accessible to individuals from a range of backgrounds and levels of expertise.
- PowerPoint File:
  - PowerPoint offers a wide range of features and design options that can help you deliver a more dynamic and engaging presentation.
  - Utilize visual aids, animations, and other multimedia elements to enhance the audience's understanding and engagement.
- Video Presentation:
  - Create a video presentation that highlights the key aspects of your project.
  - The video should be clear, concise, and engaging.
  - Aim for a duration of 5-10 minutes. Please don't have anything more than 10 minutes.
  - Provide a walkthrough of your project, explaining its objectives, implementation, and results.
  - Your presentation should be engaging and easy to understand so that anyone in the field can learn from your work.

## 6. Examples of Public Datasets

**Dbpedia**

- https://www.dbpedia.org/resources/knowledge-graphs/

**Archivo - Ontology Archive**

- https://archivo.dbpedia.org/list

**Multilingual lexical data**

- http://kaiko.getalp.org/about-dbnary/ (http://kaiko.getalp.org/static/ontolex/latest/ )

**News Data**

- https://www.gdeltproject.org/

**General data sources**

- https://www.data.gov/
- https://www.kaggle.com/datasets
- http://aws.amazon.com/datasets/

**Large Network Dataset Collection**

- http://snap.stanford.edu/data/index.html

**Airline On-time Performance**

- http://openflights.org/data.html

**Data Streams**

- 3 hourly weather forecast and observational data - UK locations
  http://data.gov.uk/dataset/metoffice_uklocs3hr_fc

**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

**OpenStreet Map**

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

**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