# MIE 1624 Introduction to Data Science and Analytics – Winter 2021

## Final Exam Project

Deadline: Thursday, April 15, 11:59pm

## Background

In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared a dataset of open sourced research papers. This data set is a resource of over 400,000 scholarly articles, including over 150,000 with full text, about COVID-19. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in new coronavirus literature, making it difficult for the medical research community to keep up and extract insights from this growing body work.

The goal of this project is to use NLP and other machine learning algorithms learned in this course to develop a tool that can text-mine this database of research articles to gain useful insights about COVID-19 and how we might be able to tackle the outbreak, contain the spread, flatten the curve and improve vaccination efforts. The overarching insights that can be acquired from this dataset are numerous and which aspect of the problem you decide to tackle is up to you. For example you may choose to use this dataset to better understand the transmission, incubation and symptoms of COVID-19, look to gain insights around which therapeutics and vaccines may hold promise and warrant further investigation, or you may wish to investigate the risk factors that make COVID-19 particularly deadly in some patients. The underlying goal of this project is to gain insights from this dataset to better inform how our healthcare system, governments, industries can tackle this growing problem.

## Learning Objectives

1. Implement functionality to parse natural language biomedical literature data according to given constraints and requirements.
2. Train and test machine learning algorithms (especially unsupervised machine learning algorithms such as clustering, dimensionality reduction, recommender systems, association rules, etc.) in order to gain insights or answer the overarching question you have chosen to pursue for this project.
3. Understand how to apply machine learning algorithms to the task of learning from a large corpus of biomedical research text.
4. Improve on skills and competencies required to collate and present domain specific, evidence-based insights. Particularly, in this case to gain insights and guide the fight against the COVID-19 pandemic.

**To do:**

1. **Data Cleaning – [1 mark]**

   The dataset of research papers is provided to you as a .csv file and starter Python code is provided to you that cleans the data by removing duplicate papers, making the text contents easier to mine by adjusting formatting, and extracting useful fields from the larger dataset such as authors, abstracts, date of publication and more. You may choose to use this starter code and clean data that it produces, or if your chosen algorithm requires a different format of data or approach, you are free to modify and/or write you own data cleaning pipeline.

2. **Data Visualization and Exploratory Data Analysis – [4 marks]**

   Depending on your overarching theme and questions that you wish to address about COVID-19 present 3 graphical figures that visualize aspects or information in the data that you will further explore with your models. How could these trends be used to help with the task of methodically extracting all information and trends of this type? Consider how accessing the data and creating these visualizations will inform how the data will need to be pre-processed and fed into your models. All graphs should be readable and presented in the notebook. All axes must be appropriately labeled. In addition to data visualizations, perform exploratory data analysis in other forms, if necessary.

3. **Model selection and fitting to data – [12 marks]**

   Select a machine learning model of your choice (you may select an unsupervised or supervised machine learning model depending on your approach) that will allow you to study some aspect of COVID-19 from the corpus of research articles. You must justify your algorithm choices and the approach you will use to fit your model using the dataset provided. You may also choose to study multiple models and report on the suitability of each in addressing your overarching question regarding COVID-19. You should also use the dataset provided to train the models selected and discuss and interpret the findings of these models. You may also use this section to improve the model depending on the findings of your models and how you interpret them.

4. **Deriving insights about policy and guidance to tackle the outbreak based on model findings – [7 marks]**

   Using the findings from your NLP model and text mining 400,000 unique biomedical research papers on the coronavirus you are now tasked with discussing and proposing how scientists, doctors, nurses, healthcare professionals, industry and governments can best use the insights from your data science model to assist in the fight against the COVID-19 pandemic. Use the insights derived about the disease from your model and your data analysis to justify proposed policies or action items.

**The order laid out here does not need to be strictly followed. A significant number of marks in each section are allocated to discussion. Use markdown cells in Jupyter notebook as needed to explain your reasoning for the steps that you take.**

## Programming Tools:

- Software
  - Python Version 3.X is required for this assignment. Python Version 2.7 is not allowed.
  - Your code should run on the Google Colab Virtual Environment or CognitiveClass Virtual Lab (Kernel 3).
  - All Python libraries and built-ins are allowed but here is a list of the major libraries you might consider: numpy, Scipy, Scikit, Matplotlib, Pandas, NLTK.
  - No other tool or software besides Python and its component libraries can be used to touch the data files. For instance, using Microsoft Excel to clean the data is not allowed.
- Required data files
  - **metadata.csv**: a corpus of published research articles and corresponding metadata on findings about the corona virus. Paper abstracts are found directly in the file whereas the full text for the papers can be acquired using the links provided within.
  - .json files containing the full texts for papers in the above dataset can be downloaded from here: https://bit.ly/3fzhlhD
  - The data file cannot be altered by any means. The IPython Notebooks will be run using local version of this data file.
- Optional starter code
  - **FinalProjectstarter.ipynb**: a simple data reading and cleaning pipeline that will extract some useful fields from the metadata.csv file including the abstracts for each paper. You can choose to use the data return from this file, modify this file to better suit your data cleaning process or write your own data cleaning file from scratch. If you want to extract the entire text for each paper you will have to access the .json files provided with the dataset or can alternatively retrieve this data from the links provided within metadata.csv.

## What to Submit:

1. Submit via Quercus portal an IPython notebook containing your implementation and motivation for all the steps of the analysis with the following naming convention:

**lastname_studentnumber_finalproject.ipynb**

Comment out any data retrieval processes (e.g., accessing full-text of articles, downloading your own additional data, etc.) in your code and replace it with code for reading the corresponding data from files. (**Submit all those data files together with your Jupyter notebook**).

Make sure that you **comment** your code appropriately and describe each step in sufficient detail. Respect the above convention when naming your file, making sure that all letters are lowercase and underscores are used as shown. A program that cannot be evaluated because it varies from specifications will receive zero marks.

2. Submit 5 slides in PowerPoint and PDF formats describing your findings from exploratory analysis, model feature importance, model results and visualizations. Use the following naming conventions **lastname_studentnumber_finalproject.pptx** and **lastname_studentnumber_finalproject.pdf**

**Late Submissions:**

- up to 2 hours late - no penalty
- one day late - 20% penalty
- more than one day late - 0 mark

**Tips:**

1. You have a lot of freedom with however you want to approach each step and which library or function you want to use. As open-ended as the problem seems, the emphasis of the project is for you to be able to explain the reasoning behind every step.
2. While some suggestions have been made in certain steps to give you some direction, you are not required to follow them. Doing them, however, guarantees full marks if implemented and explained correctly.