CSCM35, CSLM35 Big Data and Data Mining

by Dr. Jingjing Deng

Released on 30th March 2020

Submission

In this coursework, you will be given a practical data mining task to complete which consists of technical report writing and software programming. Source code must be organised and formatted neatly. Sufficient and clear comments are very welcome and necessary for markers to assess your work. The technical report must be submitted in **PDF** format. Submissions and feedback will be done via Blackboard-Tunitin system. Plagiarism will not be tolerated. Zip all your files with the following naming convention for submission:

- [Student Number]-[Last Name][First Initial]-[Coursework][Number].zip
- For example: *123456-DengJ-Coursework1.zip*

Policy

- To be completed by students working individually.
- Feedback: individual feedback is given on Blackboard within two weeks of deadline.
- Learning outcome: The tasks in this coursework are based on both your practical work in the lab sessions and your understanding of the theories and methods of data mining. Thus, through this coursework, you are expected to demonstrate both practical skills and theoretical knowledge that you have learned in this module. You will also formally present your understandings through technical writing. It is an opportunity to apply analytical and critical thinking, as well as practical implementation.
- Unfair practice: This work is to be attempted individually. You may get help from your lecturer, academic tutor and lab tutor, but you may not collaborate with your peers. Copy and paste from the Internet is not allowed. Using external code without proper referencing is also considered as breaching academic integrity.
- Submission deadline: The report and your Python 3 implementation need to be submitted electronically to Blackboard by the deadline.

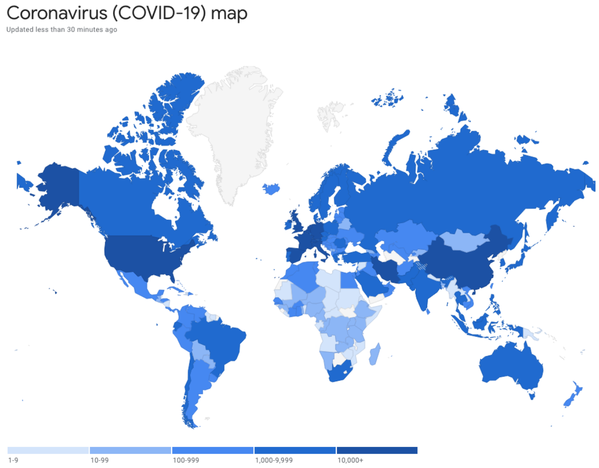
CSCM35/CSLM35 Coursework 2 Complete by 01/05/2020

Task: Data Mining Practice – COVID-19

On December 31, 2019, the World Health Organization (WHO) was informed of an outbreak of "pneumonia of unknown cause" detected in Wuhan City, Hubei Province, China – the seventh-largest city in China with 11 million residents. As of January 23, there are over 800 cases of 2019-nCoV confirmed globally, including cases in at least 20 regions in China and nine countries/territories. Infected travelers (primarily air) are known to be responsible for introductions of the virus outside Wuhan. On Jan 13 Thailand reported the first international case outside China, while the first cases within China, but outside of Wuhan were reported on January 19, in Guangdong and Beijing. On January 20, China's National Health Commission (NHC) confirmed that the coronavirus can be transmitted between humans. On the same day human infections with 2019-nCoV had also been confirmed in Japan and South Korea, and the following day cases in the U.S. and Taiwan were detected in travelers returning from Wuhan. On Jan 22, a WHO emergency committee convened to discuss whether the outbreak should be classified as a public health emergency of international concern (PHEIC) under International Health Regulations. As of March 29 2020, the coronavirus COVID-19 is affecting 199 countries and territories. There are over 677,622 cases of 2019-nCoV confirmed globally, where 141,698 cases are recovered and 31,750 cases are dead.[1]

As a data scientist, can you help the governments, global health service and institutions to make informed decisions to fight against COVID-19? In this coursework, your tasks are to:

1. Given the dataset, identify a research question and propose a feasible solution using data mining methods. For example, can we identify the next outbreak geographically?
2. Develop a prototype of your method, evaluate your method to demonstrate the proof of principle concept. For
3. Write a technical report to present the challenging problem, proposed method, experimental results and findings.



[1] Text and data are taken from: `https://systems.jhu.edu/research/public-health/ncov/`

Data: 2019 Novel Coronavirus COVID-19 (2019-nCoV)

2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository is the main data source of this coursework that is maintained by Johns Hopkins CSSE in real-time. The data are stored in *CSV* with two different formats, daily reports and time series, which are in the folder namely *csse_covid_19_data*. You can visually explore the data using geographic visualization method. There are two web-based visualisation tool-kits that maps the data onto its geographic locations that can be accessed from the links given as follows:

- `https://bit.ly/2wDkQ2K` [Direct Data Source: Johns Hopkins CSSE]
- `https://www.google.com/covid19-map/` [Direct Data Source: WHO]

The data and more details can be found in GitHub Repository `https://github.com/CSSEGISandData/COVID-19`[2]. The coronavirus disease (COVID-2019) situation reports that are published daily by WHO can be found from `https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports`. You can use other relevant datasets if needed, *i.e.* News dataset, global flight dataset *etc.* Please cite the data sources that are used in the report accordingly.

You may also find the following Python libraries useful for this coursework:

- Dataset File IO: Pandas, `https://pandas.pydata.org/`.
- Data Visualisation: Matplotlib, `https://matplotlib.org/`;
- Top 50 matplotlib Visualizations – The Master Plots, `https://bit.ly/2BBlWLE`;
- Machine Learning: Scikit-Learn, `https://scikit-learn.org/stable/`;
- Numerical Computing: `http://www.numpy.org/`, `https://www.scipy.org/`.

Assessment: Code and Report [50 marks in total]

This is an open challenge question, where you are asked to identify a data mining problem, propose a solutions and evaluate it on the datasets. You are required to write a technical report to summarise the proposed solutions, the initial results and findings. The report is expected to demonstrate your understanding of data mining in depth and what application can be derived from those methods given the dataset. Writing a convincing proposal also requires a good demo which supports your arguments and conclusions. Your report should contain the following sections:

1. Introduction[**8 marks**]: Provide a short overview of the problem that you are going to address and your solution in general. This should cover how you formulate the problem from data mining perspective, which data mining methods you are using and how to apply the problem. You can combine multiple data mining algorithms to design your own solution. Novelty of your solution will be assessed.
2. Proposed Method[**15 marks**]: Describe all technical details of how your solution is implemented and how the evaluation experiments are carried out step-by-step. Reproducibility of your experiment will be assessed. When alternative algorithms can be used, you can perform comparison study to justify the advantage and disadvantage of certain methods.

[2] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis; published online Feb 19. `https://doi.org/10.1016/S1473-3099(20)30120-1`.

Provide relevant and meaningful experimental results, such as accuracy evaluation, statistical analysis and visualisation of performance and efficiency when necessary. This section may contain multiple sub-sections.

3. Result and Discussion[**10 marks**]: Provide an in-depth discussion of the comparison results on performance, efficiency of different data mining methods and your critical analysis on findings in general through those studies.

4. References [**2 marks**]: Include references where appropriate. **The reference section is not included in the page limit.**

5. Solution Code [**15 marks**]: The code may contains multiple source files. Attach all of them in your submission including proposed solution, experimental and comparison studies, visualisation and statistical analytics when necessary. A *"README.txt"* file describing how to run your codes is required. **DO NOT** include any source code to your report.

Page Limit: The report should be **no more than 6 pages**. Font size should be **no smaller than size 10**, and the text area is approximately 9.5x6 inches. You may use images but do so with care; do not use images to fill up the pages. You may use an additional cover sheet, which has your name and student number. **Reports that exceed the specified page limit will result in penalties: 3 marks deduction for every over-length page.**

Source Code: Submit your Python 3 source code to Blackboard, together with your report, in a **Single Zip** file.