COMP20008 Project and Presentation (50 marks total)

V1.1: Updated 13 April 2016 (Phase 2 due date modified)

Hypothetical Scenario

The Victorian Minister for Data Science and the Mayor of the Melbourne City Council wish to understand more about how open data can be used to benefit Melbourne.

At a high level, they would like to see demonstrations of how open data can be wrangled to gain insight into issues affecting Melbourne, for a broad range of areas such as transport, health, business, education, tourism, the environment, communities, the arts, commerce, public amenities, employment, sport, usage of facilities, real estate, finance or urban planning.

You are a data science consultant who is hoping to convince the Minister and the Mayor about the benefits of open data. In the first phase, you will write a proposal, outlining a question in a chosen domain that is relevant to Melbourne and proposing a data wrangling project to demonstrate the benefits of processing open datasets to answer this question. In the second phase, you will commence the investigation, generate initial findings, provide an interim report on what you have learnt so far and summarise your progress on initial deployment of data wrangling techniques. In the third phase you will make a brief oral presentation about what you have done. In the fourth phase you will deliver a written report outlining your methodology and findings.

Objective

The aim of the project is to provide experience in processing some real world datasets, cleaning them, integrating them, analysing them and visualising them.

This project will be done individually. You will need to

- Choose a domain (education or sport or transport or the environment or health, etc)
- Propose a question for your domain that relates to Melbourne and for which an answer would be likely to interest politicians or policy makers in the Melbourne City Council or Victorian Government.
- Identify at least 2 open datasets that can be linked together to help shed light on this question.
- Using Python, process these datasets, integrate them and provide analysis and visualisations which help answer the question you have posed.

You are not expected to develop an interactive tool for browsing your chosen datasets. Rather, the results of your investigation can be reported as tables or graphs or static visualisations suitable for inclusion in a written report or in slides of a Powerpoint presentation.

Datasets

The LMS Project page has a list of repositories that can be used as a starting point for finding datasets. Data from any of these is fine to use.

You are also welcome to use other datasets that have been made publically available by reputable entities, or which are readily available via a registration process open to University of Melbourne staff/students.

You should not use datasets that have been illegally obtained or published (E.g. data violating copyright permissions or that has been hacked).

If in any doubt as to whether a particular dataset is ok to use, please post a question on the discussion forum or contact the lecturer for clarification.

Phase 1: Concept Formulation and Pitch (10 marks)

Your task for this phase is to write a proposal describing your domain, the question to be investigated and the datasets that will be used to answer the question.

Please submit a pdf file of no more than 2 pages (11pt A4 paper) providing answers to the following 8 pieces of information, in sequence:

1. Title of Project (choose this according to your chosen domain and question)

2. Domain: Select either one of, or a combination of: transport, health, business, tourism, sport, education, the environment, communities, the arts, commerce, public amenities, employment, usage of facilities, real estate, finance or urban planning.

3. What is the question you are seeking to answer? Who would be interested in an answer to this question and why? How might the information be used and who could it benefit?

4. In what respects will your answer to this question provide innovative information? (you do not want to have a question which is trivial, or for which the answer already publically exists and can be readily found).

5. Datasets: What are the datasets (minimum of 2, maximum of 3) that you will use? Provide a brief description of the information in each and a link (URI) to where the dataset can be downloaded.

6. What difficulties and challenges do you envisage in processing, integrating and visualising these datasets to answer your question?

7. In what ways will your processing, integration and visualization add value compared to having just the raw data?

8. How much code (in Python) do you estimate will need to be written from scratch? What are the major Python libraries that you will make use of? What other other publically available code do you intend to use?

Marking Scheme for Phase 1

Marking scheme will include consideration of the following factors:

- Completeness and clarity of answers for the 8 items.
- To what extent is answering the question likely to provide interesting and innovative results? How well does the question fit with the hypothetical scenario outlined at the beginning of this project specification document?
- To what extent is addressing the proposed question likely to be an ambitious and complex exercise?
- How feasible is the proposal?

Phase 2: Initial Investigation (5 marks)

After completing phase 1, you will start your data wrangling in order to answer the proposed question. i.e. download the datasets and commence cleaning/transforming and integrating, as well as doing some pilot analysis/visualisation.

After performing some exploratory analysis, it is likely you may need to revisit and adjust your original question and proposed methodology. Your report for phase 2 provides the chance to do this.

Your deliverable for this phase is a short progress report (2 pages in length). This report needs to provide enough detail to demonstrate that you are making progress, that you have re-evaluated/refined your initial question based on some initial results, that the remainder of the project is feasible and likely to provide interesting results.

A detailed description of the sections to include in this report, as well as marking scheme will be released end of March.

Phase 3: Oral Presentation (10 marks)

In your workshop you will make a short (5 minute) presentation on what you have found. You will be expected to develop slides in powerpoint or pdf.

A detailed description of the sections to include in this presentation, as well as marking scheme will be released early April.

Phase 4: Final Report (25 marks)

In this phase, you are asked to write a report of 8 pages (11pt A4) providing the following information

- Title of project
- Domain
- Question

- Datasets selected (minimum of 2 and maximum of 3). A description of each (what is in the dataset and the schema) and a URI of where it can be downloaded.
- A description of what methods you applied to the datasets and why. A description of any issues encountered and limitations of your methods.
- A description of how the datasets were integrated and a description of any issues encountered and limitations of your methods.
- A description of how you did the analysis/visualisation of the integrated data and any issues encountered. Presentation of the results/findings (tables/graphs and discussion) and limitations of your methods.
- An explanation of how your results help answer the question that you proposed.
- An explanation of who might be interested in the results and why. What are the potential benefits? How does your processing, integration and visualisation add value compared to having just the raw data?
- A bibliography listing any references that you have you used.
- A separate zipfile of the python code that you wrote yourself for the project. This code should include comments explaining about any libraries or external code that you used.

Indicative Marking Scheme for Phase 4 (25 marks)

A marking scheme for Phase 4 will be released in mid April. It will cover the following

- to what extent does your solution make use of the technologies, methods and data representations presented in the subject (HTML, XML, JSON, Linked Data, missing value methods, data transformation methods, outlier detection methods, visualisation methods, clustering methods, linkage and integration methods, correlation and feature ranking methods)? How complex is your use of the these technologies, methods and data representations?
- to what extent does your question and its answer reflect a complex domain?
- to what extent has your processing added value to the raw datasets?
- to what extent has your data wrangling helped answer the question that was originally posed?
- implementation: how well have you implemented your code ? The implementation should be easy enough to follow by a tutor.
- how complete is your code documentation?

Due Dates

- Phase 1: 12pm 4th April. Submission will be via the LMS.
- Phase 2: 12pm 25th April. Submission will be via the LMS.
- Phase 3: Your presentation will be scheduled during the workshop you are enrolled in. It will be held in either Week 11 (16-20 May) or Week 12 (23-27 May). A schedule will be finalised once enrolments in the subject are finalised.
- Phase 4: 12pm Friday 20th May. Submission will be via LMS.

Extensions and Late Submission Penalties: If requesting an extension due to illness, please submit a medical certificate to the lecturer. Late submissions without an approved extension will attract a penalty of 10% of the marks available for that phase per 24hr period (or part thereof) that it is late. E.g. A late submission for phase 1 (10 marks total) will be penalised 1 mark if 4 hours late, 2 marks if 28 hours late, 3 marks if 50 hours late, etc.

Phases 1,2 and 4 are expected to together require 45-50 hours work. Phase 3 is expected to require 10-12 hours work.

Academic Honesty

You are expected to follow the academic honesty guidelines on the University website `https://academichonesty.unimelb.edu.au`

Further Information

The Project page on the subject LMS will contain a list of frequently asked questions.

A project discussion forum has also been created on the subject LMS. Please use this in the first instance if you have questions, since it will allow discussion and responses to be seen by everyone.