ICS5111 Mining Large Scale Data
Group Project

Joel Azzopardi, Lizzy Farrugia

December 7, 2023

This document contains the details for the course's project which this year is aligned with theme of **Digital Health**. The goal is to identify a problem related to health/healthcare, and to use and aggregate a number of publicly available data sources to develop a solution for the chosen problem.

Individuals are encouraged to work in teams of TWO or THREE. While discussions between individual groups are encouraged, the final deliverables need to be your own and not plagiarised in any way. You can decide with whom you want to team up. However, if you do not manage to team up yourselves, then we will form the teams, and such team setups will be final. The work within each team has to be distributed fairly, and in the documentation you will need to describe how the work was distributed and who was responsible for which part of the project. The mark given to each team member is determined based on the quality of that member's contribution to the team's overall effort. Marks assigned to different members of the same team *may* vary. The mark for this project is 100% of the global mark for this study unit.

The **deadline** for this assignment is **12:00pm Friday 2nd February 2024**. Deliverables, with attached and signed plagiarism form, must be submitted on the VLE. Late projects will be penalised.

1 Introduction

The aim behind this project is the research, design and implementation of an application that can be either web-based or stand-alone. At least **two data sources** (one structured and one unstructured) need to be identified that contain information about the chosen idea.

Some of the ideas that you can consider include (but you can of course come up with your own idea related to the theme of digital health):

- Diabetes (e.g. keep track of examinations, medications etc);
- Patient Education (e.g. automatically identify patients to target with suitable educational content related to their medical checkups);

- Data analysis (detect anomalies in a patient's data);
- Healthcare emergency support (e.g. resuscitation emergency);
- Mental health support (e.g. personalised time/goal management);
- Hospital HR system (e.g. to support optimized scheduling of beds).

What you need to do:

i. Select ONE of the provided use cases or come up with a related use case of your own to tackle;

ii. Identify at least TWO related research paper (check out Papers with Code[1]);

iii. Identify (and use) at least TWO data sources (structured and unstructured) that are openly available;

iv. Identify the technologies and techniques for your solution. These should include those used during the lectures, but may include others as well;

v. Develop an architecture/design explaining the how, why, what technologies/techniques are used and how your solution leverages on the data;

vi. Design and develop a working Proof of Concept (POC) through which you will be able to explain and showcase how the solution would work in real life;

vii. Write a scientific report detailing the work done.

2 Datasets and Data Sources

At least TWO sources need to used and these should include a mix of structured in nature (e.g. CSV, or databases) and unstructured (e.g. text repositories, documents, scientific papers or news reports).

Here are some relevant data sources to start you off:

- https://healthdata.gov/
- https://bchi.bigcitieshealth.org/
- https://www.who.int/gho/en/
- https://www.europeandataportal.eu/data/datasets?locale=en&categories=heal&keywords=human-health-and-safety&page=1

[1] https://paperswithcode.com/

- https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/
- https://allenai.org/data

However, once you choose the topic of interest, you can use any dataset/s relating to that topic. If possible, do make use of data from the local sphere, but this is **not** a necessary requirement and should not restrict the scope of the proposed solution.

3 Deliverables

The following is information about the deliverables that you will need to present at the end. Each component is individually marked.

3.1 D1: Proof of Concept and Documentation

You need to design and implement a POC that solves the identified problem and that uses techniques and technologies that were discussed in class. You can of course, also make use of other techniques and technologies that you think are suitable for the solution.

The POC needs to be packaged in a way that it is easy for us to execute and assess. The use of Jupyter[2]/Google Colab[3] notebooks is encouraged.

Documenting and presenting your work is very important. You are to discuss your contribution as a scientific report that is formatted using the ACM LaTeX class template[4]. The maximum page limit for this report is of 8 pages (including figures, tables and references).

The title of the report should be related to the selected topic. Furthermore, the report is expected to include the following sections, however you can include further sub-sections if you think that these will improve the structure and readability of the report:

a. *Introduction*: a brief explanation of the problem that is being addressed. This should be concise and highlights the main aim and objectives *(10 marks)*;

b. *Related Work*: briefly discuss existing research related to the problem addressed. Include literature that uses similar technologies and technique/s *(20 marks)*;

c. *Design and Implementation*: this section should include details about the design and implementation of your solution. Consider discussing: *(60 marks)*.

[2] https://jupyter.org/

[3] https://www.acm.org/publications/proceedings-template

[4] https://www.acm.org/publications/proceedings-template

- how the data was handled (discuss tasks such as data scraping, data collection, data storage, pre-processing, missing values etc);
- the architecture/design explaining the how, why, what technologies are used and how your solution leverages open data;
- the development of the POC to explain and showcase how the solution would work in real life;
- any experiments performed and how these address objectives;
- any challenges that were encountered and how they were resolved;

d. *Conclusion*: provide a critical appraisal of the solution: what were the strong and the weak points the approach used? what worked well and what did not and how this work can be extended in the future *(10 marks)*;

Deliverable will be marked out of 100%, however it is equivalent to **80% of the total mark**

3.2 D2: Presentation and Demo

A session will be held during which you will be allocated 10 minutes to do a short presentation and 5 minutes to demo the solution.

Deliverable will be marked out of 100%, however it is equivalent to **20% of the total mark**

3.3 Summary of deliverables

D1:	Design and development of the POC and report	**80 marks**
D2:	Presentation and Demo	**20 marks**

You need to upload the following onto the VLE:

i. A PDF of the scientific report, together with duly filled-in plagiarism form;

ii. An archive file which includes the developed prototype (source/executable/notebook and any other relevant intermediary data)

iii. the 10-minute presentation.

4 Final Note

If you have difficulties do not hesitate to contact us and/or post questions on the VLE forum.

Good luck!!