

# COMP47670 Assignment 1: Data Collection & Preparation

**Deadline:** Monday 23rd March 2020

## Overview:

The objective of this assignment is to collect a dataset from one or more open web APIs of your choice, and use Python to preprocess and analyse the collected data.

The assignment should be implemented as a single Jupyter Notebook (not a script). Your notebook should be clearly documented, using comments and Markdown cells to explain the code and results.

## Tasks:

For this assignment you should complete the following tasks:

### 1. Data identification:

- Choose at least one open web API as your data source (i.e. not a static or pre-collected dataset). If you decide to use more than one API, these APIs should be related in some way.

### 2. Data collection:

- Collect data from your API(s) using Python. Depending on the API(s), you may need to repeat the collection process multiple times to download sufficient data.
- Store the collected data in an appropriate file format for subsequent analysis (e.g. JSON, XML, CSV).

### 3. Data preparation and analysis:

- Load and represent the data using an appropriate data structure (i.e. records/items as rows, described by features as columns).
- Apply any preprocessing steps that might be required to clean or filter the data before analysis. Where more than one API is used, apply suitable data integration methods.
- Analyse, characterise, and summarise the cleaned dataset, using tables and plots where appropriate. Clearly explain and interpret any analysis results which are produced.
- Summarise any insights which you gained from your analysis of the data. Suggest ideas for further analysis which could be performed on the data in future.

## Guidelines:

- The assignment should be completed individually. Any evidence of plagiarism will result in a 0 grade.
- Submit your assignment via the COMP47670 Brightspace page. Your submission should be in the form of a single ZIP file containing the notebook (i.e. IPYNB file) and your data. If your data is too large to upload, please include a smaller sample of the data in the ZIP file.
- In the notebook please clearly state your full name and your student number. Also provide links to the home pages for the API(s) which you used.

- Hard deadline: Submit by the end of Monday 23rd March 2020
  - 1-5 days late: 10% deduction from overall mark
  - 6-10 days late: 20% deduction from overall mark
  - No assignments accepted after 10 days without extenuating circumstances approval and/or medical certificate.