COMP41680 Assignment 1: Data Collection & Preparation

Deadline: Monday 25th March 2019

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.

The assignment can be completed either individually or in pairs.

Tasks:

In this assignment you should complete all of the following tasks:

1. Choose at least one open web API as your data source (i.e. not a static dataset). If you decide to use more than one API, these APIs should be related in some way.
2. 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.
3. Parse the collected data, and store it in an appropriate file format for subsequent analysis (e.g. plain text, JSON, XML, CSV).
4. 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/filter/combine the data before analysis. Where more than one API is used, propose suitable data integration methods.
5. Analyse and summarise the cleaned dataset, using tables and visualisations where appropriate. What insights does this analysis offer about the dataset?

Guidelines:

- The assignment can be completed either individually or in pairs. Any evidence of plagiarism will result in a 0 grade.
- Submit your assignment via the COMP41680 Moodle page. Your submission should be in the form of a single ZIP file containing the notebook and your data. In the notebook please clearly state your full name(s) and student number(s). Students working in pairs only need to make one submission from either student.
- Hard deadline: Submit by the end of Monday 25th March 2019
    - 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.