COMP41680 Assignment 1: Data Collection & Preparation

Deadline: 5pm, Friday 31st March 2017

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 pre-process and analyse the collected data.

The assignment can be implemented as either an IPython Notebook or a standalone Python script. Your code should be clearly documented, using comments and/or Markdown cells.

Tasks:

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

1. Choose one or more open web APIs as your source of data. If you decide to use more than one API, the APIs should be related in some way.

2. Collect data from your chosen API(s) using Python. Your dataset should contain at least 100 records/items in total. 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 pre-processing steps that might be required to clean/filter/combine the data before analysis.

5. Analyse and summarise the cleaned dataset, using tables and visualisations where appropriate.

Guidelines:

- Submit your assignment via the COMP41680 CS Moodle page. Include your full name and student ID number with your submission.

- Include all (or a sample of) your data with your submission. Your submission should be in the form of a single ZIP file containing code and data.

- The assignment should be completed individually. Any evidence of plagiarism will result in a 0 grade.

- Hard deadline: Submit before 5pm on Friday 31st March 2017
    - 1-5 days late: 10% deduction from overall mark
    - 6-10 days late: 20% deduction from overall mark
    - Assignments will not be accepted after 10 days without an extenuating circumstances form and/or a medical certificate.