Case Studies in Government Analytics Assignment

Due Date: Thursday March 21, 2019

Assignment Goal:
The assignment will test the ability of students in applying their data analytics skills and knowledge to solving problems in the Government domain. The task will provide an opportunity for students to learn how to access and analyse unstructured data from Social Media pages (i.e. Facebook, Twitter) or alternatively access, integrate and analyse available datasets on open data platforms.

Students are expected to tackle ONLY ONE of these two tasks specified below.

Task 1 - Determining the trending policy topics in Galway based on analysis of Facebook pages or tweets from Local Media house (Newspaper or Radio station)

The task here entails first developing an R or Python program to extract the public posts (and related comments if available), second identifying posts and comments containing mentions of policy keywords, third counting the posts/comments associated with each policy and fourth creating a chart to display the results.

Challenges:
There are four primary challenges associated with this task:

1. **Accessing the Posts on a Facebook page or tweets from twitter (20 Marks)**
   Students will need to identify the appropriate API for connecting to a Facebook or Twitter page (e.g. https://www.facebook.com/galwaybayfm) and extract public posts/comments in the last 6 months.
2. **Identifying relevant posts from those obtained in step 1 (30 Marks)**
   This step entails determining a set of relevant keywords that will be used to filter posts and then determining an efficient operation for filtering the relevant posts/comments obtained from step 1.
   a. Choose keywords from public open data sources, manual inspection of data samples.
   b. Leverage existing topic modelling techniques (e.g. LDA)
3. **Analysis of the relevant posts to determine to provide required information (counts of posts per policy area) (20 Marks)**
   This step involves counting the posts/comments related to each policy area and storing the summary information in table (a CSV file).
4. **Visualisation of the result (30 Marks)**
   This step requires the student to determining the best way to visualise the tabular data obtained from step 3 which best shows the popularity of a policy.

Task 2 – Analysis of "OP Waiting List" datasets available at National Open Data Portal
The task for the student here is to develop an R or Python program to process OP Waiting List* available on National Open Data portal (http://data.gov.ie/). This involves connecting to the CKAN platform and accessing the datasets and associated files for each year on the portal. In addition, the student will be required to integrate files associated with the dataset, analyse them by aggregating the counts by hospital and finally creating a chart to display the changes in OP waiting list over time https://data.gov.ie/dataset/op-waiting-list-by-group-hospital (consider datasets from 2014 to 2018):

Challenges:
There are four primary challenges associated with this task:

1. **Accessing the datasets available on Open Data Platform (20 Marks)**
   Students will need to connect to the National Open Data portal at data.gov.ie using the CKAN API for R to access `OP Waiting List By Group Hospital dataset` and fetch all the related files.
2. **Integrating the datasets into one dataset (30 Marks)**
   This step entails aggregating the selected files obtained in step 1 into one single file. Thus, the resulting dataset will contain consolidated data from different years.
3. **Aggregate counts (20 Marks)**
   The third step involves generating a summary table for `OP Waiting list` showing the changes years wise across several dimensions. This summary table should be available as a CSV file or any other open data format.
4. **Visualisation of the result (30 Marks)**
   This final step involves presenting the data in the Table generated in Step 3 as a chart.

Groups

This is a group assignment. Each group should have 2 students. Groups are self-selected. Please email your selection to Tarek at tarek.zaarour@insight-centre.org by Monday 11th March. If you cannot form a group, please inform Tarek.

Report Format

Please submit a short report (approximately 2-3 pages in addition to your code) specifying:

- Your name, class and student numbers.
- A description of your algorithm and design decisions
- Your tests and results
- Conclusions and observations

Submission Instructions

- Please put your code into a single .zip archive with name "YourNames_CaseStudyAssignment3_code.zip", submit via Blackboard
- Please include the signed "Group Assignment Submission Form" available in the assignment folder on blackboard.
- Include a screenshot of the output of your application for any relevant output.
- Include all source code files (that is, files with name ending .java etc) required to compile and run your code.
- Use comments to explain your source code. Insufficient comments can lead to mark deductions.
- Please put your report into a single .pdf file with name "YourNames_CaseStudyAssignment1_report.pdf", submit via Blackboard
- Please note that all submissions (both code and report) will be checked for plagiarism.
- Plagiarism (from another group or other sources) are not allowed and would be treated seriously.