THE UNIVERSITY OF SYDNEY

PROJECT

ONE

Project Stage 1: Data Finding, Cleaning and Loading

Due date: 10:00pm Thursday of Week 6 (2017-04-13)

This task is worth 5% of your final assessment.

Project Stage 1

For this task, you need to obtain a data set. This may be any data that interests you. We prefer that you use publicly available data (so we can check your work if we need to) but it is OK for you to work on privately-owned data as long as you have permission to use it, and permission to reveal it to the markers. As you will see in the marking scheme, it is desirable that the data be sufficiently large that automated processing shows genuine benefits.

You are then expected to do whatever transforming and cleaning is appropriate, to get the data so it can be analysed by the tool of your choice. The details of this aspect all vary a lot, depending on both the format of the data you obtained, and the tool you will use for processing. For example, you might have "scraped" data in the form of an HTML page source, and want to get it into CSV format to use with a Unix pipeline, or alternatively you may have a JSON file and want to process it with Python. In any case, there will almost certainly be a need to do some data cleaning (such as removing instances that have corrupted or missing values, or correcting obvious spelling mistakes, etc).

Finally, we ask you to show one very simple analysis, that picks out some subset of the data and reports on some aggregate summaries. This is not intended to be a detailed exploration of the data (that will come in Stage Two), but simply a demonstration that the data is now in a form where you can work with it.

There are four deliverables in this Stage of the Project.

- Submit a written report on your work, in pdf. This should be submitted through Turnitin, via the link in the eLearning site. The report should have a three-section structure that corresponds to the marking scheme: a section that describes the data source, the contents of the data, and what your interest is in this; a section that describes the initial transformation and cleaning that you did (if you did this automatically, include here the code that you used, or a description that is detailed enough to be reproduced); and a section that describes and explains a simple analysis that you have done (including saying which tool you used, and illustrating the output of the analysis).
- Submit a copy of the raw data as you obtained it. This should be submitted through the eLearning system, as a single file (if you got multiple files from your source, you need to tar them into a single file for submission)
- Submit a copy of the cleaned and transformed data set, that you will use in your tool of choice. This should be submitted through the eLearning system, as a single file.
- Submit a copy of the simple analysis you have done. This should be submitted through the eLearning system, as a single file. The nature of the file will vary depending on the tool you chose: if you are processing with Excel, then you submit a spreadsheet; if you are processing with Unix, submit a text file that contains the processing pipeline; if you are processing with SQL, submit a text file that contains the sequence of queries; if you are processing with Python, submit a Python program.

Here is the mark scheme for this assignment.

- There is 1 mark for the work on obtaining a data set (as described in Section 1 of the report, and as evidenced in the submitted raw data set). A pass (adequate) score indicates that the data is genuine, that you have clearly showed where you obtained the data, that you have described the contents of the dataset (explaining clearly both the format, and the meaning of the various aspects). A distinction level score (good work) is awarded if, in addition to the above, the amount of data is at least 100 items, and your description shows clearly that you have appropriate rights to use the data in the ways that you do use it, and your explanation shows sensible reflection of the strengths and limitations of the data that you obtained. Full marks (excellent work) indicates that you have achieved all the distinction-level requirements and in addition, that your data set has at least 500 items, and that your data set is produced by combining data from more than one source[1]
- There are 2 marks for the work on transforming and cleaning the data set to support later processing in the tool of your choice (as described in Section 2 of the report, and as evidenced in the changes between the raw data set and the cleaned data set). A pass score indicates that you have produced a version of the data that is able to be used by your tool. A distinction score indicates that you have passed and also that you have carefully examined the source data set for data quality and format difficulties, and that you have dealt reasonably with several of these issues. Full marks is awarded if, in addition, your transformation and cleaning was to a substantial extent, an automated process. If you have found a dataset where the data is already clean, you can instead show how you check the data cleanliness and quality properties.
- There are 2 marks for the simple analysis work (as described in Section 3 of the report, and evidenced in the submitted analysis). A pass score is awarded if you have written an analysis that can produce output that correctly derives some aggregate value over some subset of the data (for example, it might give the maximum value of one attribute, among all items with a given value for some other attribute). A distinction score is given if you can produce output that gives the aggregates over multiple subsets (for example, the maximum value of some attribute in each subset corresponding to a value of another attribute; this might be output all the aggregates in one table like a SQL group-by or Excel pivot table, or it might have a changeable parameter that determines which subset is to be aggregated over, like a Unix shell command with a command-line argument that sets the value of the attribute). Full marks would be awarded for doing the above where your analysis combines and connects data that originated in different data sources [for example, the attribute you aggregate may be from a different source than the attribute used to determine the subset; clearly this is only possible if you used data from more than one source].

Academic Honesty

This assignment forms part of the assessment for INFO1903. As such, the work you submit must be your own (except where you explicitly acknowledge sources). Please make sure that you have understood the University's Academic Integrity expectations. By submitting work, you will be declaring that you are aware of the policy, that the work is your own, and that you understand that the University may reproduce the work and communicate it to others, in order to run similarity detection software. We urge that if you haven't yet done so, you complete AHEM1001, the online module on Academic Integrity that the University is offering.

[1] To be considered for full marks, there must be a real challenge in relating the data values in the two sets. It is not enough to simply take two datasets from the same authority (that use the same definitions of attributes etc), nor is it ok just to use unrelated data, where there is not connection made across the information.