COMPSCI 747 — Computing Education
Assignment 02

Visualising Data from Learning Tools

Due: Midnight, Thursday, 23$^{\text{rd}}$ April 2015

> *If the statistics are boring, then you've got the wrong numbers.*
>
> Edward Tufte[1]

1 Introduction

Digital technology and online tools are being used increasingly in the classroom. In many cases, these systems can collect fine-grained data about the way that students engage with the corresponding learning activities. In turn, this data can be analysed to help researchers and instructors understand *how* students are using the tools, and to evaluate their effectiveness.

> *The most exciting part of it is the data that we're gathering. Thousands of interactions per student per class. We can start analysing it, and what we learn from this is when the real revolution will come.*
>
> Peter Norvig[2]

While normally experiments are conducted to test a particular hypothesis, it is possible to mine existing data in an exploratory fashion to see if any interesting patterns or relationships exist. This should be done with some care however, as the more the data are investigated, the more likely it is that some interesting pattern will be observed by chance. Nonetheless, this approach can identify general questions of interest which can then be studied more formally with a controlled experiment.

Your task in this assignment is to explore possible relationships within a moderately large set of data collected from an online learning tool used in authentic classroom settings. Not all of the data provided to you may be relevant to what you are exploring. In addition, even within the sets that are relevant, you may want to exclude certain items. One of the challenges presented by this assignment is how to effectively manage, clean and process a significant amount of data, and how to clearly describe your methodology.

[1] The Visual Display of Quantitative Information, Graphics Press, 1983

[2] TED talk, "The 100,000 student classroom", 2012

2 Provided resources

You should begin by downloading the data set "CS747_PeerWise_Asst2.zip" (approx 100MB). This file is password protected and the password will be given in class. The archive contains data relating to courses that were created on PeerWise in the last year or so (specifically, since January 1st, 2014). This data includes:

- 6,942,083 submitted answers
- 488,333 authored questions
- 113,561 "game" scores for students
- 3,482 administrators/instructors

Details of each of these files will be covered in class. If there is any additional data you would like, please discuss this with Paul.

3 What you should do

Your goal is to produce one or more visualisations (for example, a graph or plot) of the provided data. Ideally, you should attempt to highlight some relationship or trend existing in the data. The motivation for this is that particularly interesting relationships may inspire future research. However such patterns can be hard to find and challenging to represent graphically – so have fun!

3.1 A trivial example

As a trivial example, you might expect that questions with a greater number of answer options (each question is permitted between 2 and 5 options) tend to be more difficult than questions with fewer answer options. To investigate this, you need to decide how to measure the *difficulty* of each question. You could do this by using the average student-assigned difficulty rating (students can rate each question they answer on an easy-medium-hard scale), or by using the proportion of submitted answers that were correct. In the latter case, you would need to decide how to determine the correct answer – would you rely solely on the question author's choice, or would you also take into account other student's responses to the question?

Either way, questions that have only been answered by a small number of students may be poor choices for analysis, because the difficulty measure may not be very meaningful in such cases. You might therefore choose to restrict your analysis to only questions that have received a minimum number of responses. Further, you may wish to restrict your analysis to only one course, in which case you would need to use the `course_id` information to filter the data set (each separate course has a unique ID). Finally, to visualise your findings, you might plot the average question difficulty against the number of answer options. Although this is a trivial example, it illustrates that there are several choices you will need to make when selecting data for inclusion in your analysis.

4 Potential questions to explore

You are free to explore any subset of the provided data you wish. Some questions might seem more interesting, but be harder to examine. You also may find it challenging to produce a visualisation that shows any clear trend or pattern – don't worry too much about this as long as you clearly describe the process you have taken to achieve your analysis.

The short list below includes a few examples of questions that might be interesting to explore. There are many other ideas – you are not restricted to only looking at questions on this list!

- How closely do student-selected difficulty ratings agree with more objective measures of difficulty (such as the proportion of responses that are correct)?
- Do longer questions (that require more reading) tend to be rated as more difficult?
- Do questions that have a more detailed explanation (as measured by the length of the explanation) tend to receive higher ratings for *quality*?
- How are question quality ratings distributed for courses of various sizes? In other words, what proportion of questions for a given course have quality ratings in various ranges (e.g. 0.0-1.0, 1.0-2.0, etc.)?
- Students can change their original answer, after receiving feedback on their selection, and can even *confirm* which answer they believe is correct. How often do these two things happen, and how much disagreement is there amongst students over the *initial* answers compared with the *confirmed* answers?
- Presumably, when a student chooses to change their original answer to something else (after they have received feedback on their first choice), they are more likely to change to the "correct" answer than to an incorrect one. Is this true?
- Do students who "confirm" their answers to questions more frequently tend to have more accurate initial answers?
- Students earn 10 "answer points" for answering correctly, and they receive a small penalty for a response which is likely incorrect (this penalty is roughly proportional to the number of answer options). What is the relationship between the total number of questions answered by a student and their "answer score"? What are the lowest "answer scores" amongst students who have answered a large number of questions?
- A student's "reputation score" increases over time as their contributions are validated by their peers. The earlier a student begins contributing to their course, generally the higher their "reputation score" will become. What is the relationship between when students make their first contribution to their course (or, possibly, how many distinct days they are active) and their final "reputation score" for that course?
- How do students answering patterns change over time? How are the questions that a student answers for a given course distributed over time? Are they all submitted on a single day, or over a longer period?
- For a given course, what proportion of questions are answered by all students and how does this vary by course size?
- For a given course, what proportion of students answer all available questions, and how does this vary by course size?
- Some students appear to answer questions very rapidly – what is the relationship between answer accuracy and the average amount of time that elapses for a given student between submitting subsequent answers?
- For a given course, are the most frequently answered questions the ones that have the highest quality ratings?

5 What to hand in

There are two deliverables:

1. *One or more visualisations* [format: pdf or image] — these charts or plots should be clear and neatly presented. They should include a title that summarises the question or relationship that you explored, and a key if necessary.

2. *Accompanying report* [format: pdf] — describe in detail how you selected and processed data for inclusion in your analysis. This should include a description of how you parsed the data, any software you used or scripts you developed to analyse the data, and any problems that you encountered during this process. You should also justify any decisions you have made regarding cleaning or filtering of the data, and what data you have specifically included in your analysis. You should conclude with a short statement that summarises your findings.

When you have finished, submit your visualisations and accompanying report via email directly to Paul (paul@cs.auckland.ac.nz).

6 Assessment

This assignment will be assessed using the following criteria:

- Presentation quality of your visualisation(s) [2.5 marks]
- Clarity of the question(s) chosen to investigate [1.5 marks]
- Challenge of the question(s) chosen to investigate [1.5 marks]
- Quality of accompanying report – description of motivation for your question(s), detail of your process/methodology for data cleaning and analysis, justification for inclusion/exclusion of data [2.5 marks]
- Appropriate approach and analysis for the question(s) investigated [1.5 marks]
- Bonus – any interesting findings [0.5 marks]

This assignment contributes 10% towards your final grade. If you have any questions about any part of this assignment, please contact Paul at any time either by email (paul@cs.auckland.ac.nz) or by phone (+64 9 373-7599 x 87087).

Have fun!