COMP6237 Coursework 2

Module:	*Data Mining*	**Lecturers:**	*MB and SH*
Assignment:	*Coursework 2*	**Weight:**	*20%*
Deadline:	*20/03/20*	**Feedback:**	*10/04/20*

Instructions

Overview

In this coursework your need to perform exploratory/descriptive data mining on a data set that we provide. You will need to write scripts to parse the data into a usable format, perform some kind of feature extraction and then apply standard techniques to explore relationships between data items, such as K-Means and Hierarchical Clustering, and data-analytic visualisation techniques like Multidimensional Scaling. Finally you need to put together a report that details your approach and your findings.

Details

The data you will be using for this assignment is a set of 24 texts about Antiquity (both classical and secondary literature); the original books have been scanned and run through an Optical Character Recognition system to produce an HTML document for each page. The scans and OCR data were produced by Google as part of the Google Books Library Project https://en.wikipedia.org/wiki/Google_Books_Library_Project.

You can download the Zip file containing the HTML pages with the OCR results here https://secure.ecs.soton.ac.uk/notes/comp6237/data/gap-html.zip. Inside the zip file, there are 24 folders representing the 24 texts, with each page represented by the sequentially numbered HTML files. The original scanned images are not included here due to their size (around 4GB), however, you can browse the original scans here https://secure.ecs.soton.ac.uk/notes/comp6237/data/gap-images/ if you wish.

The aim of this coursework is for you to explore how these 24 texts are related by applying appropriate data mining techniques. You'll need to create software to extract the contents of the HTML files and build some form of feature representation to which you can apply standard descriptive data mining techniques. At a minimum, we expect you to experiment with Hierarchical Clustering and Multi-Dimensional Scaling, however you should also explore other approaches.

Deliverable

You need to produce a concise 2-page “working notes” paper (see http://ceur-ws.org/Vol-1043/ for examples of standard academic working notes papers) using the standard 2017 ACM conference proceedings style (use the sigconf style option for the template). The two page limit on the paper is final; no additional pages or appendices are permitted. We are expecting the paper to illustrate (with pictures as appropriate) what you have done and also demonstrate your ability to interpret what the data mining techniques are showing.

Marking and Feedback

Full details of the marking scheme are given below:

Learning Outcomes

Solve real-word data-mining, data-indexing and information extraction tasks
Demonstrate knowledge and understanding of:
Key concepts, tools and approaches for data mining on complex unstructured data sets
Theoretical concepts and the motivations behind different data-mining approaches

Mark Scheme

Good working notes papers not only effectively apply techniques and describe results, but also offer critical insight into the findings of the analysis in the context of the underlying data. In particular you need to demonstrate that you understand the data, and, in the context of that understanding, that you can rationalise and reflect on why the analytic techniques are giving the results they do. The working notes paper will be marked using the following criteria:

Criterion	Description	Marks
Experimentation	Analyse the problem and define suitable preprocessing and feature extraction operations	28
Application of techniques	Show ability to apply exploratory data mining techniques	28
Analysis	Reflection on what can be understood from the data through the application of exploratory techniques	28
Reporting	Clear and professional reporting	16

Standard ECS late submission penalties apply.

Written individual feedback will be given covering the above points, and will be emailed out once marking is complete. We’ll also use one of the lecture slots for a further group feedback as well as a discussion about the data and the analysis.

Tools

You can use any available existing tools, programming environments and software libraries for this coursework. It is however important that you include full details in your report - this must include details about which specific variant of the standard techniques are being used, with references as appropriate. Also include any details of the implementation doing something non-standard (for example making approximations in the sake of efficiency), and all parameters.