COMP6237 Data Mining

Coursework 2: Understanding Data

Maintained by Dr Jonathon Hare.

COMP6237 Data Mining

Coursework 2: Understanding Data

Brief

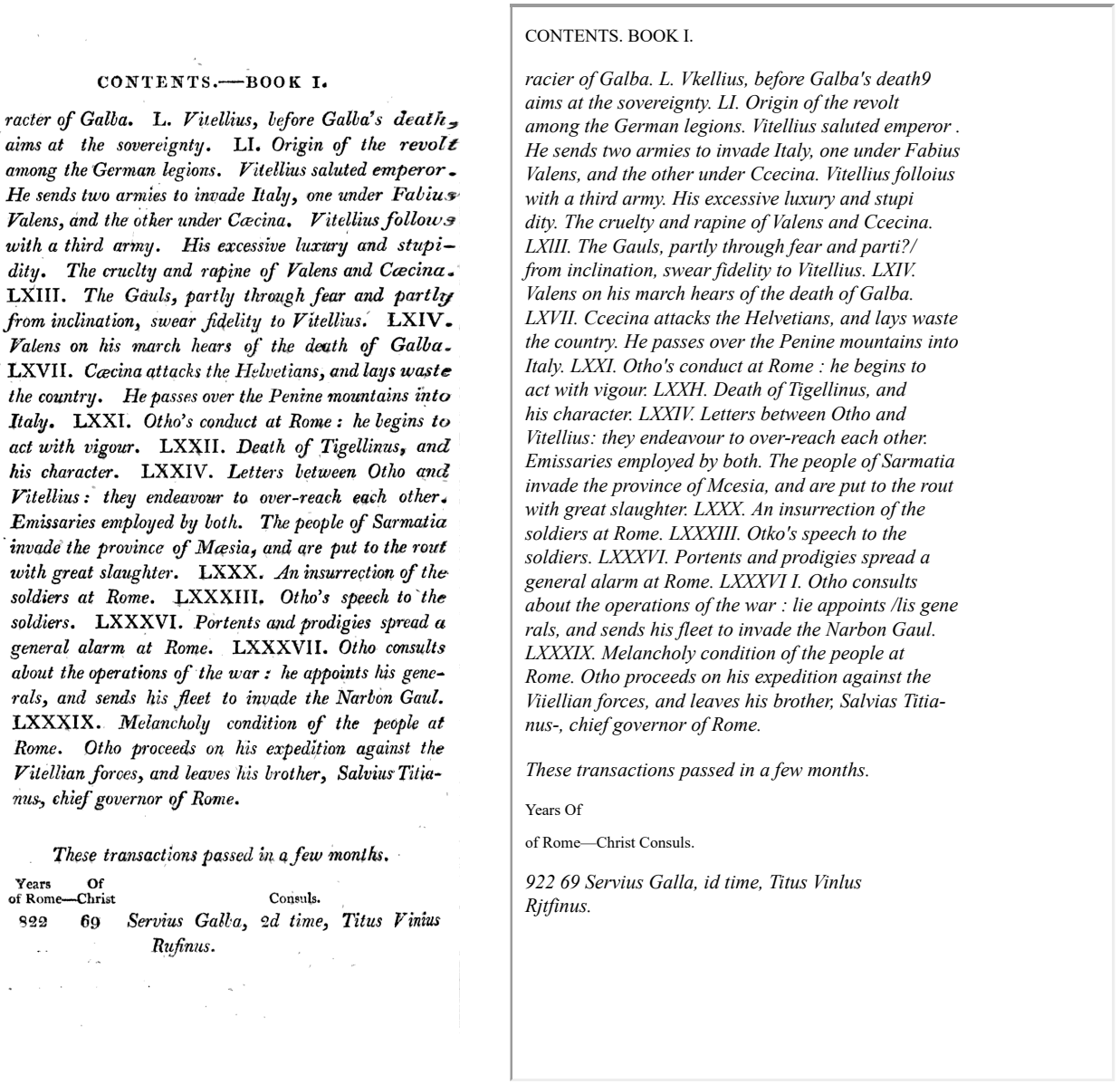
Due date: Thursday 17th March 2016, 16:00.
Handin: 1516/COMP6237/3/
Required files: report.pdf
Data: gap-html.zip
Credit: 20% of overall module mark

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. A typical scanned page and its OCR result is shown below:



CONTENTS.—BOOK I.

racter of Galba. L. *Vitellius, before Galba's death, aims at the sovereignty.* LI. *Origin of the revolt among the German legions. Vitellius saluted emperor. He sends two armies to invade Italy, one under Fabius Valens, and the other under Cæcina. Vitellius follows with a third army. His excessive luxury and stupidity. The cruelty and rapine of Valens and Cæcina.* LXIII. *The Gauls, partly through fear and partly from inclination, swear fidelity to Vitellius.* LXIV. *Valens on his march hears of the death of Galba.* LXVII. *Cæcina attacks the Helvetians, and lays waste the country. He passes over the Penine mountains into Italy.* LXXI. *Otho's conduct at Rome: he begins to act with vigour.* LXXII. *Death of Tigellinus, and his character.* LXXIV. *Letters between Otho and Vitellius: they endeavour to over-reach each other. Emissaries employed by both. The people of Sarmatia invade the province of Mœsia, and are put to the rout with great slaughter.* LXXX. *An insurrection of the soldiers at Rome.* LXXXIII. *Otho's speech to the soldiers.* LXXXVI. *Portents and prodigies spread a general alarm at Rome.* LXXXVII. *Otho consults about the operations of the war: he appoints his generals, and sends his fleet to invade the Narbon Gaul.* LXXXIX. *Melancholy condition of the people at Rome. Otho proceeds on his expedition against the Vitellian forces, and leaves his brother, Salvius Titianus, chief governor of Rome.*

These transactions passed in a few months.

Years of Rome	Of Christ	Consuls.
822	69	*Servius Galba, 2d time, Titus Vinius Rufinus.*

CONTENTS. BOOK I.

racier of Galba. L. Vkellius, before Galba's death9 aims at the sovereignty. LI. Origin of the revolt among the German legions. Vitellius saluted emperor . He sends two armies to invade Italy, one under Fabius Valens, and the other under Ccecina. Vitellius folloius with a third army. His excessive luxury and stupi dity. The cruelty and rapine of Valens and Ccecina. LXIII. The Gauls, partly through fear and parti?/ from inclination, swear fidelity to Vitellius. LXIV. Valens on his march hears of the death of Galba. LXVII. Ccecina attacks the Helvetians, and lays waste the country. He passes over the Penine mountains into Italy. LXXI. Otho's conduct at Rome : he begins to act with vigour. LXXH. Death of Tigellinus, and his character. LXXIV. Letters between Otho and Vitellius: they endeavour to over-reach each other. Emissaries employed by both. The people of Sarmatia invade the province of Mcesia, and are put to the rout with great slaughter. LXXX. An insurrection of the soldiers at Rome. LXXXIII. Otko's speech to the soldiers. LXXXVI. Portents and prodigies spread a general alarm at Rome. LXXXVI I. Otho consults about the operations of the war : lie appoints /lis gene rals, and sends his fleet to invade the Narbon Gaul. LXXXIX. Melancholy condition of the people at Rome. Otho proceeds on his expedition against the Viiellian forces, and leaves his brother, Salvias Titia- nus-, chief governor of Rome.

These transactions passed in a few months.

Years Of

of Rome—Christ Consuls.

922 69 Servius Galla, id time, Titus Vinlus Rjtfinus.

Book scanning example.

You can download the a Zip file containing the HTML pages with the OCR results here. Inside the zip file, their are 24 folders representing the 24 texts, with each page represented by the sequentially numbered HTML files. We haven't included the original scanned images due to their size (around 4GB), however, you can browse the original scans here 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're expecting you to experiment with Hierarchical Clustering and Multi-Dimensional Scaling, however you might 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 ACM proceedings style. The two page limit on the paper is final; no additional pages or appendices are permitted. We're expecting the paper to illustrate 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

Criterion	Description	Marks
»	Theoretical concepts and the motivations behind different data-mining approaches	

Mark Scheme

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.

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 specific 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.

Questions

If you have any problems/questions then email or speak to Jon in his office or after the lectures.