Project 3
Data Mining & Visualization

In the 3rd project you will apply latent semantic analysis to determine the similarity between documents and visualize the result using force directed placement.

You can select one or several of the subsequent datasets:
- IU Course Descriptions: http://www.indiana.edu/~deanfac/blfal01/
  (and other semesters)
- Grade Distribution Reports: http://wwwreg.indiana.edu/Gradedistribution/
Or very similar data sets – please check with me!

How to proceed

• Select a user group and a data set this group is interested in viewing.
• Do an analysis of your intended user group & their tasks.
• Decide how to parse & clean the data. What additional data is required to serve your users and their tasks?
• In this project you have to apply LSA for data analysis and FDP to layout data objects.
• Create a web page that shows a sketch of the final visualization, discusses the results of user and task analysis as well as visualization design.
• You can work in teams of two or by yourself.

When you mine your data:
• Try different parser & filter to generate the optimal input for LSA & visualization.
• Run LSA with different parameter settings and watch the effects.
• Visualize the similarity matrix using the Similarity Matrix Visualization Java applet accessible at http://ella.slis.indiana.edu/~katy/L697/code/simvis.html

Explore different data filtering, analysis, and visualization algorithms. Try to maximize:
• Space Efficiency - are they good at utilizing screen area?
• Information Abstraction - do they abstract information and avoid information overload?
• Simplicity - are they based on simple, elegant ideas which are easy to implement?
• Navigation – is it intuitive and efficient?
Try out what works best!

In Lab 7 – next week

Present your ideas and first sketches/visualizations in Lab 7. Feel free to hand in draft web pages of your approach for additional feedback.
Final Handin

The final project handin is due 03/05/02 and comprises a web page that contains/links to the
• User and task analysis results.
• Input data set (delimited text, EXCEL or MS Access file is fine).
• Detailed explanation of data mining, IV sketch, & justifications of design decisions for IV. (Problems & strengths of your visualization & interaction techniques. Desirable modifications & extensions.)
• Snapshot of the final visualizations.
• The xml version of the data sets you used and any code you wrote/modified.
• Discussion of the IV & conclusions (In how far does the IV match your users and their tasks? Which psychological features are met? Discuss complexity & scaling issues.)
• References - complete citations!

Handin one copy per team via http://ella.slis.indiana.edu/~mukkama/login97.html

Project 3 is graded according to

• Quality of user and task analysis.
• Appropriate selection of data set, parsing & filtering data algorithm, and visualization design that match your users and their tasks.
• Discussion of data mining, visualization & conclusions (discuss all the problems that could not be resolved in 2 weeks time – there should be many!).
• The quality of the content, including the accuracy and completeness of information, the expressiveness and clarity in communication of ideas, and the appropriateness of the attribution(s) for the work of others.
• Overall quality of the project (organization, implementation, presentation, references).