CSCI-4502 Data Mining Project Description

Objectives

- Work in a group of size 2-4 students
- Go through the full data mining process: defining interesting questions, data collection, data preprocessing, data mining technique.

You will work in a group to answer interesting questions by data mining a large dataset.

Pick your own project idea - this is the exciting part!
Discuss ideas with the instructor and other students.

Help each other - if you find a good tool to use let the rest of us know! If you know of a good data set another group could integrate for their project - let them know!

Use tools that are available. This class is *not* about reinventing the wheel. But you still want to understand how the tools work so you can understand the results.

The Project is divided up into 7 parts.

Part 1 - Team & Topic Slides
Part 2 - Proposal Paper
Part 3 - Progress Report
Part 4 - Project Final Report
Part 5 - Project Code and Descriptions
Part 6 - Project Presentation
Part 7 - Peer Evaluation & Interview Question

Remember, you all have access to your group's GitHub account.
So if it is submitted late, it affects the grade of everyone in the group. Work together!

Part 1 .

Team & Topic & Dataset

One of the most important parts to this project is to make sure you have a dataset. While there are many very interesting questions you could ask, if you cannot access any data to mine for those answers then you will not be able to use those questions for your project! Make sure you come up with your interesting questions along with ensuring you have access to a dataset that can answer your question. Download the dataset to be sure, t*his has been a problem in the past!*

Size of data set - hopefully millions of data points. A few thousand is too small.
But always start by testing with a small data set to test and get going (medium-sized sample) so you can see the results within half an hour. Make sure getting code correct before testing takes a long time.

For Part 1 you will create a set of slides that you will present in class and that will be saved as a .PDF for the Part 1 submission with the following information (be brief):

- **Title:** Descriptive title of your project
- **Team members**: first and last name of each member
- **Description:** 2-3 sentence paragraph project description - *what interesting questions do you intend to answer?*
- **Prior Work:** What prior work has been done on your idea
- **Datasets:**
  - List of datasets to use
  - Where found (URL and who is supplying the data, e.g., NASA)
  - Whether it you have it downloaded (on who's machine)
- **Proposed work**: what do you need to do?
  - Data cleaning:
  - Data preprocessing:
  - Data integration:
  - etc.
- **List of tool(s)** you intend to use
- **Evaluation:** How you can evaluate your results

Note: There are no right or wrong questions, only interesting questions.

Submission Requirements:

- Create a group github account.
- Get everyone in the group set up as a contributor.
- Submit to your group github account a PDF named **Group#_ProjectTitle_Part1.PDF** (E.g., **03_MiningForGoldInStocks_Part1.PDF**). *Do not put this in any subdirectory.*

Part 2 .

Project Proposal Paper

For part 2 you will write up a proposal following the ACM SIG paper format (ACM_SigConf.pdf but use 11 point font. Keep 1.1 line spacing and other formats).
https://www.acm.org/publications/proceedings-template

Your project proposal should be approximately 3 pages using the template and contain the following main sections:

- **Problem Statement/motivation** (like homework 1, what knowledge and how would you apply that knowledge, what is interesting that you hope to find)
- **Literature survey** (previous work) describe and cite.
- **Proposed Work**
    - E.g., what do you need to do for data collection, preprocessing (cleaning, integrating, transforming, etc.), process for derived data, design, evaluation…
    - Describe how it is different than what has been done previously from your literature survey (or if replicating).
- **Data set** (make sure you *have* the data set!). Provide URL and details about the data set (similar to homework 1, chapter 2, etc.)
- **Evaluation Methods**
    - E.g., metrics, existing solutions, …
- **Tools**
- **Milestones** What you plan to have done by when

Submission Requirements:

- Submit to your group github account a PDF named **Group#_ProjectTitle_Part2.PDF** (E.g., **03_MiningForGoldInStocks_Part2.PDF)**. *Do not put this in any subdirectory.*

Part 3 .

Project Progress Report

Your project progress report should be approximately 6 pages using the 2-column ACM SIG template and modification to your part 2 proposal. Add the following:

- Updated, extended version of initial proposal,
    - Proposal review: motivation, proposed work, tools, evaluation, milestones
- **Milestones Completed:** What you have achieved so far (in milestones section as a subtopic)
- **Milestones Todo:** What remains to be done (in milestones section as a subtopic)
- **Results so far** (new topic at end)
    - Any graphs, correlations, etc. if any

Submission Requirements:

- Submit to your group github account a PDF named **Group#_ProjectTitle_Part3.PDF** (E.g., **03_MiningForGoldInStocks_Part3.PDF**). *Do not put this in any subdirectory.*

Part 4

Project Final Report

Your project final report should be approximately 10-12 pages using the template and modification to your part 3 report. Include the following:

- **Abstract**
  - What interesting question(s) were you seeking to answer?
  - What is a brief summary of your results?
- **Introduction**
  - Description of your question(s)
  - Why they are important
- **Related Work**
- **Data Set**
  - Where from
  - Attribute features
  - etc.
- **Main Techniques Applied**
  - Data clean/preprocess/etc.
  - Data Warehouse/cube/etc.
  - Classification/Clustering/etc.
- **Key Results**
  - What did you discover/learn?
- **Applications**
  - How can the knowledge gained be applied?
- [Optional Extra Credit] **Visualization**
  - Visualization of results (and/or link to interactive site to view or link to a video demonstrating the interactive site)
    - *Note: not just a bunch of graphs, but a way to depict the interesting knowledge in a useful/meaningful way.*
    - Can use d3js.org, processingjs.org, or other such library (anything really)

Submissiøn Requirements:

- Submit to your group github account a PDF named **Group#_ProjectTitle_Part4.PDF** (E.g., **03_MiningForGoldInStocks_Part4.PDF**). *Do not put this in any subdirectory.*

Part 5 .

Project Code and Descriptions

All of your source code used for the project should be submitted to your group github repository. Include the following:

- All source code in github
- Create a README to display on your github main page with:
    - Project title
    - Team members
    - Description of the project
    - Summary of the question(s) sought and the answers
    - Application of this knowledge
    - Link to the video demonstration
    - Link to your final project paper

Part 6 .

Project Presentation

Present your project to the class.
Your presentation should include:

- Project title
- Team members
- Question(s) sought to answer
- Data preparation work
- Tools used
- Classification/clustering/etc applied,
- Knowledge gained
- How that knowledge can be applied.

Presentation tips:

- Use 20-point font minimum
- *Pictures say a thousand words* much easier to convey info than lots of words on each slide
- Make sure we can read your content.
- Everyone must speak during presentation.
- You have only 6 minutes to present.

You will need to do the following:

- Submit a video to Github labeled **Group#_Project_Title_Part6_Video.[extension]** discussing the topics listed above.
- Submit your slide deck to Github labeled **Group#_ProjectTitle_Part6.PDF**

Part 7 .

Peer Evaluation & Interview Question

Each person individually fills out the peer evaluation form (see Moodle) and submits into Moodle as an Excel spreadsheet (do not convert to .PDF). This will be incorporated into the individual project grade for groups that have someone not helping much at all or someone who put considerable extra effort helping others and/or developing the project as a good team player. However, taking over the project and not being a good/helpful team player is bad!

Please note: Giving everyone the same number does not count as filling out this form!

You will need to fill in the comments to support your evaluation, as well as answer the Team Dynamics question. For Team Dynamics, enter the letter a, b, or c for “Which one” to specify which question you are answering.

Submit to **Moodle** as:
Group#_YourLastName_YourFirstName_PeerEval.xls e.g., 03_Boese_Elle_PeerEval.xls