CMPUT 697 Project Proposal Resources

What you can do for a CMPUT 697 project is very flexible. You can adapt/define/apply a method for data mining and use it to study a data set in a domain of your interest, or you could try to develop and evaluate a "new" idea for a method or modification of a method, or a clever combination of methods that you come up with. The methods chosen can be supervised, unsupervised or semi-supervised, and they are not limited to what we have discussed in class.

Individual projects or group projects with groups of two students are allowed. Group projects are encouraged. For group projects, all participants will receive the same mark, unless a deviation is agreed on by all group members.

Project Related Tasks

A. You first need to register a course project for yourself or a group via a spreadsheet (link also available directly on eClass). To register a project, you need to enter the following information:
   - A title for the project
   - One to two sentences describing your intention
   - Name of participants

B. After deciding on a topic, you(r group) will need to submit a more detailed one-page course project proposal (worth 5% of your final grade) via an upload link that will be available on eClass (before a given deadline). When you set your goals, keep in mind that the project must be finished before the semester ends, which means that you will only have about one and a half months to work on the project. The proposal should outline your project idea and goals in more detail. It must include:
   - A short description of the overall objective/problem/task
   - The steps and subtasks that you can anticipate and through which you plan to achieve the goal (which typically includes finding and reading papers related to your task, planning how you are going to use/adapt/integrate/improve/ methods and prepare/pre-process data sets, implement and test your ideas, etc.).
   - A plan on how you intend to evaluate your solution (e.g. - what are you going to measure and what experiments are you going to conduct?)
   - A timeline for the project, detailing when you expect to do each subtask.

C. You(r group) will have to schedule a project demo with me, for which there will be a sign-up sheet posted on eClass. The demo should include a short presentation of the project showing what was achieved as well as a demonstration of your implementation and how you run it. (worth 15% of your final grade).

D. You will need to write and upload a detailed project report (worth 20% of your final grade).

Below are pointers to some data sets and associated research/analysis questions that you could take as a starting point to elaborate into a more specific course project. However, these are just examples. Try to find and use your own data set/project.

1 - NSERC Dataset

The government of Canada makes many datasets about its finances available. Among these datasets is the NSERC awards dataset composed of awards that were funded between 1991 and 2017 by NSERC. The goal of projects with this data would be to collect the NSERC dataset, pre-process it, and use data mining methods to extract insights from it. To answer perhaps some of the following questions:

- Is there an awarded research topic that gets more money than other topics?
- What are the award distributions with respect to:
    - Institutions
    - Provinces
    - Gender
    - Language

  and are there any interesting relationships between them?
- Are there meaningful clusters, outliers?
- What are the topics that are more frequent in a particular year?
- Is there a relationship between topics from year to year?
- ...

Dataset available at:
https://open.canada.ca/data/en/dataset/c1b0f627-8c29-427c-ab73-33968ad9176e

2 - NYC Taxi Dataset

The city of New York makes available data about taxis and For Hire Vehicles (Uber, Lyft, etc.). This data comprises trips that occurred from 2009 to 2019, including start location, end location, cost, area, among others. This is a huge dataset (more than 1 billion trips) and you might have to work with a subset of it. Some interesting questions that could be answered in a course project using this data set include:

- What are the most popular parts of the city given a time of the day?
- Taxis vs Uber
    - Is there an area where Uber overcame taxis?
    - Does that change over time?
    - What are the most popular times for either services?
- What are "hotspot" areas of pickup and dropoff? Do they change over time?
- Can we identify any privacy issues in this dataset?

Dataset available at: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page