Fall 2023	**CSC 466:** Knowledge Discovery from Data	Alexander Dekhtyar

Analytical Project

Overview

The **analytical course project** is viewed as an equivalent to a final exam. The project is to be performed in teams of three to four people. Due to constraints on the presentation time, we cannot have teams smaller than three-person teams. Team formation is left up to you. Please note, that the amount of work done for the project must be commensurate with the size of the team.

Due Dates:

What	Date
Proposal	November 8 (Wednesday)
Presentation	TBD
Report	December 13 (Wednesday)

Assignment

This quarter's analytical project is fairly straightforward. Each team is asked to do the following:

1. Find one or more datasets of interest.
2. Ask analytical questions about each dataset.
3. Translate the analytical questions into applications of the methods, algorithms, and techniques studied in the class.
4. Prepare and submit a brief proposal outlining your intended project.
5. Apply the appropriate methods, algorithms, and techniques to the data.
6. Collect results, visualize and explain them.
7. Prepare and submit a report describing your findings.
8. Present your findings during the finals exam time slot.

Step One: Finding Datasets

Each team needs to determine what data it wants to analyze. A number of options are possible here.

Use of existing datasets. You may use any existing dataset that is (a) publically available, or (b) can be legally procured by the team members. The only condition is that the dataset has not been used in the course (if you are not sure whether a specific dataset will be used in the outstanding labs in the course, feel free to consult me). The course web page will have links to some dataset repositories that may give you an idea where to start your search for datasets.

Generation of own datasets. Your team may decide to build a dataset of its own. This can be done in a number of different ways. For example, a team may choose to scrape content of a web site[1], or build a database based on some observed information. Some teams in the past built their own datasets via observations of a physical world[2], or via surveys or other ways of engaging human respondents [3].

A number of existing interesting datasets are fairly easy to obtain. For example, DBLP, a collection of Computer Science bibliographic records is available as an XML document from the DBLP site. A number of Machine Learning data repositories have interesting datasets, both large and small that may be of interest to you. Large text collections exist as well: Wikipedia is one of those - it is downloadable in its entirety. Enron emails collection is also publicaly available. US Census Bureau has a wide range of demographic information available about the US, that can lead to some interesting analytics. Kaggle is an emerging repository of diverse data collections.

Steps 2,3,5: Analytical Questions and Analytical Methods

It is expected that the analytical questions you ask involve use of the KDD methodolgy discussed in the course.

As part of your solution approaches you can conduct any statistical analyses of the data you seem fit, as well as any KDD tasks discussed in the course, or discovered by you independently.

The ground rules for what you can and cannot do are set below.

Allowed Activities

As part of your preparatory and analytical activities you are allowed to do the following:

- Use any programs you (members of the team) created during this course.
- Use any programs other students (outside of your team) created during this course, **with the explicit permission of the authors of the programs**.
- Use any existing code for "menial" tasks (parsing data, reporting) as well as for tasks such as visualization of output. You **must be allowed to use the code by the licensing agreement of the code**.
- Use any existing code for KDD methods both covered and not covered in class, subject to the following two conditions:
  1. You must be allowed to use the code by the licensing agreement of the code.
  2. You **must gain sufficient understanding of the methodology implemented by the code**. For example, if you decide to use some open source software for learning neural networks from data, I will expect at least one member of the team to be able to coherently explain to me what neural networks are, and what specific types of networks are being constructed by the software used.
- Study new (not covered in class) methods for solving KDD problems discussed in class.
- Study new (not covered in class) KDD problems and methods for addressing them.
- Write new code.
- Enhance code created earlier during this course.

[1] You must not violate the site's usage policies when doing so, though.

[2] One team staked out a local donut shop and collected information on the donut orders.

[3] One team ran a survey asking students to specify their preferences of operating systems and collected around 400 responses.

- Use any supporting architectural solutions (e.g., MySQL DBMS, or math/stats packages like R or MatLab) and use any analytical and KDD techniques available through them, subject to the same condition:

  You **must gain sufficient understanding of the methodology being used**.

Disallowed Activities

The following is a list of **no-no**s for this project. Any of the activities below conducted as part of the project **are considered equivalent to academic cheating!**

You may not:

- Use ANY code you have not been authorized to use (by the authors, or by the licensing agreements).
- Use ANY KDD/analytical techniques (or their implementations), when you did not gain sufficient understanding of the technique.
- Actively seek, and peruse information about the datasets, that contains the answers to your analytical questions.

  Note: some of the datasets you may wind up using are well-known data mining/machine learning datasets, which have been used by many different research teams to test their methods. KDD models developed for such datasets may be discoverable via some targeted web search.

  Note: Some of the datasets are featured in multiple publications. Typically, it is safe to peruse such publications in your work on the project. If a paper publishes, in addition to the evaluation results, the actual models built by the KDD methods for the dataset, you are still allowed to use the paper on the following two conditions:

  - You explicitly acknowledge the source of the model.
  - If the model addresses your analytical questions, you still use tools available to you to generate it.

  (I do not want this assignment to turn into a hunt for existing models. I want you to build your own.)
- Solicit help with your analysis from anyone outside of this class. (In particular, do not ask dataset owners or researchers who used the dataset in their work for help.) If you believe you need to get in touch with the data owners/other researchers because you have a bona fide question or concern, **bring your question(s)/concern(s) to me**, and let me initiate the contact. (this, among other things, will increase the probability and timeliness of the response).

What about LLMs?

Your project can involve the study of Large Language Models/Large Generative Models, including those like ChatGPT or Bard or Dall-E available publicly. However, please be aware that your project cannot solely rely on artifacts produced by the LLMs and the like.

That is - if you want to compare the behavior of an LLM to a classifier you have trained yourselves - this is allowed. If you need to build a codebase to analyze the outputs of an LLM - this is allowed. But if your entire project relies solely on the work done by an LLM, without further work on your part - this will not be allowed.

As usual, when in doubt - please consult the instructor.

Step 4: Proposal

By **November 8** each team will submit a short *project proposal*. The project proposal shall include the following information:

- Names of all team members.
- Dataset information: describe the dataset/datasets you plan on using for the project.
- Questions to study. Outline the analytical questions you want to work on as part of the project.

The proposal shall be word-porcessed and submitted as a PDF document. It should be relatively short.

You will receive responses to your proposal by **November 9**. If I approve your proposal fully, you can start working on the project. If I have comments/corrections to your proposal, you shall take those into account when proceeding with your work. We will use November 9 lab period for discussing proposals with individual teams (as needed).

Submit your proposal using handin:

```
$ handin dekhtyar 466-proposal <proposal.pdf>
```

Please note, the deadline. I will print all proposal by the end of the work day on November 9.

Step 7: Report

Each team shall submit report of all the findings. The report shall be typeset, written in a word-processing software (Word, or Word analogs, or LaTeX), be submitted in PDF format, and be formatted as an academic paper/technical report.

The report shall have a title, include a list of authors, a short abstract, an introduction section in which you discuss the overall approach the team took to the assignment, multiple sections describing the datasets you used, the questions you asked, the methods you deployed and the results you observed. Finally, your report shall have a conclusions section in which you summarize your team's experiences with analysis of data.

Step 8: Presentation

The project involves delivering a 15-20 minute presentation describing your project to the entire class. Typically we do this during the finals week. As our final exam is scheduled for a Friday afternoon, this creates a challenge. We will discuss the logistics of setting up presentations some time in class in the next week, and the instuctions for preparing the presentations, and the timeline will be published at that time.

Deliverables and Submission

Each team shall produce the following artifacts.

- A written report (see above).
- PDF of your presentation
- Code and instructions.

Submit all information using handin as follows:

```
$ handin dekhtyar 466-project <files>
```

GOOD LUCK!