Spring 2018	**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 four people (due to changes in the class roster, one team of three people is allowed). 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 Date: ***June 14, 2018***.

See full set of deadlines below.

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.

In addition, it is possible that teams will be asked to present their results during the finals week (the information about whether we will be doing oral presentations will be provided to you some time in November).

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.

The rule-of-thumb for this assignment is "one dataset per person", although in some cases, if you are building or using very large datasets and are asking multiple analytical questions of them, this requirment can be reduced. This will be finalized in your proposal.

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.

[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.

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.
- 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).

Extra Credit

You get 10-15% extra credit for each dataset **created by you** that you submit to the instructor. In order to submit the dataset:

- You must have the rights to it. That is, you must be the creator/designer/owner (whichever is more appropriate) of the collected data.

- You have to give the instructor the (non-exclusive) rights to use the dataset for research and instructional activities. This can be accomplished by including with your dataset submission a README file that explicitly grants such permission.

The size of the extra credit depends on the size and the versatility of the dataset, where "versatility" is understood as appropriateness of the dataset for a variety of different analytical tasks (going beyond those covered in your project).

Step 4: Proposal

By **May 14** 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.

You will receive responses to your proposal by **May 18** (in fact, I will probably send written comments to each proposal on May 18 or May 20). If I approve your proposal fully, you can start working on the project. If I can comments/corrections to your proposal, you shall take those into account when proceeding with your work.

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 May 14.

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

On June 14, during our final exam time, each team will in turn have 20 minutes to present their project to the rest of the class.

Deliverables and Submission

Each team shall produce the following artifacts.

- A written report for each dataset selected by the team. The report shall, at the very least, contain the following:
  - Description of the dataset used. If this one of datasets provided by the instructor, you need to simply name it. If this is a dataset you creates/selected, provide a full description of the dataset.
  - Description of the analytical question(s) your team studied.
  - A narrative explaining which analytical methods your team used.
  - Results of the use of the methods visualized where possible.
  - Conclusions you team drew from the results.
- Any datasets you would like to submit.

Submit all information using handin as follows:

```
$ handin dekhtyar project01-466 <files>
```

GOOD LUCK!