| Spring 2016 | **DATA 301:** Introduction to Data Science | Alexander Dekhtyar |
|---|---|---|

Analytical Project

## Overview

The **analytical course project** is the final deliverable for the class. The project is to be performed in teams of two–four people. 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 - I will expect to see a bit more done by a team of four people, than by a team of two people.

**Due Date:** ***June 8, noon***.

## Assignment

The 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 a data science process for collecting, cleaning, modeling and analyzing the data using the methods, algorithms, and techniques studied in the class.
4. Follow the steps of the data science process to analyze the data.
5. Collect results, visualize and explain them.
6. Prepare and submit a report describing your findings.

In addition, it is possible that teams will be asked to present their results during the finals week. It seems unlikely, but I will leave the final decision until the final week of classes.

## 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 the physical world, or via surveys or other ways of engaging human respondents.

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.

## Steps 2-4: Analytical Questions and Analytical Methods

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

As part of your solution you can choose to collect and clean your data in any way you see fit. You may integrate multiple datasets[2] to obtain your eventual data. You may choose any convenient way to store and model data.

Last, but not least, you can conduct any statistical analyses of the data you seem fit, as well as any machine learning tasks discussed in the course, or discovered by you independently.

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

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

[2] For example you can integrate a weather dataset with a US census dataset to find out if there is a relationship between climate/weather and population density

**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 functionality from NumPy and MatPlotLib.
- Use any existing code for statistical analysis/ machine learning/data mining/information retrieval/collaborative filtering 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.
- The latter part covers your ability to use SciKit-Learn and NLTK functions. Both packages have a lot of different learning algorithms implemented (some are covered later in CSC 466, STAT 419, DATA 401, CSC 582 and some other courses). If you choose to use any of the functions to analyze data, **you must take your time to study what these functions are actually doing**. I can help with appropriate references.
- Study new (not covered in class) methods for solving machine learning problems discussed in class.
- Study new (not covered in class) data analytical 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, Jupyter) and use any analytical and machine learning/statistical 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 data 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 of 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).

### Step 5: 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.

## 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 code/Jupyter notebooks visualizing the work performed by the team to collect, clean, model, analyze and visualize the data.
- Any datasets collected in the process.

Submit all information using handin **on the CSL machines** as follows:

```
$ handin dekhtyar project-301 <files>
```

In addition, if you have Jupyter noteboks to submit, submit the notebooks, and all necessary code/data via `nbgrader` on the Jupyter server as follows

```
$ nbgrader submit --course=dekhtyar Project
```

(you may want to do `nbgrader fetch --course=dekhtyar Project` first to create the project directory, then copy all your files into it and submit).

**GOOD LUCK!**