15-388/688: Final Project

October 11, 2016

Important Due Dates (due 11:59pm, except for video presentation)

- **10/24** – Proposal (200 words max) due
- **11/11** – Project midterm report due
- **12/9** – Final project report due
- **12/14** (time TBD) – Final project video presentations

Introduction and requirements

A major component of 15-388/688 is a final project that allows you to investigate some applied data science problem in more detail. While we are very open regarding the topic and type of analysis that you do in the project, we require the following:

1. The focus of the project should be on analyzing a data set to ask some underlying question about the data itself. While you can use advanced algorithms to help you answer this question, the focus on the final project should not be solely on the algorithms themselves, but should be grounded in some practical question you want to understand from the data itself.

2. Owing to the above, the project must analyze a *real* data set. You cannot generate a purely synthetic data set for the project and then analyze that. To be clear, though, you *can* still collect your data from some computational process ("real" does not mean that it has to be generated by physical, non-computational systems). If you have any doubts about this point and your topic, talk to the instructors.

3. You cannot use a pre-curated data set, even if it is "real" data. This would include data sets, for instance, that are released as part of a Kaggle contest or something similar (unless you *substantially* build upon the data as it was originally released). The goal of the project is to gain experience with the full data science pipeline, and using a pre-developed data set (often with pre-specified metrics), removes a great deal of the challenge.

4. You *can* pick topics that overlap with your existing areas of research.

Groups

The final project should be done in groups of 2-3 students. In rare cases we will approve students who want to work on a project alone, but there must be a well-founded reason for this beyond simply not being able to find a group to work on a particular topic of interest: for example, one student mentioned working with data covered by federal privacy regularization, which they had access to through existing research; this would be a reasonable cause for an exception. We won't approve any groups of four students (people in this scenario are encouraged to simply split into two groups, even if they end up working on two very similar projects).

You should have your group set by the date that project proposals are submitted (each group will submit only one proposal, listing all members of the group). There can still be changes if they are absolutely necessary after this point, but every effort should be made to have the final group assignments by the time that you submit the proposal.

Final report - due 12/9

Like the tutorial, the final report will be presented in the form of a Jupyter notebook. You can upload a full package with all the data (and any additional code, for instance if you want to include some code in a separate Python file for convenience), **but your Juptyer notebook should be readable as a narrative explaining your project *without* requiring the reader to run any code.** In particular, the instructors will *not* be running your code when they grade the project, just reading through the notebook itself (in fact, we will typically be viewing them in the `nbviewer` app, which only displays a static rendering of the notebook. Thus, you should think of the integration between code, figures, and text really just as a means to make concrete the algorithmic aspects you are describing in your report. The notebook is still very much a "traditional" report, just one that mixes together prose, code, and figures, and which *can* be run directly by anyone who is interested in diving more deeply into your work.

See the example tutorial we have released on GIS systems (`https://nbviewer.jupyter.org/url/www.datasciencecourse.org/GIS%20Tutorial.ipynb`) for an example of this type of report. Note that even if you don't run any of the cells in the notebook, you can still follow the discussion about how the various libraries work. Since this example is targeted towards the tutorial assignment and not the project assignment, it's naturally much more focused on explaining certain operations, installing libraries, etc (your final project does not need to discuss any of these), but hopefully it illustrates how notebooks can be used in a purely "read-only" mode.

The final report has a constraint of *3000 words* of prose, and no limit on the amount of code. However, if you develop very large code blocks as part of the project, you will may want to include them in a separate Python file, rather than include all the code in the notebook (though the later is acceptable too, if you put the code in a separate appendix section in the notebook).

Midterm report - due 11/11

The midterm report should mark a "halfway point" in developing your project. It should be a notebook with similar goals as the final report, but with at most 1000 words. Given that a substantial portion of the project involves understanding some real-world data, the main goal of the midterm report should simply be to show that you have come fairly far along in getting the data, performing some basic analysis, and understanding the types of questions that you want to

answer. It is fine if you want to change the text substantially between the midterm report and the final.

Project proposal - due 10/24

The project proposal requires that you write a maximum of 200 words describing the overall subject and aim of your project. Like with the tutorial, the main goal here is to arrive on a topic, and receive brief feedback from the instructors about whether this seems like a good project to work on (the feedback will likely be slightly more detailed than the tutorial feedback, but not much). We'll post a Google Docs link for students to submit these proposals.

Video presentations - held on 12/14, time TBD

After the final reports have been submitted, students will be required to record and present a 2 minute video highlighting their work. The reason for these videos (over say a poster session), is to ensure that every group gets approximately the same amount of time with the class to present their work. This video can be a recording of a Powerpoint presentation, a screen-capture of some aspect of the system you built, or virtually anything else, as long as it's a video file and it lasts 2 minutes (and describes your project). The file should be submitted in MP4 format, and you should check to ensure it can be played by "standard" media players on Mac and Windows.

Note that the date above is *not* the scheduled time of the final exam. Since the final exam date was scheduled quite late in the finals period, we opted to hold these presentations earlier. However, since this may preclude some students from attending the presentations, we are additionally ensuring that 1) we will have a live-stream of the event so that remote students can participate in real-time, and 2) those who absolutely cannot participate in real time can also view the videos separately.

As a student in the course, you'll be required to provide feedback on 5 other student presentations (selected randomly). This feedback will not affect the grade of the other projects (unlike as with the tutorial), but we will provide (anonymously) the feedback to the other students. Detailed forms for providing feedback will be made available online, and we strongly request that anyone who can attend the in-person presentations, and fill our your feedback there.

Grading

Grading will be done on a standard scale (A, A-, B+, B, etc), and will be based largely upon how well you address the following points:

- Does the project describe what is the problem you are addressing in your project, what is the relevant data, and how you can use this data to answer the question at hand?
- Does the project use techniques presented in the course (or clearly related to topics covered in the course) to understand and analyze the data for this problem.
- Does the report situate the work clearly within the related work in this subject area.
- Does the report clearly demonstrate the work that was done, and provide directions for further investigation.

The final report will count for the majority of the total grade on the assignment (65%, based only upon instructor feedback and not other student feedback), with the proposal adding an additional 5%, midterm report 10% (in both cases mainly just checking to ensure that groups made a reasonable effort to complete these milestones), 15% for the video, and 5% for your comments on other students' videos (again, just a check that you made a reasonable effort).