Instructions for the SI630 Project Update

Version 1.0

Put your name here

1 Introduction

The course project is intended to provide an opportunity for students to dive deeper into one problem or topic of their choice and write a very small scale study on the topic. Projects typically take two forms: (1) the student has some data, problem, or algorithm in mind and proposes a study to investigate these or (2) students pick an existing NLP task and try a new approach to solving it. Tasks for the latter are detailed more in Section **??**. In both cases, projects should be *feasible* for completing in the available time frame. The latter part of the course has a lighter workload to allow more time for working on the project, but we want to ensure that students pick projects that help them learn real, practical skills in NLP without being trivial. Ideally, your course project is a chance to develop something you can show off to future employers or could serve as a pilot study for a full research project.

For this project, you're expected to use this LaTeXtemplate. You're welcome to copy this template directly off of Overleaf as well using this URL: https://www.overleaf.com/read/kqktwvvmypfh or clone it using `git` at https://git.overleaf.com/13769810nctdwybfjxtf, which hopefully has enough examples of how things can be written to get you started. If you're having any issues getting LaTeXto do what you want please feel free to ask the instructors or know that there's a great resource on WikiBooks https://en.wikibooks.org/wiki/LaTeX and a whole StackExchange site dedicated to answering questions https://tex.stackexchange.com/. LaTeXis a common method for writing technical documents so we are using it for 630 to help get you started on using it for your career. It also is pretty awesome for citation management.

Finally, please remember that the 630 instructors are here for you and will gladly offer suggestions and advice on projects. We want your projects to succeed, to be fun to work on, and to spark your intellectual curiosity!

2 What to do (10 points total)

The project update serves two purposes. The first is that it helps you write an initial rough draft of your final project report. The second is that it requires you to have all of your data in hand and to develop a simple baseline for your proposed task. Both of these are important steps at keeping your project on track. Following, we outline the key requirements for your update.

Introduction (1 point) You should have a rough draft of the introduction that clearly states what the problem is and provides some broader context. We recommend writing the introduction last after you finish the Problem Definition and Related Works sections. You should also include a statement on why solving this problem matters–who would care if you solved it and what effect would solving it have? In the project update, you should include details of your proposed method for solving the problem (even if you haven't implemented it).

Data (3 points) The section describes what data you will use for the project. For the update, you should already have the data on hand. You should describe the source of the data, how you obtained it, what type of preprocessing steps you took, and (if not sensitive data) include a few examples. We strongly encourage you to include some very rough statistics (e.g., how many instances you have, the class distribution if doing classification, relevance score distribution) in a table format. If you had to create your own dataset or needed to annotate a ground truth relevance, this section should specify how you did it and provide details on the relevance scores.

If your data doesn't have the required ratings to evaluate, e.g., you're building an NLP system to predict something but the data doesn't have labels for that construct, you will need to annotate. The amount of annotation depends on the task; in general, to get meaningful performance most NLP systems will require large hundreds or low thousands of examples. We will require upper hundreds and check with us before going down this path to get a sense of scope and necessity.

Related Work (1 point) The related work section should describe how other people have thought about the problem you're working on. How did they approach it? What makes their problem different from yours? Why do you think your approach will be better? For your update, you should have at least **five** papers related to your current problem and a few sentences describing what they did to solve the problem. We recommend using Semantic Scholar or Google Scholar to help find related papers.

Your description of each related work should be *at least* a few sentences each. Highlight the main aims and, where possible, how the model was evaluated and what was learned. You will want to contrast with these systems (and possibly their evaluation setups!) in your Results and Discussions sections.

Methodology (1 point) This section will describe how you solve your problem. Go into algorithmic details and be sure to describe what various kinds of preprocessing steps you did. Someone should be able to recreate your exact methodology from the description. Be specific about what each step does. For example, it's insufficient to say "we trained a classifier;" instead say something like "we trained a Random Forest classifier using 250 trees and requiring a minimum of 5 items per leaf"

For the update, you should include a *detailed* outline of the method you plan to try, even if you haven't implemented yet. Your update describe to a reader what you want to do and *why* you want to do it. Think of this part as an exercise in writing a full description of how you plan to solve the problem. This update also lets us give you feedback on different parts of your plan. You do not need to have implemented anything—your plan will likely evolve in the coming weeks as you get exposed to new deep learning models but the act of writing a plan will be critical for ensuring you know what to look for in new techniques. We will also be able to give you feedback on the plan too during our review to help spot any important changes that will need to be made.

Evaluation and Results (3 points) This section provides an overview of how you evaluated your method on the data. What methods did you compare against? How successful were you? Describe the exact evaluation setup and what kinds of steps were taken.

For the update, you should clearly define one or more baselines to compare your system against. One baseline should be random performance. A second baseline should be something reasonable that doesn't require much knowledge or learning. For example, if you're doing a classification, always choosing the most frequent class is a useful baseline; or if building a summarization system, using the first sentence of a document is a reasonable "simple summary." *Generating a result on your actual data with a real baseline method is the most important part of the update and will have the biggest effect on your grade.* You need to demonstrate that you can work with the data to solve the problem (even poorly with a baseline method!) so that when you actually try to solve it with your own method, you know how to work with the data and know how to evaluate your system. If you're having trouble coming up with a baseline, please see one of the instructors immediately. If you don't have data for the problem you're working on, consider switching to a different task where you can get the data.

You should have at least one figure or table showing your baseline's results. Please make sure to label all your axes and make the font size legible without having to zoom in excessively.

Work Plan (1 point) For the update, describe your workplan for the semester in terms of (i) what you've done so far and (ii) what you intend to do to finish the project. Be specific and lay out weekly objectives/milestones that you can use to keep track of progress. This part should an exercise in reflecting on project planning and seeing how you can improve your estimates for how long things will take (it's difficult!). We won't hold you to this workplan for the final project, but we have found that the act of creating such a workplan generally helps students quantify the remaining work involved and scope their project better (fewer April surprises!).

Acknowledgments

If you got help from anyone or had substantive discussions, please acknowledge those people here and describe how they contributed. The work you do for your project should be entirely your own.

References

Jochen WL Cals and Daniel Kotz. 2013. Effective writing and publishing scientific papers, part vi: discussion. *Journal of clinical epidemiology* 66(10):1064.

Sheela P Turbek, Taylor M Chock, Kyle Donahue, Caroline A Havrilla, Angela M Oliverio, Stephanie K Polutchko, Lauren G Shoemaker, and Lara Vimercati. 2016. Scientific writing made easy: A step-by-step guide to undergraduate writing in the biological sciences. *The Bulletin of the Ecological Society of America* 97(4):417–426.

Note that you must cite all your references

A Supplemental Material

If you want to put longer examples of data and code, put it here in the appendix.

B How to cite a paper

You should use BibTeX to cite papers with requires putting the information make for your reference in your `references.bib` For example, the references on pun interpretation in the bibtext is

```
@inproceedings{miller2017semeval,
  title={SemEval-2017 Task 7: Detection and
         interpretation of English puns},
  author={Miller, Tristan and Hempelmann,
         Christian and Gurevych, Iryna},
  booktitle={Proceedings of the 11th
             International Workshop on
             Semantic Evaluation (SemEval-2017)},
  pages={58--68},
  year={2017}
}
```

In text, you then cite this work by including `\cite{miller2017semeval}`. See the original Latex for this file linked on Overleaf (though canvas) for the example. If you want to citation to appear as "Author (Year)" (which is the appropriate style when referring to someone by name), the you can use `\newcite{miller2017semeval}`

C Writing Notes

There are many other good guides to writing a paper-like project report and I encourage you to read these examples.

1. http://www.people.vcu.edu/~rbfranklin/science%20writing.pdf

2. http://bit.ly/2ElGoPT

3. http://onlinelibrary.wiley.com/doi/10.1002/bes2.1258/full

The introduction should summarize the problem you're working on and set the stage for the reader on what to expect. Turbek et al. (2016) notes that

> the Introduction sets the tone of the paper by providing relevant background information and clearly identifying the problem you plan to address. Think of your Introduction as the beginning of a funnel: Start wide to put your research into a broad context that someone outside of the field would understand, and then narrow the scope until you reach the specific question that you are trying to answer. Clearly state the wider implications of your work for the field of study, or, if relevant, any societal impacts it may have, and provide enough background information that the reader can understand your topic. Perform a thorough sweep of the literature; however, do not parrot everything you find. Background information should only include material that is directly relevant to your research and fits into your story; it does not need to contain an entire history of the field of interest.

Cals and Kotz (2013) notes that

> Start thinking about the discussion even before collecting the first data. Many aspects and "pearls and pitfalls" of the study, as well as its relation with other studies in the field, will be discussed when developing, carrying out the research and analyzing the data, and in project group meetings. Make notes and a list of keywords as a reminder of these useful discussions, while remembering your story line at all times. Having such a list will greatly facilitate writing the first draft of the discussion section and will serve as a skeleton for this section of the paper (see "How to start writing").
>
> Start by presenting the main findings, by answering the research question in exactly the same way as you stated it in the introduction section (see "Introduction"). If you cannot present the main findings in three sentences, it may mean that you have forgotten the storyline of the paper. Do not waste words by repeating results in detail, and only use numbers or percentages if they are really necessary for your message. Do not ignore or cover up inconvenient results. Reviewers will pick them up anyway, and it weakens your paper if you try to hide them. Also, do mention unexpected findings by explicitly stating that they were unexpected and did not relate to a prior hypothesis; such honesty will strengthen your paper.