Instructions for the SI630 Project Proposal

Version 1.0

Put your name here

Abstract

This document summarizes the main parts of the project proposal, which sections are needed, and what's needed in them. Your task will be to fill in these parts, which are marked in red. There is additional commentary to help you in the scientific writing process. The proposal, progress update, and final submission all build upon one another so the work you put into the proposal will directly carry over to each and help provide a roadmap of what's to be done.

For now, your abstract should just summarize what's the problem you're working on.

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 2. 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!

1.1 Writing notes

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.

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

1.2 What to do

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?

2 Problem Definition and Data

The section describes what specific problem you plan to solve and goes into much more detail than the Introduction. You can also add details about what your problem is not to help guide the reader's expectations. You should also go into detail about how you define success, i.e., what are the evaluation metrics you will use to evaluate. How will you know if you have a good solution to your problem?

Second, you should describe what data you will use for the project. Ideally, you should already have the data on hand or spend a few minutes getting it. Projects are *much* more successful when most of the time is spent solving the problem rather than searching for the right data. If you don't have data at proposal time, please be very specific on how you plan to get it. We *might* be able to recommend something but we encourage you to come to talk to us during office hours or after class. If you have the data, you should include a few examples and can also include some very rough statistics (e.g., how many instances you have).

Since not everyone has a burning question they're dying to answer with NLP, we've included a few potential NLP problems people could work on in Table 1. For these tasks, the problem, data, and evaluation criteria are already provided–though you should be sure to describe them here as a part of your proposal.

Important Note: For the final project submission, student-proposed NLP tasks and existing tasks are evaluated slightly differently. If you choose an existing NLP task for your project (e.g., from Table 1), you will be expected to submit a final report that has a complete solution (a working algorithm) for the task. This expectation does not mean you have to have a high-performing or super novel system, just that you have completed the project. If you pitch your own problem to work on, we realize that sometimes a task can be a bit more daunting than expected, or the data might not be what you expected. In these cases, you're expected to report in detail how far you got and what went right and what went wrong. Ideally, in these negative results papers, you should still be presenting a substantial analysis of the problem or data, it's just that this analysis doesn't ultimately lead to a solution. Also, the instructors for this course will do whatever possible to help guide you in your projects so you don't get stuck!

For the proposal, you should be have a clearly stated problem and a very high-level description of the data with a few examples shown.

3 Related Work

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?

You should have at least three papers related to your current problem and a few sentences describing what they did to solve the problem. Since you haven't tried solving the problem yet, you don't need to compare with them at all.

4 Methodology

This section will describe how you solve your problem. Go into algorithmic details and be sure to describe what various kinds of preprocessing

Task	Reference Paper	Website
Natural Language Inference	(Bowman et al., 2015)	https://nlp.stanford.edu/projects/snli/
Question Answering	(Rajpurkar et al., 2016)	https://rajpurkar.github.io/SQuAD-explorer/
Scientific Keyphrase Extraction	(Augenstein et al., 2017)	https://scienceie.github.io/
Detection and Interpretation of English Puns	(Miller et al., 2017)	http://alt.qcri.org/semeval2017/task7/
Sentiment Analysis in Twitter	(Rosenthal et al., 2017)	http://alt.qcri.org/semeval2017/task4/
Named Entity Recognition in Noisy Text	(Derczynski et al., 2017)	http://noisy-text.github.io/2017/emerging-rare-entities.html
Multilingual Emoji Prediction[1]	none yet!	https://competitions.codalab.org/competitions/17344
Irony detection in Tweets	none yet!	https://competitions.codalab.org/competitions/17468
Hypernym Discovery	none yet!	https://competitions.codalab.org/competitions/17119

Table 1: Examples of NLP tasks that you could choose to work on for your project. The last three tasks are very recent tasks that have details and data on the website but no papers yet. You can work on something cutting edge!

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"

As a part of this section, please describe *why* you chose what you did. What was your design process and how did you approach solving the problem?

For the proposal, you should include a very general description of what method(s) you plan to try.

5 Evaluation and Results

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.

You should also clearly define some 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.

For the proposal, you should propose a very simple baseline to compare your model against.

6 Discussion

The discussion section is where you start to unpack the results for the reader to help them understand what was learned. You may not have many results at the moment (but you should have some!) so you can discuss here what has gone wrong and right in your current setup. For example, maybe you realized you needed more data, or maybe you realized that the SVD was not helping your analysis. The discussion should point the way to what work will be done next.

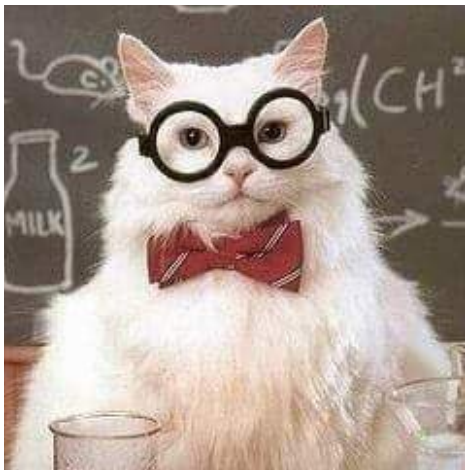

Figure 1: Eventually, you'll have cool figures to show here.

You can leave this section blank.

6.1 Writing Tips

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.

7 Work Plan

Normally, you would have a conclusion to end the work, but as this is a project proposal, you'll end with a plan for how you'll accomplish your project. We hope this section can help you think at a high-level about which tasks are necessary to get to the point where you have a working model. Of course, plans are often made and then changed when new information or challenges emerge, so we won't hold you you to this. However, the act of writing a plan can greatly help you figure out how think about the process and, in general, projects that are proposed with more concrete work plans tend to be more successful.

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

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853* .

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* .

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

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. pages 140–147.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. Semeval-2017 task 7: Detection and interpretation of english puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. pages 58–68.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* .

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. pages 502–518.

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.