CSC 2511: Course project option

This year there is a course project option for graduate students enrolled in CSC2511 who:

- work better in teams,
- want to work in a team,
- would prefer a large unified research project instead of relatively isolated assignments, or
- some combination of the above.

The purpose of these projects is to produce **original research** in natural language computing. In fact, teams that produce **high-quality** projects will be encouraged to submit their work to **relevant journals or conferences**. Although this option replaces the need to perform the assignments, it may involve more programming, experimentation, and analysis in total.

Teams must consist of 1 or 2 humans – no more and no fewer! Projects must contain a significant programming component; you are free to choose your programming languages, but your code must run on the CDF servers. Projects must utilize large and relevant linguistic corpora or tools. *Projects must involve at least one of* **corpus statistics**, **statistical machine translation**, *or* **automatic speech recognition**. Many of the important concepts in these themes won't be discussed in class until it is too late to incorporate them into your projects, so we will advise you to get the required knowledge 'early'.

Some possible ideas for projects include but are not limited to:

- A novel speech-to-speech machine translation application that converts speech in language A to equivalent speech in language B for $A \neq B$.
- A novel method of using data in language A to train a classification system in language B for $A \neq B$. This would be useful, for example, if there is not enough data in language B (e.g., Klingon) to train a competent system, but lots of data exist for language A (e.g., English), and some relationship between A and B is known (e.g., the word *Qapla'* $\equiv$ *success*).
- Processing of **health-related** information on the web.
- Processing of **duplicitous, deceitful, or devious** text on the web.

Marking scheme

The marking scheme for projects is to a large degree at the discretion of the markers. However, marks will be assigned according to these proportions:

Corpus analysis	15%	Judicious use of relevant corpus statistics
Programming	20%	Appropriate, correct, and original developmental effort
Experiments & analysis	30%	Appropriate experiments, experimental procedures, and thoughtful analysis of results
Report – literature review	10%	Extensive and correct survey of relative prior work
Report – technical quality	15%	Extensive and correct technical description of your work
Report – presentation	10%	Overall quality of academic-style monograph

Requirements and checkpoints

There will be 5 (+1) scheduled checkpoints over the course of this project at which your team must submit your progress to the course instructor by email to `csc401-2021-01@cs.toronto.edu`. You should stay in touch with your instructor between checkpoints to ensure sufficient progress – he is available to advise you at each step of the process. Feel free to complete each checkpoint earlier, if you wish, as long as you don't sacrifice quality. As this is a research project you are naturally able to correct your plans and methodologies between checkpoints, within reason (e.g., if it becomes clear that some set of data is unavailable, you will need to select some other set of data).

Each submission (except the zeroth) should be submitted as PDFs with 12pt main body text in Times New Roman font, single column, double-spaced, on letter-sized paper with 2.54cm borders all around. References should indicate all relevant fields (e.g., see `http://nwalsh.com/tex/texhelp/bibtx-7.html`) and be ordered in a references section by the last name of the first author. Citations in the text should be in the author-year format (e.g. "I am the walrus (Lennon, 1967)"). The style of the text should be concise and clear.

0. Notification – 18 January

If you decide to undertake this project option, you need to notify the course instructor by email before 16 January. That email only needs to consist of the names of your team members, confirmation that each team member is a graduate student enrolled in CSC2511, and an agreement that you have decided to take the project option rather than perform the assignments. *You may switch back to the assignment stream at any time if you wish, but you will not be given special considerations or extensions on those assignments if you do.* If your team consists of $N = 2$ human beings, only 1 team member needs to send this email, but it should be CC'd to the other team member.

1. Proposal – 29 January

Your proposal should be between $\frac{1}{2}$ **and 1 page** of text. Indicate a temporary title for the project. Indicate clearly the challenge you will attempt to solve. Outline the type of data you will need to use for this project as specifically as possible (e.g., size, content, source). Note that collecting your own data can be time-consuming and will typically involve a review by the University of Toronto's Research Ethics Board, which can be quite an involved process. Estimate the steps you will take to meet your goals in terms of computational models, algorithms, or other resources (e.g., pre-existing machine learning software). What are your hypotheses? How will those hypotheses be tested and how will your results be evaluated? This checkpoint should be fairly abstract and introductory.

2. Literature review – 5 February

Your literature review will survey and synthesize a substantial collection of existing work that pertains to your chosen project topic. The document you submit should be between **5 and 10 pages** of text, plus any additional pages needed for references. This document should encapsulate a revision of your proposal from the previous checkpoint as the introduction.

Your literature review at this stage should consist of a core of at least 10 *relevant* scientific articles and publications that are closely related to your intended project. You are responsible for finding these publications yourself. You should summarize the methods and findings of each of these core publications, compare and group them as appropriate, and point out any substantial advances or failings you may encounter. These core publications may also be accompanied by references to other material such as textbooks or websites that help to justify or define your claims but these sorts of general sources are not sufficient for the core references which should outline recent progress in your chosen field and would ideally be from the last 3 to 5 years.

3. Experimental setup – 5 March

This checkpoint will describe in detail the programming component of your project and your planned empirical methods, which should be largely complete by this stage. The document you submit should be between **10 and 20** pages of text, plus any additional pages needed for references. This document should encapsulate revisions of your proposal and literature review as the introduction and background sections, respectively. Feel free to add new references as the need arises.

Your experimental setup should unambiguously describe the methods you will use to conduct your experiments so that they can be easily recreated by other researchers. This includes a thorough description of your data (specifying which data will be used for which experiments), your mathematical models (with relevant formulae), your algorithms, and your proposed experiments (including some indication of what you expect to find and how your experiments are designed to test your hypotheses). You should include more than one primary experiment, ideally each of a distinct nature. For example, one experiment may compare the performance of various classifiers on one set of data and another experiment may compare the effects of certain features extracted from the data.

4. Initial analysis – 26 March

This checkpoint will describe the results of your experiments, which should be largely complete by this stage. The document you submit should be between **20 and 30** pages of text, plus any additional pages needed for references. This document should encapsulate revisions of your proposal, literature review, and experimental setup, which should all flow together in a relatively coherent document.

Your analysis should indicate a good depth of scientific inquiry – a few metrics of success is not sufficient. Typically, results will include a few tables of numerical results and a few graphical figures as the need warrants. These will be accompanied by text that explains the meaning and significance of these results (including any statistical tests of significance, where appropriate). Some comparison to the prior work in the literature review would be appropriate, as would some claims about the degree to which your hypotheses have been satisfied (or falsified).

5. Final report – 9 April

This checkpoint finalizes the nearly-complete document from the previous checkpoint in a form that can be submitted for scientific publication to a relevant conference or journal. The document you submit should be between **25 and 35** pages of text, plus any additional pages needed for references. This will involve the addition of an abstract (no more than 500 words) that summarizes the important points in the document, a discussion which extrapolates your findings and provides high-level insight into your work, and suggestions as to follow-up work which may be undertaken as a result of your work. It will be important at this stage to concentrate on the quality of your written presentation.

When you submit your final report PDF, you must also submit all source code that was used in your experiments or a pointer to a world-readable section on the CDF or Computer Science servers so that the instructor can test your claims and examine your data and results. You may be asked to meet with the instructor to 'walk through' your code and results after the final exam so that your grade on the project can be properly evaluated.

If you work in a team of 2 members, your final report must also be accompanied by a text file, *contributions.txt*, in which the specific contributions of each team member are enumerated.