

CSC 401/2511: Natural Language Computing

Instructors: Gerald Penn (GP), Raeid Saqr(RS), and Sean Robertson(SR)
Email: csc401-2024-01@cs.toronto.edu
Lectures: **M** 10h-11h (LEC0101) at MP 102; 11h-12h (LEC0201) at KP 108.
W 10h-11h (LEC0101) at MP 102; 11h-12h (LEC0201) at KP 108.
Tutorials: **F** 10h-11h (LEC0101) at MP 102; 11h-12h (LEC0201) at KP 108.
Web-page: <http://www.cs.toronto.edu/~raeidsaqr/csc401/>
Forums: Piazza, Quercus
Office hours: Gerald Penn: F 12-14h at PT 283; SR/RS: M 12-13h at BA 2270/Zoom

This course presents an introduction to natural language computing in applications such as neural language models, information retrieval and extraction, intelligent web searching, speech recognition, and machine translation. These applications will involve various statistical and machine learning techniques.

Prerequisites: CSC 207 or 209 or 228, and STA 247 or 255 or 257 and a CGPA of 3.0 or higher or a CSC subject POST. MAT 223 or 240, CSC 311 (or equivalent) are strongly recommended.

Evaluation policies

CSC401/2511 students will be marked on three homework assignments and a final exam. The relative proportions of these marks are as follows:

| | | |
|----------------|-----|-----------------------|
| Assignment 1 | 20% | language: Python |
| Assignment 2 | 20% | language: Python |
| Assignment 3 | 20% | language: Python |
| Ethics surveys | 1% | (2 surveys 0.5% each) |
| Final exam | 39% | |

All assignment submission code must run on the teaching machines.

Note that a 24-hour ‘silence policy’ will be in effect – we do not guarantee that the instructors or TAs will respond to your request within 24 hours of an assignment’s due time.

Lateness

A 10% deduction is applied to late homework one minute after the due time. Thereafter, an additional 10% deduction is applied every 24 hours up to 72 hours late at which time the homework will receive a mark of zero. No exceptions will be made except in emergencies, including medical emergencies, at the instructor’s discretion.

Final exam

The final exam will be a timed 3-hour test. A grade of 50% or higher on the final exam is required to pass the course. In other words, if you receive a grade lower than 50% on the final exam then your final grade in the course will be no higher than 47%, regardless of your performance in the rest of the course.

Academic offences

No unauthorized collaboration on the assignments is permitted. The assignment materials (starter code, solutions etc.) **must not be shared publicly** (e.g. via github) in perpetuity without explicit approval by the instructors. The work you submit must be your own. ‘Collaboration’ in this context includes but is not limited to sharing of source code, correction of another’s source code, copying of written answers, and sharing of answers prior to submission of the work (including the final exam). Failure to observe this policy is an academic offence, carrying a penalty ranging from a zero on the homework to suspension from the University. See the academic integrity page of the University of Toronto at <https://www.academicintegrity.utoronto.ca/>.

Readings

- Optional Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. MIT press, 2016 *Deep Learning, 1st ed.*. **Available at:** <https://www.deeplearningbook.org/>
- Optional Christopher D. Manning and Hinrich Schütze (1999) *Foundations of Statistical Natural Language Processing*, MIT Press. **Available free at:** <http://go.utlib.ca/cat/10552907>
- Optional Daniel Jurafsky and James H. Martin (2021) *Speech and Language Processing, 3rd ed.*. **Available free at:** <https://web.stanford.edu/~jurafsky/slp3/>

See website for additional readings.

Planned topics

1. Introduction to corpus-based linguistics
2. N -gram models and linguistic features, word embeddings
3. Entropy and information theory
4. Intro to deep neural networks and neural language models
5. Machine translation (statistical and neural) (MT)
6. Transformers, attention based models and variants
7. Large language models (LLMs)
8. Acoustics and phonetics
9. Speech features and speaker identification
10. Dynamic programming for speech recognition.
11. Speech synthesis (TTS)
12. Information Retrieval, Summarization

Planned course calendar

See Academic dates & deadlines for undergraduate students and Sessional dates for graduate students for any changes.

| | |
|----------------|--|
| 8 January | First lecture |
| 16 January | Last day to add CSC 2511 |
| 22 January | Last day to add CSC 401 |
| 9 February | Assignment 1 due |
| 19 February | Last day to drop CSC 2511 |
| 19–23 February | Reading week – no lectures or tutorial |
| 9 March | Assignment 2 due |
| 13 March | Last day to drop CSC 401 |
| 5 April | Last lecture |
| 5 April | Assignment 3 due |
| 5 April | Project final report (if applicable) due |
| TBD April | Final exam |

See course website for details and updates.