CSC 401/2511: Natural Language Computing

**Instructor:** Frank Rudzicz

Lectures: MW 10h–11h, AH 100  
Office hours: M 11h30–12h30, Vector Institute  
Email: frank@cs.toronto.edu  
Tel: 416 597 3422 x7971  
Tutorials: F 10h-11h, AF 100  
Forum: [https://piazza.com/class/jb4cummc8pi5pl](https://piazza.com/class/jb4cummc8pi5pl)

**TAs:**  
A1 Willie Chang and Ramin Zaviehgard  
A2 Mohamed Abdalla and Hamed Heydari  
A3 Bai Li and Lin Gao

This course presents an introduction to natural language computing in applications such as 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 POS. MAT 223 or 240 is 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
- Final exam 40%

Graduate students enrolled in CSC2511 will have the option of undertaking a course project instead of the assignments in teams of at most two students for 60% of the course mark; all graduate students need to pass the final exam, which is worth 40% of the final mark. All code must run on the CDF machines.

**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**

A grade of at least D– on the final exam is required to pass the course. In other words, if you receive an F on the final exam then you automatically fail the course, regardless of your performance in the rest of the course.
Academic offenses

No unauthorized collaboration on the assignments is permitted. 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 offense, 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 http://www.artsci.utoronto.ca/osai/students.

Readings


Planned topics

- Introduction to corpus-based linguistics
- N-gram models and data features
- Entropy and information theory
- Hidden Markov models
- Statistical machine translation
- Articulatory and acoustic phonetics
- Automatic speech recognition
- Speech synthesis
- Information retrieval
- Text summarization
- Classification with linguistic data

Planned course calendar

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 January</td>
<td>First lecture</td>
</tr>
<tr>
<td>17 January</td>
<td>Last day to add CSC 401</td>
</tr>
<tr>
<td>22 January</td>
<td>Last day to add CSC 2511</td>
</tr>
<tr>
<td>12 February</td>
<td>Assignment 1 due</td>
</tr>
<tr>
<td>19-23 February</td>
<td>Reading week – no lectures or tutorial</td>
</tr>
<tr>
<td>26 February</td>
<td>Last day to drop CSC 2511</td>
</tr>
<tr>
<td>9 March</td>
<td>Assignment 2 due</td>
</tr>
<tr>
<td>14 March</td>
<td>Last day to drop CSC 401</td>
</tr>
<tr>
<td>4 April</td>
<td>Assignment 3 due</td>
</tr>
<tr>
<td>4 April</td>
<td>Project final report due</td>
</tr>
<tr>
<td>TBD April</td>
<td>Final exam</td>
</tr>
</tbody>
</table>

See course website for details.