This course presents an introduction natural language computing in applications such as information retrieval and extraction, intelligent web searching, speech recognition, and multi-lingual systems including machine translation. These applications will involve techniques such as \( n \)-grams, part-of-speech tagging, semantic distance metrics, indexing, entropy, hidden Markov models, corpus analysis, and IBM Watson.

**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 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:

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Percentage</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignment 1</td>
<td>20%</td>
<td>Python</td>
</tr>
<tr>
<td>Assignment 2</td>
<td>20%</td>
<td>Matlab</td>
</tr>
<tr>
<td>Assignment 3</td>
<td>20%</td>
<td>Matlab</td>
</tr>
<tr>
<td>Final exam</td>
<td>40%</td>
<td></td>
</tr>
</tbody>
</table>

Undergraduates are requested to form teams of two to complete the assignments. Instructions will be given in class. Each team member gets the same mark on each assignment.

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 team's 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.utoronto.ca/academicintegrity/academicoffenses.html.

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

11 January First lecture
24 January Last day to add CSC 401
25 January Last day to add CSC 2511
12 February Assignment 1 due
15–19 February Reading week – no lectures or tutorial
  1 March Last day to drop CSC 2511
  11 March Assignment 2 due
  13 March Last day to drop CSC 401
  8 April Last lecture
  8 April Assignment 3 due
  8 April Project final report due
  TBD April Final exam

See course website for details.