COMP/LING 596, 483, 682: From Natural Language to Data Science, Fall 2020

Tue/Thu 13:05 (1:05 PM) – 14:25 (2:25 PM)

Instructors

Siva Reddy
siva@cs.mcgill.ca
Office hours: By Appointment

Tim O'Donnell
timothy.odonnell@mcgill.ca
Office hours: By Appointment

Teaching Assistant

Ian Paroda
ian.porada@mail.mcgill.ca
Office hours: By Appointment

1 Overview

The last decades have seen phenomenal increases in our scientific and engineering understanding of language from a computational perspective. A large part of this success has been the rapid and unprecedented expansion of different kinds of language data as well as new computational tools for dealing with this data. This course provides an introduction to the data science of language. The emphasis will be on learning basic tools for working with language data for both engineering and scientific applications.

2 Goals

The goal of this course is to learn how to think about language data (predominantly raw text) and to work with it computationally. Along the way, we will learn a number of mathematical and computational tools for data collection, representation, processing, modeling, and analysis. The main emphasis of the course will be on transferring theory into practice. By the end of the course, you will know how to visualize large amounts of text, study the nature of words, build thesaurus, search for information, predict outcomes from social media text while also studying the ethical implications. And one of the side goals is to have fun with text!!

Students will:

- Learn to visualize text.
- Learn how to query linguistic databases.
- Learn how to analyze data using simple statistical methods.
- Learn how to build simple models of language domains.
- Learn how to predict outcomes based on text.
- Learn the biases of current NLP technology and their implications.
- Learn how to process different kinds of language data using Python.
- Learn how to query linguistic databases.

3 Prerequisites

Required: Programming background at the level of COMP 202 or equivalent.

Recommended: Programming background at the level of COMP 250 or higher. Mathematics background at the level of MATH 240. Basic calculus and linear algebra will be helpful but not critical.

4 Course structure

The course will be taught in 8 modules, divided between the two instructors.

I: Language data and applications (Reddy, O'Donnell)

II: Searching, querying, and organizing linguistic data (Reddy)

III: How to make sense of text data (Reddy)

IV: Modeling sequence structure (O'Donnell)

V: Supervised learning: Classification and Regression (O'Donnell)

VI: Information Retrieval and Extraction (Reddy)

VII: Language Variation and Change (O'Donnell)

VIII: Ethics (Reddy)

Readings will be posted on myCourses as they are needed, and lecture slides will be posted after each lecture.

There will be 7 problem sets, one for each of the first 6 modules, and one for the last two modules. All problem sets will use the python programming language.

Problem set questions can consist of: (i) programming problems in Python (ii) mathematical problems and (iii) short answer problems. Problem sets are due before the beginning of class at 13:05 in the afternoon on the due dates specified below. Note that we cannot debug code that does not run and problem sets whose code does not run will receive a 0.

Participation and interaction is a encourage in this course. We will sometimes use 15–20 minutes of class time to cover questions on preceding problem set or to review other difficult material.

Class discussion and announcements will take place through the myCourses site.

5 Evaluation

Note that details below are subject to change.

- Warm-Up Problem Set (10%): The first problem set will be a warm up to get used to implementing Python and submitting.
- Problem Sets (90%): 6 problem sets equally weighted (15% each).

6 Logistics

Course Website: https://mycourses2.mcgill.ca/d2l/home/473665

7 Lecture Recordings

This fall this course will be delivered online. Lectures will be delivered live, at the scheduled time on Tuesdays and Thursdays. However, in order to make sure that the course content is available for students located in other time zones who are unable to be in Montréal this term, lectures will be recorded using Zoom and uploaded to course website on myCourses. Note that these videos may only be available on myCourses for a limited time after the delivery of each lecture.

Attendance: Although these videos will be available, students are strongly encouraged to attend the live lectures, if it is at all possible. The live lectures provide an opportunity for interaction that will otherwise be lacking and will allow the lecturer to more finely tune the material depending on questions and feedback.

Cameras: Although it is not required, we ask that students please leave on their cameras whenever possible. Having more members of the class visible greatly increases the feeling of interaction, feedback, and community that can sometimes be missing from online course delivery. It also allows the lecturer to better gauge how students are following the material.

Participation: Students are strongly encouraged to participate in class by asking questions and participating in discussions during lectures. This can be done by using the hand raising function in Zoom, and then once the lecturer has called on you, turning on your mic and asking the question. Students may also ask questions in the Zoom chat during lecture, although to encourage discussion the first option is preferred.

Video Privacy: To protect the privacy of students, lecturers, and teaching assistants, all class participants must ensure that videos and associated materials are not reproduced or placed in the public domain. This means that the materials can be used for educational and research purposes, but cannot be shared with others by putting them up on the internet, by giving them away or selling them, or by allowing others to copy them or otherwise make them available. Please refer to McGill's Guidelines for Instructors and Students on Remote Teaching and Learning for further information.

Changes: Since this is the first time this course has been given entirely online, many of the details are experimental. We may make changes to delivery, participation, or recording if we discover better ways of doing things.

8 Readings

Reading will be posted to myCourses as needed.

9 McGill Policy Statements

McGill University values academic integrity. Therefore, all students must understand the meaning and consequences of cheating, plagiarism and other academic offences under the Code of Student Conduct and Disciplinary Procedures (see www.mcgill.ca/students/srr/honest/ for more information).

In accord with McGill University's Charter of Students' Rights, students in this course have the right to submit in English or in French any written work that is to be graded.

Instructor generated course materials including recorded lectures, readings, notes, summaries, etc. are protected by law and may not be copied or distributed in any form or in any medium without explicit permission of the instructor. Note that infringements of copyright can be subject to follow up by the University under the Code of Student Conduct and Disciplinary Procedures.

10 Course schedule

Note that the exact dates below are subject to change, depending on how quickly we make progress through topics in the course.

- Thursday, September 3, 2020
  - ✎: Problem Set 1 Released
- Tuesday, September 8, 2020
- Thursday, September 10, 2020
- Tuesday, September 15, 2020
- Thursday, September 17, 2020
  - ✎: Problem Set 1 Due
  - ✎: Problem Set 2 Released
- Tuesday, September 22, 2020
- Thursday, September 24, 2020
- Tuesday, September 29, 2020
- Thursday, October 1, 2020
  - ✎: Problem Set 2 Due

  - ✎: Problem Set 3 Released
- Tuesday, October 6, 2020
- Thursday, October 8, 2020
- Tuesday, October 13, 2020
- Thursday, October 15, 2020
  - 🖉: Problem Set 3 Due
  - ✎: Problem Set 4 Released
- Tuesday, October 20, 2020
- Thursday, October 22, 2020
- Tuesday, October 27, 2020
- Thursday, October 29, 2020
  - 🖉: Problem Set 4 Due
  - ✎: Problem Set 5 Released
- Tuesday, November 3, 2020
- Thursday, November 5, 2020
- Tuesday, November 10, 2020
- Thursday, November 12, 2020
  - 🖉: Problem Set 5 Due
  - ✎: Problem Set 6 Released
- Tuesday, November 17, 2020
- Thursday, November 19, 2020
- Tuesday, November 24, 2020
- Thursday, November 26, 2020
  - 🖉: Problem Set 6 Due
  - ✎: Problem Set 6 Released
- Tuesday, December 1, 2020 [Last Class]
- Thursday, December 3, 2020 [After Term]
- Tuesday, December 8, 2020 [After Term]
  - 🖉: Problem Set 7 Due