From Natural Language to Data Science

- **Course codes:** COMP 345 & LING 345 (Winter 2025)
- **Instructors:**
  - Siva Reddy (https://sivareddy.in/)[office hours: Tuesdays 1:15pm–2:30pm ENGMC 104N, starting Jan 14th]
  - Morgan Sonderegger (https://people.linguistics.mcgill.ca/~morgan/) [office hours: TBA, starting late February]
- **Teaching Assistants:** Parishad BehnamGhader, Mehar Bhatia, Gaurav Kamath, Arkil Patel, Gandharv Patil
- **Classroom:** Macdonald-Harrington Building G-10 (MDHAR G-10)
- **Time:** Tuesdays and Thursdays, 11:35 am – 12:55 pm
- **Links:** MyCourses: announcements, slides (https://mycourses2.mcgill.ca/d2l/home/764854)

Description

Course Format: The course format is in person unless notified.

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.

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 thesauruses, search for information, and predict outcomes from social media text while also studying the ethical implications. An important side goals is to have fun with text!

Students will

- Learn to visualize text.
- Learn how to analyze language data using simple statistical and machine learning 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.

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.

Grading

There are six problem sets and no exam.

- Warm-Up Problem Sets 1 and 2 (20%): These will be a warm-up to get used to implementing Python and submitting (10% each).
- Problem Sets (80%): 4 problem sets equally weighted (20% each).

Because they are the primary method of assessment, **successful completion of each of problem sets 3-6 is required to pass the course**. In other words, if you do not submit or receive an F on one of these problem sets, you may fail the course.

The details above are subject to change. If there are changes to the means of assessment after the add/drop date, you will have the option to decide whether the original or modified means of assessment will determine your final grade.

Topics (Tentative)

Language data and applications

1. Data (Web documents, Reviews, Social Networks)
2. Applications (Information retrieval/extraction, Sentiment analysis, Recommendation systems)

Searching through data

1. Regular expressions and corpus query language
2. Symbolic and distributional representations
3. Tree regular expressions and semantic regular expressions
4. Hashing

How to make sense of data

1. Mutual information, collocation
2. Keywords
3. Vector space models
4. Distributionally similar words
5. Topic models

Language Modeling

1. Spell checkers
2. Detecting fake content
3. Language generation

Language to decisions

1. Feature-based models (logistic regression)
2. Black-box models (neural)
3. Sentiment analysis
4. Robustness of models (build it and break it)

Information Retrieval

1. Indexing
2. Pagerank
3. Reading comprehension systems

Information Extraction

1. Knowledge representation
2. Knowledge bases
3. Question answering

Social Networks (Twitter and Facebook data)

1. Representations
2. Search through structured data
3. Applications

Ethics

1. Biases in data
2. Biases in machine learning models
3. Applications and ethical questions

Data analysis: supervised

1. Classification
2. Regression
3. Ensenble methods

Data analysis: unsupervised

1. Probabilistic clustering
2. Hierarchical clustering
3. Dimensionality reduction

Speech and information

1. Information theory
2. Speech technology: overview
3. Speech data science

Schedule

Lecture	Date	Topic	Instructor	Notes
1	Jan 7	Introduction	Reddy, Sonderegger	
2	Jan 9	Regular Expressions	Reddy	Assignment 1 (Reddy) – Python basics – 10% Exercise (https://regex.sketchengine.eu/basic-exercises.html), Advanced (https://regex.sketchengine.eu/advanced-exercises.html)
3	Jan 14	Keywords, Association metrics	Reddy	Add or drop deadline
4	Jan 16	Vector space model	Reddy	
5	Jan 18	Vector space model, LSA	Reddy	
6	Jan 21	LSA / Word embeddings	Reddy	
7	Jan 23	Compositionality / Sentence Representations	Reddy	Assignment 1 due Assignment 2 (Reddy) – Corpus Query Language – 10%
8	Jan 28	Document Representation / Topic Models / Contextuality	Reddy	
9	Feb 30	Contextuality	Reddy	
10	Feb 4	Language Modeling	Reddy	
11	Feb 6	Language Modeling	Reddy	Assignment 2 due Assignment 3 (Reddy) – Vector space model, topic models - 20%
12	Feb 11	Dialogue Systems/Semantic Parsing	Reddy	
13	Feb 13	Ethics and bias	Reddy	
14	Feb 18	Neural Networks	Reddy	

Lecture	Date	Topic	Instructor	Notes
15	Feb 20	Neural Networks	Reddy	Assignment 3 due Assignment 4 (Reddy) – language modeling, semantic parsing, information retrieval, and bias – 20%
16	Feb 25	Classification and Regression Models	Sonderegger	
17	Feb 27	Classification and Regression Models	Sonderegger	
18	Mar 4	Reading week		Reading week
19	Mar 6	Reading week		Reading week
20	Mar 11	Classification and Regression Models	Sonderegger	
21	Mar 13	Classification and Regression Models	Sonderegger	Assignment 4 due Assignment 5 (Sonderegger) – Data analysis and regression – 20%
22	Mar 18	Classification and Regression Models	Sonderegger	
23	Mar 20	Unsupervised learning	Sonderegger	
24	Mar 25	Unsupervised learning	Sonderegger	
25	Mar 27	Unsupervised learning	Sonderegger	Assignment 5 due Assignment 6 (Sonderegger) – clustering, language phylogeny – 20%
26	Apr 1	Information Theory	Sonderegger	
27	Apr 3	Dimensionality Reduction	Sonderegger	
28	Apr 8	Speech Technology	Sonderegger	
29	Apr 10	Data Science for Speech	Sonderegger	Assignment 6 due

Generative AI Policy

If you use any Generative AI tool for your submitted work (e.g., ChatGPT, Github Copilot, Claude), you must cite it and submit a detailed statement describing its use, as well as a log of the chat where you used it.

You may not use AI tools for:

- Copying AI-generated prose or non-trivial code chunks – this is plagiarism
- Writing whole assignments or code files
- Writing large chunks of an assignment or code
- Using AI without citing it in your assignment

Inappropriate use of Generative AI may result in penalties on grades or referral to disciplinary authorities. If you have any question about appropriate use of AI applications for course work, please contact the instructors in their office hours.

Language of Submission

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.

Academic Integrity

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)

Inclusivity

As the instructors of this course we endeavor to provide an inclusive learning environment. However, if you experience barriers to learning in this course, do not hesitate to discuss them with one of us, or the Student Accessibility and Achievement office. Note that accommodation requests for a given problem set must be made before the problem set is due.

Extraordinary Circumstances

In the event of extraordinary circumstances beyond the control of McGill University, assessment tasks in a course are subject to change, provided students are sent adequate and timely communications regarding the change.

Tags: Winter 2025

Categories: Teaching

Updated: January 4, 2025