Introduction to Literary Text Mining

LLCU 255 – Winter 2023
M/W 10:05 - 11:25, Rm. 491 680 Sherbrooke

Professor Andrew Piper
Office: Rm 484, 680 Sherbrooke
Phone: 514-398-4400 x094504
Email: andrew.piper@mcgill.ca
Office Hours: M 3-4 pm, W 11:30-12:30 pm

Course Description
This course will serve as an introduction to the new tools and techniques being developed to study literature and culture at a vastly greater scale. How does the ability to analyze several hundred to hundreds of thousands of documents give us new insights into the history of literature and culture? This course will introduce you to the major concepts and methods of text mining and natural language processing and the ways in which they are being applied to the study of literature. You will learn how to computationally study large numbers of documents as well as structure experiments in order to better understand the distinctive behavior of different kinds of literary documents. Weekly assignments will introduce you to the R software environment and will culminate in a final project of your choosing. No prior programming experience is required.

Reading List

All readings are available through myCourses or links in the syllabus. All code is available at the following repository.

Weekly Assignments

Wk. 1 01.04 No Class (Covid)

Wk. 2 01.09 No Class (Covid)

01.11 *Course Intro: What is Text Mining?*

Wk. 3 01.16 *What's it for? Some examples*
- Daniels, "The Largest Vocabulary in Hip Hop"
- Shendruk, "Analyzing the Gender Representation of 34,476 Comic Book Characters"

01.18 *Introduction to Literary Modeling*
- Lancashire, "Vocabulary Change in Agatha Christie"
- Christie, Excerpts from two novels

Wk. 4 01.23 *Introduction to R*
- R tutorial on udemy (Sections 1 & 2)

01.25 *Working with text data: The* TM *library*

Wk. 5 01.30 *Corpus Comparison*
- Piper/So, "Women Write About Family, Men Write About War"
- Introduction to hypothesis testing (VIDEO, 8 MINS)

02.01 *Discovering distinctive features*
- **Assignment #1 Due**

Wk. 6 02.06 *What are Vector Space Models and what can you do with them?*
- Jurafsky, "Vector Semantics", pp. 1-10

02.08 *Clustering Documents and Word Embeddings*
- **Assignment #2 Due**

Wk. 7 02.13 *Learning about machine learning*
- Introduction to machine learning (VIDEO, 17 MINS)

02.15 *SVMs and Random Forests*
- **Assignment #3 Due**

Wk. 8 02.20 *Sentiment Analysis: Emotional Arcs*
- Vonnegut, "The Simple Shapes of Stories" (VIDEO, 5 MINS)
- Jockers, "Revealing Sentiment and Plot Arcs with the Syuzhet Package."
- Piper/So, "Quantifying the Weepy Bestseller"

02.22 *Sentiment Analysis in R*

***** Study Break *****

Wk. 9 03.06 *Topic Modeling*
- Sarraf/Chen, "Queer Fans: The Difference Fanfiction Makes"

03.08 *The* topicmodels *library in R*
- **Assignment #4 Due**

Wk.10 03.13 *Understanding Characters*
- Kraicer, "Social Characters"

03.15 *Studying Characters using NER & bookNLP*

Wk.11 03.20 *Literary Spaces*
- Moretti, *Atlas of the European Novel* (pp. 12-29)

03.22 *Analyzing literary space using NER and bookNLP*
- **Final Project: Project Description Due**

Wk.12 03.27 *Data visualisation*

	03.29	*Data visualisation* **- Final Project: Data Description Due**
Wk.13	04.03	*Final Projects* **- Final Project: Model Description Due**
	04.05	*Final Projects*
Wk.14	04.12	*Final Projects*

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 http://www.mcgill.ca/integrity/ for more information).

Course Requirements

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.

Reading Assignments	10%
Programming Assignments (4x)	40%
Final Paper (7-8 pp.)	50%

Reading Assignments. During a week when there is an assigned reading, you will be askd to submit a one page 3-2-1 reading assignment document for **one** reading that week. "3-2-1s" involve documenting three key issues in the reading, two things you found unclear, and one big question you want to ask the author. A 3-2-1 guide will be available on myCourses.

Programming Assignments. Programming assignments are designed to introduce you to the R software environment for text analysis. Assignments will vary between implementing existing code to exploring data and measurements in the form of 1-2 pp papers. In each case you will be provided with a choice of data sets and a particular script which you will learn how to "tune." The aim of these assignments is to give you a hands-on understanding of how computational analysis works and how to critically analyze your results.

Final Paper. The final paper will consist of the following steps: a) design an experimental study; b) choose your data; c) implement one or more R scripts for analysis; d) write a detailed and thoughtful engagement with your results. The aim of this paper is to have you work through the entire analytical process, from the choice of appropriate data, the relevance of your analytical techniques, to the potential significance of your findings. What data did you choose to work with and why? What has your method told you about your texts? What challenges did you encounter? What do you remain uncertain about? Why is this an important question to be asking in the first place? As with the weekly assignments you may choose an existing data set or create your own. To faclitate completion of the project you will be required to hand in 1-page statements for the

three stages of the project (1. project description; 2. data description; 3. model description). These are mandatory for submission of the final paper. Late or unsubmitted statements will reduce your final paper grade by 1/3 of a grade.

Late papers will lose 1/3-grade for every day late. Students who receive a grade of D,F, or J will not be allowed to do supplemental work. All papers will be submitted to the text-matching software per university policy. Three or more missed classes will result in a lowering of the student's overall grade. According to Senate regulations, instructors are not permitted to make special arrangements for final exams. Please consult the Calendar, section 4.7.2.1, General University Information and Regulations at www.mcgill.ca. In the event of extraordinary circumstances beyond the University's control, the content and/or evaluation scheme in this course is subject to change. © Instructor generated course materials (e.g., handouts, notes, summaries, exam questions, 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.