

Big Data and Literature: Introduction to Literary Text Mining

LLCU 255 – Fall 2014

T/Th 10:05 - 11:25, Rm. 295 688 Sherbrooke

Optional Laboratory: Th 2:35-4 pm, Leacock 212

Professor Andrew Piper

Office: Rm 483, 688 Sherbrooke

Phone: 514-398-4400 x094504

Email: andrew.piper@mcgill.ca

Office Hours: T 2.30 – 3.30 pm, Th 1 – 2 pm

Course Description

This course will serve as an introduction into the new tools and techniques being developed to study literature at a vastly greater scale. How does the ability to analyze hundreds of thousands of texts give us new insights into the history of literature and culture? How might thinking about literature as data change our understanding of foundational literary categories like authorship, textuality, narration, plot, character or even language? In order to address these questions, we will move between the close study of a sample text, Virginia Woolf's *To the Lighthouse*, which will serve as our literary handbook for the semester, and techniques that allow for a broad-scale understanding of literature using computational analysis. In doing so, we will have three main goals: a) introduce you to the basic concepts and practices of text mining (sentiment analysis, topic modeling, social network analysis, etc); b) introduce you to the major categories of literary analysis (discourse, narration, plot, character, setting and space); and c) allow you the opportunity to develop your own experimental projects. A voluntary lab will be held each week in which you will learn how to program in R, the most popular text analysis platform. No previous programming experience is required, just enthusiasm and curiosity.

Reading List

Please purchase the following, which is available at the McGill Book Store:

Virginia Woolf, *To the Lighthouse*

Matthew L. Jockers, *Text Analysis with R for Students of Literature* (optional, for Lab only)

Weekly Assignments

Wk. 1 09.02 *What is Text Mining?*

09.04 *What's it for? Some examples.*

- Lancashire, "Vocabulary Change in Agatha Christie."

- Matt Daniels, "The Largest Vocabulary in Hip Hop":

<http://rappers.mdaniels.com.s3-website-us-east-1.amazonaws.com>

- Michel, "Quantitative Analysis of Culture Using Millions of Digitized Books."

- Wk. 7** 10.14 *Sentiment Analysis* **Woolf, Time Passes, 1-10**
 - The Hedonometer: hedonometer.org
 - Jockers, "A Novel Method for Detecting Plot."
<http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/>
 - Liu, "Sentiment Analysis and Subjectivity."
****Mid-Term Paper due in class**
- 10.16 *Sentiment Analysis*
- Wk. 8** 10.21 *Topic Modeling* **Woolf, The Lighthouse, 1-4**
 - Mohr, "Topic Models: What they are and why they matter."
 - Tangherlini, "Trawling in the Sea of the Great Unread."
- 10.23 *Topic Modeling*
 - Jockers, "Significant Themes in 19C Literature."
 - Underwood/Goldstone, "The Quiet Transformations of Literary Studies."
- Wk. 9** 10.28 *Intertextuality* **Woolf, The Lighthouse, 5-8**
 - Genette, Palimpsests, pp. 1-13.
 - Barthes, S/Z, pp. 3-20.
 - Kristeva, "Word, Dialogue, Novel."
- 10.30 *Intertextuality*
 - Cordell, "Infectious Texts."
 - So/Long, "Network Analysis and the Sociology of Modernism."
- Wk.10** 11.04 *Intertextuality* **Woolf, The Lighthouse, 9-13**
 - Case Study: *To the Lighthouse*.
- 11.06 *Social Networks*
 - Newmann, Networks, "Introduction."
 - Extra Credit: Newmann, chap. 4.
- Wk.11** 11.11 *Social Networks*
 - Piper, "Communities of Detection: Social Network Analysis and Detective Fiction."
 - Elson, "Extracting social networks from literary fiction."
- 11.13 *Social Networks*
- Wk.12** 11.18 *Geospatial Analysis*
 - Wilkens, "The Geographic Imagination of Civil War Era American Fiction."
 - Moretti, Atlas of the European Novel

- 11.20 *Visual Language Processing*
- Piper, "Footnote Detection."
- Wk.13** 11.25 *The End of Books?*
- Robert Coover, "The End of Books," pp. 705-710.
- 11.27 *Books Without End*
- Sinclair/Rockwell, "Ubiquitous Text Analysis."
****Final Paper due December 2 by 4 pm**

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.

Class Participation	15%
Mark-Up Project	15%
Midterm Paper (4-5 pp.)	25%
Final Paper (9-10 pp.)	45%

Class Participation. You are expected to attend every class and actively participate in class discussions with observations and questions derived from close and thoughtful reading of each weeks' texts, both primary and secondary. Our aim is to engage critically with existing quantitative studies of literature and to engage creatively with our primary text to develop new avenues for research.

Lab (Optional). Lab attendance is not required, but recommended. You will learn how to use the R statistical platform to analyze large collections of texts. This will allow you to conduct your own research, but also better understand the conditions and limitations of this kind of text analysis more broadly. No programming background is necessary, just a willingness to learn and experiment.

Mark-Up Project. For this assignment, you will be responsible for choosing a particular literary feature and identifying it throughout our primary novel, Virginia Woolf's *To the Lighthouse*, using an online mark-up system. The feature can be based on a particular week's reading, or it can be something you have identified on your own. As part of the assignment, you will present the feature you have chosen, why it is you have chosen it, and what problems it poses for measurement. The aim of this exercise is to have you think about issues surrounding the identification and measurement of literary features. The final submission will consist of a fully marked-up version of the novel which will be integrated with the other members of the class

which can be shared with researchers in the field. Our goal is to create a digital critical edition of this novel. You may choose to integrate this assignment into your final paper.

Mid-Term Paper. The mid-term paper will consist of a close analysis of a selected passage of *To the Lighthouse* according to one of the theoretical criteria explored in the first half of the semester (style, diction, point of view, setting, time and duration). The aim of this exercise is to hone your skills of literary analysis, the necessary basis of all large-scale analysis. It is also an effort in teaching you how to write in a concise and pointed way.

Final Paper. For the final paper you have two options: a) write a critical review of a particular study we've encountered this semester or b) undertake your own study using the tools learned in lab. The aim of this paper is to have you think through the entire research process, from data collection and preparation to the interpretation of results, whether your own or those in a study we've read.

Late papers will lose a half-grade for every class 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.