Big Data and Literature: 
Introduction to Literary Text Mining
LLCU 255 – Fall 2015
M/W 2:35 - 3:55, Rm. 295 688 Sherbrooke

Professor Andrew Piper
Office: Rm 484, 688 Sherbrooke
Phone: 514-398-4400 x094504
Email: andrew.piper@mcgill.ca
Office Hours: M 4 - 5 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 at a vastly greater scale. How does the ability to analyze several hundred to 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 author, text, work, narrative, plot, character or even language? In order to address these questions, this course will introduce you to the basic concepts and practices of text mining (stylometry, distributional semantics, sentiment analysis, topic modeling, and social network analysis) and the ways in which they have been applied to the study of literature. Weekly assignments will introduce you to the R software environment and will culminate in a final project of your own choosing. No prior programming experience is required.

Reading List
All readings are available through myCourses.

Weekly Assignments

Wk. 1 09.04 What is Text Mining?
09.09 What’s it for? Some examples.
   - Lancashire, “Vocabulary Change in Agatha Christie.”
   - Underwood, “Seven Ways Humanists are Using Computers to Understand Text.”

Wk. 2 09.14 How to model a text (Vector Space Models)
   - Turney et al. “From Frequency to Meaning: Vector Space Models of Semantics.”

09.16 How to test for significance (Regression, analysis of variance, predictive modeling)

**Assignment 1:** Create a vector space model in R of the English Novel Data Set.

**Wk. 3 09.21  Style**
- Jannidis, “Burrows Delta and its Use in German Literary History.”

**09.23  Style**
- Eder et al. “Stylometry with R.”

**Assignment 2:** Apply the stylo package in R to a corpus of your choice.

**Wk. 4 09.28  Diction**
- Underwood/Sellers, “The Emergence of Literary Diction.”

**09.30  Diction (Distributional Semantics)**
- Piper, “The Unbearable Nostalgia of High Culture: On Prizewinning Novels.”

**Assignment 3:** Analyze a dictionary of your choice in R on a selected corpus.

**Wk. 5 10.05  Plot: An Intro to Narrative Analysis**
- Genette, Narrative Discourse, Chap. 5
- Bal, Narratology

**10.07  Plot: Sentimental Arcs (Sujet)**

**No Assignment**

**Wk. 6 10.12  Thanksgiving: No Class**

**10.14  Plot: Character Frames (Fabula)**
- Elsner, “Character-Based Kernels.”
- Piper, “Detecting Narrative Frames using Characters.”

**No Assignment**

**Wk. 7 10.19  Plot: Language Frames**
- Piper, “Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel.”

10.21  Plot: Review

Assignment 4: Use subject package, sentiment or character dictionaries in R to analyze narrative frames of a single work or compare two works.

Wk. 8 10.26  Topoi: Topic Modeling
- Mohr, “Topic Models: What they are and why they matter.”
- Underwood, Goldstone, “The Quiet Transformations of Literary Studies.”
- Jockers, “Significant Themes in 19C Literature.”

10.28  Topic Modeling in R

No Assignment

Wk. 9 11.02  Intertextuality: Topics, Quotations, Memes, Vectors
- Cordell, “Virtual Textuality.”
- Leskovec, “Meme-Tracking.” Website.
- - Commonplace Cultures.

11.04  Intertextuality (cont’d)
- Piper/Algee-Hewitt, “The Werther Effect 1.”

Assignment 5: Use the topicmodels package in R to analyze a corpus of your choice.

Wk.10 11.09  Networks: An Introduction
- Newmann, Networks, “Introduction” and Chap. 4

11.11  Citation and Publication Networks
- Healy, “Gender and Citation in Philosophy.”
- So/Long, “Network Analysis and the Sociology of Modernism.”

No Assignment

Wk.11 11.16  Reader and Character Networks
- Piper et al., “Communities of Detection: Social Network Analysis and Detective Fiction.”

11.18  Networks: Review

Assignment 6: Use the iGraph for R package to study the social network of a short story.
Wk.12 11.23  
*Geo-Space*
- Moretti, *Atlas of the European Novel*

11.25  
*Page-Space*
- Piper, “Footnote Detection.”

Wk.13 11.30  Review of Final Projects

12.02  Review of Final Projects

Wk.14 12.07  
- Sinclair/Rockwell, “Ubiquitous Text Analysis.”

**Final Paper due as a hard copy on December 11 by 4 pm in Room 425**

**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/](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%
- **Weekly Assignments (6x)** 40%
- **Final Paper (6-8 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. Our aim is to engage critically with existing quantitative studies of literature and to think creatively about new ways of understanding texts.

**Weekly Assignments.** Weekly assignments are designed to introduce you to using the R software environment for text analysis. You will move from the straightforward implementation of existing scripts to the analysis of results in the form of a 1-2 pp paper. In each case you will be provided with a choice of data sets and a particular script which you will learn how to “tune.” You are free, indeed encouraged, to construct your own data sets. The aim of these assignments is to give you a hands-on understanding of how computational analysis works and the in’s and out’s of critically analyzing results.
**Final Paper.** The final paper will consist of the following steps: a) design of an experimental study; b) choice of data; c) implementation of 1 or more R scripts for analysis; d) 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 one of your own.

**Weekly Tutorial (Optional).** A trouble-shooting tutorial for using R will be held every week at a specified time. This is an opportunity for you to improve your R programming skills.

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.