Big Data and Literature: Introduction to Literary Text Mining

LLCU 255 – Fall 2016
M/W 11:35 - 12:55, BURN 1B23

Professor Andrew Piper
Office: Rm 484, 688 Sherbrooke
Phone: 514-398-4400 x094504
Email: andrew.piper@mcgill.ca
Office Hours: M 1:30 - 2:30 pm, W 2 - 3 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 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 (vector space models, 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 09.02 *What is Text Mining?*
- Matt Daniels, "The Largest Vocabulary in Hip Hop." Poster. http://poly-graph.co/vocabulary.html

09.07 *What's it for? Some examples.*
- Lancashire, "Vocabulary Change in Agatha Christie."
- Piper/So, "Quantifying the Weepy Best-Seller." https://newrepublic.com/article/126123/quantifying-weepy-bestseller
- Leskovec, "Meme-Tracking." http://www.memetracker.org
- *Extra Credit*: Michel et al., "Quantitative Analysis of Culture Using Millions of Digitized Books."

Wk. 2 09.12 *Feature Selection*
- Piper, "Fictionality."

09.14 *Feature Comparison (t.test, rank-sum test, fisher's exact test)*

Assignment 1: Compare a single feature between two corpuses of your choice.

Wk. 3 09.19 *Vector Space Models*
- Turney et al.

09.21 *Clustering*

Assignment 2: Create a vector space model in R of a selected data set using the Stylo package in R.

Wk. 4 09.26 *Machine Learning*
- Intro to Machine Learning in R

09.28 *Classification*

Wk. 5 10.03 *Genre*
- Underwood, "The Life Cycles of Genres." http://culturalanalytics.org/2016/05/the-life-cycles-of-genres/

10.05 *Genre*
- Piper, "How Cultural Capital Works." http://post45.research.yale.edu/2016/05/how-cultural-capital-works-prizewinning-novels-bestsellers-and-the-time-of-reading/

Assignment 3: Build a classifier using the kernal package in R and apply it to a corpus of your choice.

Wk. 6 10.10 **Thanksgiving: No Class**

10.12 *Narrative Structure 1: Story Arcs and Sentiment Analysis*
- Reagan et al, "The emotional arcs of stories are dominated by six basic shapes."
- Media Coverage: http://www.theatlantic.com/technology/archive/2016/07/the-six-main-arcs-in-storytelling-identified-by-a-computer/490733/
- Rebuttal: http://sappingattention.blogspot.ca/2016/07/plot-arceology-emotion-and-tension.html

Wk. 7 10.17 *Narrative Structure 2: Social Networks*
- Healy, "Using Metadata to find Paul Revere." https://kieranhealy.org/blog/archives/2013/06/09/using-metadata-to-find-paul-revere/
- Sack, "Character Networks for Narrative Generation."

10.19 *Regression Analysis*

Wk. 8 10.24 *Introduction to Social Network Analysis*

- Newmann, Networks, “Introduction” and Chap. 4

10.26 *iGraph for R*

Assignment 4: Use the iGraph for R package to study the social networks of a selected group of texts.

Wk. 9 10.31 *Publication and Citation Networks*

- So/Long, “Network Analysis and the Sociology of Modernism.”
- Healy, “Gender and Citation in Philosophy.” https://kieranhealy.org/blog/archives/2015/02/25/gender-and-citation-in-four-general-interest-philosophy-journals-1993-2013/

11.02 *iGraph for R*

Wk.10 11.07 *Topic Modeling*

- Mohr, “Topic Models: What they are and why they matter.”
- Underwood, “Topic Modeling Made Just Simple Enough.”
- Jockers, “Significant Themes in 19C Literature.”

11.09 *Topic Modeling in R*

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

Wk.11 11.14 *Geo-Space*

- Wilkens, “The Geographic Imagination of Civil War Era American Fiction.”

11.16 *Page-Space*

- Piper, “Footnote Detection.” http://txtlab.org/?p=395

Wk.12 11.21 *Model Building*

11.23 *Model Building*

Wk.13 11.28 Review of Final Projects

11.30 Review of Final Projects

Wk.14 12.05 *Final Discussion: The End of Books? Books Without End.*

- Sinclair/Rockwell, “Ubiquitous Text Analysis.”

****Final Paper due as a hard copy on December 12 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/ 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 (5x)	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 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 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 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.