Introduction to Critical Text Mining

Instructor Mark Algee-Hewitt

Time: T/Th 8:00-9:50

E-Mail mark.algee-hewitt@stanford.edu

Description
Text mining, the practice of using computational and statistical analysis on large collections of digitized text, is becoming an increasingly important way of extracting meaning from writing. Whether working on survey data, medical records, political speeches or even digitized collections of historical writing, we are now able to use the power of computational algorithms to extract patterns from vast quantities of textual data. This technique gives us information we could never access by simply reading the texts. But determining which patterns have meaning and which answer key questions about our data is a difficult task, requiring not only knowledge of statistics and computer science, but a critical approach to analyzing the graphs, charts and numbers that we receive as output from a quantitative analysis of text. In this course, we will explore the different methods through which text mining can be used to “read” text in new ways: including authorship attribution, sentiment analysis, genre studies and named entity extraction. While students will learn the basic skills of text mining, including developing appropriate research questions, assembling and cleaning textual corpora and employing different quantitative methodologies, our focus will be on the analysis and interpretation of our results. How do we formulate research questions and hypothesis about text that can be answered quantitatively? Which methods fit particular needs best? And how can we use the numerical output of a text analysis to explain features of the texts in ways that make sense to a wider audience?
This course will combine lectures, discussion groups and hands-on lab work in the Stanford Literary Lab to explore these questions and introduce students to the practice of literary and critical text mining.

Course Layout
This course will be divided between the classroom and the Literary Lab, with one class per week in each setting. In the classroom, students will be exposed to the various methods, procedures, outcomes and challenges presented by literary text mining through lectures on and discussions of key concepts in the emerging field of Digital Humanities. The lab component will involve practical hands-on approaches to the methods that we discuss in class as students will have the opportunity to use the various tools and example corpora to design research experiments in quantitative textual analysis.

Course Outcomes
Students will learn to compare the results from a quantitative analysis of a large-scale corpus with a critical reading of a work from that corpus. We will emphasize ways that we can learn about a single text by analyzing a corpus of many texts and how our reading of a single text sheds light on an analysis of an entire corpus. Moreover, we will learn how to do this in a methodologically sound way that will still be accessible to the methods and strategies of both statistics and literary criticism. Students will also learn to design thoughtful and meaningful quantitative tests of textual corpora in order to answer specific research questions, including
assessing the results of these tests for significance. Finally, students will gain new skills in critical thinking by assessing the meaning of quantitative results using both numerical and visual information.

Course Texts
Edward Tuft, *The Visual Display of Quantitative Information*
Franco Moretti, *Graphs, Maps and Trees*
Matt Jockers, *Macroanalysis*
*An Introduction to R* (Available online)

Software Required (either PC, Mac or Linux)
Instructions will be given during the first class on how to obtain and install the following software/packages.
The R software environment for statistical computing (open source)
www.r-project.org
Assorted packages for R: TM, stylo, ggplot2, topicmodels, klaR

Work and Assignments:

1. Participation (online and in class/lab) 20%
2. Short Assignments (1 per week) 50%
3. Final Project 30%

Participation
As this class is split between discussions of the methodologies and hands-on explorations of these methods, you are all tasked with keeping the spirit of experimentation alive. This is another way of saying that participation is mandatory: your voice must be heard in class contributing, questioning or challenging or in the lab as we work together or separately to learn the techniques of literary quantitative analysis.

Short Projects
While the goal of this class is to explore the ways in which quantitative analysis can assist the study of textual or literary material, a prerequisite of this is your ability to use many of the new techniques we are studying to do basic corpus analyses. Lab time will be devoted to learning the basic programing and statistics in R that will enable you to do this and each week you will receive a very short assignment based on what we have covered in class or in lab for you to do on your own for a total of 50% of your grade. These assignments will help mark your progress and formalize the skills we learn in class.

Final Project
In your final project, you will combine the theoretical knowledge of how the digital humanities can offer critical insights to literary/textual problems with your hands-on knowledge of text analysis in R to perform your own analysis/critical reading of the class corpus. This project will require you to perform, interpret and write up a quantitative analysis: in particular, you will extract critical meaning from the results of your digital work. More details will be given in the formal project assignment.
Syllabus

Introduction
Class 1: Why do we mine? Reading vs Quantitative Analysis
Class 2: Introduction to the Lab / Programming basics
Text: Jockers, Chapter 2; Selections from An Introduction to R

Fundamentals of Text Mining
Class 3: Building a Corpus
Text: David Berry, “The Esthetics of Hidden Things”
Class 4: Practical Lab on Corpus Building (Cleaning and Tagging)
Class 5: Research Question Design
Text: Jockers, Chapter 6
Class 6: Introduction to Statistics in the Lab
Text: An Introduction to R

Text Mining Strategies
Class 7: Authorship Attribution and Forensic Authorship Analysis
Class 8: Lab – Authorship Attribution Tests: the Stylo Package, Burrows’ Delta
Class 9: Frequency Analysis
Text: Selections from Moretti Graphs, Maps and Trees
Class 10: Lab – Frequency Analysis and Genre
Class 11: Topic Modeling (Advanced Frequency Analysis)
Text: Berry, Analysis Tool or Research Methodology: Is There an Epistemology for Patterns?
Class 12: Lab – Topic Modeling
Class 13: Names, Places and Times
Text: Jockers, Chapter 8
Class 14: Lab – Classification; Variable Selection and Discriminant Function Analysis

Beyond the Numbers: Visualizing the Results
Class 15: Graphs and Plots
Text: Tufte
Class 16: Lab – Extracting Meaning from Visualizations
Class 17: Maps and Topographies
   Text: Algee-Hewitt, *Reading Topographically*

Class 17: Lab – Alternate Visualizations

Class 18: Networks and Relationality
   Text: Tufte

Class 19: Lab – Network Creation and Analysis

**Conclusion**

Class 20: Synthesis: Reading and Reading Quantitatively