Literary Text Mining
An Introduction to Quantitative Text Analysis

Instructor Mark Algee-Hewitt

Couse Description:

This course will allow students to explore a variety of applied methods for computationally and statistically analyzing texts for humanities research by introducing them to both the available tools and their underlying practices that are fundamental to this area of digital humanities research. Strategies such as text mining, content analysis, sentiment analysis and entity extraction are becoming fundamental to research in the humanities, especially as they applied to large and diverse digital corpora. Equally important, however, is the recognition of the limits of these methods and the need to integrate them within a holistic approach to humanities inquiry. The skills students will gain will include basic programming for textual analysis, applied statistical evaluation of results and the ability to present these results within a formal research paper or presentation. Students will learn to recognize patterns within their data, test the significance of these patterns and explain this significance within the context of humanities research. As an introduction, students in this course will also learn the prerequisite steps of such an analysis including corpus selection and cleaning, metadata collection, and selecting and creating an appropriate visualization for the results.

Course Layout

Class time 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

By the end of the course, committed students will be able to demonstrate their technical knowledge of a variety of digital textual analysis methods, describe the differences between these methods, identify appropriate use cases for each method discussed, and, most importantly demonstrate both their ability to apply these methods to humanities-based research questions and describe the humanities implications of their computational analysis. Students will also be able to generate and analyze meaningful visualizations of their data and describe, in detail, the methodological foundations that underlie the tools that we will explore during the course.

Course Texts

Edward Tufte, *The Visual Display of Quantitative Information*
Franco Moretti, *Graphs, Maps and Trees*
Matt Jockers, *Macroanalysis*
Taylor Arnold and Lauren Tilton, *Humanities Data in R*
An Introduction to R (Available online)
Selections from: Dawn Archer *What's in a Word-List? Investigating Word Frequency and Keyword Extraction* (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
RStudio software (https://www.rstudio.com/)
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, Chaper 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
Text: Archer, "Word Frequency, Statistical Stylistics and Authorship Attribution"

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 Logistic Regression

Beyond the Numbers: Visualizing the Results

Class 15 Graphs and Plots—Selecting and Designing the right Chart
Text: Tufte

Class 16 Extracting Meaning from Visualizations

Class 17 Natural Language Processing
Text: Matthew Wilkens, “Canons, Close Reading, and the Evolution of Method.”

Class 19 Lab – Part of Speech Tagging, Named Entity Recognition

Class 19 Networks and Relationality
Text: Tufte

Class 20 Lab – Network Creation and Analysis