**English 184E**
**Literary Text Mining**
**An Introduction to Quantitative Text Analysis**

**Instructor** Mark Algee-Hewitt

**Time/Location** T/Th 10:30-12:20; Lathrop 290

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

**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 need to integrate these methods within a holistic approach to humanities inquiry. This class will therefore balance its approach between learning the details of text mining methods, and exploring how they can be effectively used in humanities research.

   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 alternate between discussion and lab work.  Rather than a strict day-to-day division, all classes will involve both aspects as we discuss methods and approaches and then try them out together. This means that you will be responsible both for completing the readings assigned for each class, and for coming prepared to experiment with the methods we will discuss.

**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, *Distant Reading*
   Taylor Arnold and Lauren Tilton, *Humanities Data in R*
   *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
   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) | 25% |
| 2. | Short Assignments (1 per week) | 50% |
| 3. | Final Project | 25% |

**Participation**
   As this class will interweave discussions of the methodologies with 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**

**Preparation**

Sept 26: The quantitative and the qualitative: what is literary data?
The Basics of Programming in R

Sept 28: Reading and Distant Reading
Text: Selections from Moretti's *Distant Reading*
Reading and cleaning the text

**The Fundamentals of Text Mining**

Oct 3: Single Text Analysis 1
Dispersion plots and the text as visual object
Text: Algee-Hewitt and Piper "The Werther Effect"

Oct 5: Single Text Analysis 2
**K**ey **W**ords **I**n **C**ontext: searching and sorting
Text: Algee-Hewitt and Piper "The Werther Effect"

Oct 10: Frequency and Counting 1
The text as object and the (power) laws of text

Oct 12: Frequency and Counting 2
The multidimensional corpus
Text: Moretti "Style Inc" (in *Distant Reading*)

**Working with Data**

Oct 17: Metadata and the corpus
Tables and filenames
Text: Tufte, *The Visual Display of Quantitative Information*

Oct 19: The aesthetics of Data
Visualization: the grammar of graphics
Text: Tufte (continued)

**Style and Genre**

Oct 24: Authorship
MFW and the meaning of the author
Text: Foucault "What is an author?" (online)

Oct 26: Style
Dendrograms and the limits of authorship
Text: Jan Rybicki "The great mystery of the (almost) invisible translator

Oct 31: Statistical Frequency Analysis 1
        Distinctive words and the question of genre

Nov 2: Statistical Frequency Analysis 2
        Principal Component Analysis the display of data

**Mining and Models**

Nov 7: Modeling 1
        Topic models and the work of probability
        Text: Blei: "Probabilistic Topic Models" (online)

Nov 9: Modeling 2
        Supervision, Classification and Regression
        Text: Underwood "The Life Cycles of Genres"

Nov 14: The Text as Network 1
        Nodes, Edges and the Visualization of Graphs
        Text: Scott Weingart, "Demystifying Networks"

Nov 16: The Text as Network 2
        The mathematics of connections
        Text: Agarwal et al. "Social network analysis of Alice in Wonderland."

Nov 21 & 23    **No Classes (Thanksgiving)**

**Advanced Methods**

Nov 28: Natural Language Processing 1
        Tokens, types and dependencies
        Text: Arnold and Tilton: *Humanities Data in R*

Nov 30: Natural Language Processing 2
        Named entities and co-reference resolution
        Text: Arnold and Tilton: *Humanities Data in R*

Dec 5: Vector space 1
        The meaning of proximity

Dec 7: Vector space 2
        Word embeddings and vector math