PHIL 4500: Data and Culture -- Course Syllabus

Spring 2017, T Th 2:00-3:15. Rotunda 150.

Office hours (Humphreys, Cocke Hall 105) T 11-12, R 11-12, 3:30-4:30

Course Description

This is a new course that combines hands-on textual analysis with broader philosophical and cultural issues concerning the epistemology of data, the relations between the digital humanities and data science, and the tensions between traditional methods of the humanities and contemporary computational techniques. The course text is *Text Analysis with R for Students of Literature* by Matthew Jockers. The course is self-contained and is aimed at undergraduates in the humanities and social sciences who would like to acquire skills in these new areas. The course is restricted to third and fourth year students. We shall learn and use enough of the R programming language to write basic programs to analyze text. No previous knowledge of R is required but some familiarity with the basics of programming is expected, as is openness to discussions of interpretation.

The class is interactive and you are expected to keep up with the assigned reading and exercises. Because discussing coding techniques with others is helpful, when possible, we shall pair students having a lower level of exposure to programming with students having a higher level. We hope this will benefit both students.

Requirements

The requirements for the course are weekly exercises and a term project. The former will be based on standard text files that accompany Jockers' book and other sources. For the term project, students should choose a set of texts that interest them and produce a detailed computational analysis of those texts using the techniques taught in the class.

Required Texts

Matthew L. Jockers, 2014, *Text Analysis with R for Students of Literature*. Springer.

Schedule

Week 1: Introduction

R 19 JAN / Day 1: Introduction to the Course

Topics

- Thematic Overview
  - On Culture and data, knowledge and understanding.
  - What computers can do that humans can’t and vice versa.
  - Rationale for textual analysis.
- Technical Overview
  - Choice of language. (Why R?)
  - Introducing Jockers' book -- rationale and warnings
- Syllabus Review
  - Learning goals.
  - Structure of the course -- studio, reflection, projects.
- Getting Started
  - Download and install R and Rstudio.
  - Create class directory and add supporting materials.

Readings / Homework

- Jockers 2014, Ch. 1: 1-5

Week 2: First Foray into Text Analysis

T 24 JAN / Day 2: R and RStudio Basics

Topics

- Overview of RStudio
- Interacting with R -- doing simple math, printing results, etc.
- Creating your first file. Working with files and the file system. (Paths, names, etc.)
- Create a Project?
- R help resources.

Readings

- Jockers 2014, Ch.1: 6-7

R 26 JAN / Day 3: First Foray (i)

Topics

- Importing files from outside sources.
- Separating content from metadata.
- Stripping out punctuation.
- Vectors, strings, lists, matrices, and functions.
- Numerical and character vectors.

Reflection

- What happens to text when it become digital?
- Syntax and Semantics
- Philosophical perspectives on meaning (meaning as use, pragmatism)

Readings

- Jockers 2014, Ch.2: 1-2
- J.L. Borges, The Library of Babel.

Week 3: First Foray (cont'd)

T 31 JAN / Day 4: First Foray (ii)

Topics

- Reprocessing Content.
- Beginning the Analysis.
- Counting word types, tokens and type/token ratios.
- Locating words in vectors.
- Tables.
- Word frequencies.
- Graphing functions in R.

Readings

- Jockers 2014, Ch.2: 3-4

R 2 FEB / Day 5: Word Frequencies

Topics

- Recycling.
- Saving and reusing code.
- Zipf's Law.
- Examples of word distributions.

Reflection

- Power laws -- what do they mean?
- Do patterns need explanations?
- Positivism.

Readings

- Jockers 2014, Ch. 3
- Zipf’s word frequency law in natural language: A critical review and future directions (selections) Steven T. Piantadosi

Week 4: Word Distributions

T 7 FEB / Day 6: Word Distributions (i)

Topics

- Dispersion plots
- Using `grep()` to parse a text.
- Regular Expressions.
- Logical vectors.
- Regular expressions.

Readings

- Jockers 2014, Ch. 4:1-2

R 9 FEB / Day 7: Word Distributions (ii)

Topics

- Understanding algorithms.

- If/else routines.
- How to use `for’ loops.

Short Reflection

- Algorithmic understanding and human understanding

Readings

- Jockers 2014, Ch. 4:3
- Tiel and Latour, 1995, "The Hume Machine"

Week 5: Correlation

T 14 FEB / Day 8: Correlation (i)

Topics

- `apply()` and `do.call()` operations
- Matrices, matrix operations, and `cbind()`

Readings

- Jockers 2014, Ch. 4:4

R 16 FEB / Day 9: Correlation (ii)

Topics

- Correlation
- Data frames
- Randomization
- Histograms

Reflection

- Understanding correlation and significance
- Mutual information

Readings

- Jockers 2014, Ch. 5

- FiveThirtyEight, "Science is not Broken." (Reading on p-values.)

Week 6: Clustering

T 21 FEB / Day 10

Topics

- User-defined functions.

Readings

- TBD

R 23 FEB / Day 11

Topics

- Clustering and the Euclidean distance.
- Lists, data frames, and data matrices

Reflection

- Traditional Statistics vs Data Science
- "Machine as Horizon of Interpretation"

Readings

- Jockers 2014, ch11:6-7

Week 7: Classification

T 28 FEB / Day 12

Topics

- Dendrograms
- Text segmentation
- gsub
- Retaining possessives

Readings

- Jockers 2014 Ch 11:8-9, 12:3-5

R 2 MAR / Day 13

Topics

- Cross tabulation
- Mapping data to metadata
- Author identification

Reflection

- Machine learning

Readings

- Jockers 2014 Ch 12:6-9
- Raf. Alvarado, "Digital Humanities and the Great Project: Why We Should Operationalize Everything -- And Study Those Who Are Doing So Now."

-------- Spring Break --------

Week 8: Topic Modeling

T 14 MAR / Day 14

Topics

- What is topic modeling?
- Mallet
- Chunking texts- how big should the chunks be?

Readings

- Jockers 2014 Ch 13:1-4 with class notes*
- David M. Blei , "Topic Modeling and Digital Humanities." http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei

R 16 MAR / Day 15

Topics

- Stoplists
- CSV files.
- Training: How many iterations?
- How many topics?

Reflection

- Supervised and unsupervised learning
- What is a model?
- Understanding what topic models can and cannot do.

Readings

- Jockers Ch 13:5

Week 9: Topic Modeling continued

T 21 MAR / Day 16

Topics

- Understanding the output of topic models
- Understanding LDA.
- Enhancements to LDA (theory only).
- Wordclouds

Readings

- Jockers Ch 13:6-7

R 23 MAR / Day 17

Topics

- Topic probabilities
- The pros and cons of bags of words.[Can we ignore meanings?]

Reflection

- What have we learned?

Readings

- David Mimno, The Details: Training and Validating Big Models on Big Data (video). http://journalofdigitalhumanities.org/2-1/the-details-by-david-mimno/

Week 10: Additional Topics

T 28 MAR / Day 18

Topics

- Stemming and the difference it can make

Readings

- Class handout

R 30 MAR / Day 19

Topics

- Some ways in which topic modeling can mislead

Readings

- Benjamin M. Schmidt, "Words Alone: Dismantling Topic Models in the Humanities." http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/

Week 11: PROJECTS 1 / Jockers Visit

T 4 APR / Day 20

Topics

- The epistemology of data: How much theory is needed?
- The epistemology of digital humanities and its differences from data science.

Readings

- Chris Anderson, "The End of Theory." https://www.wired.com/2008/06/pb-theory/
- Reading on traditional statistics vs. predictive analytics and the problem of interpretability TBD.

R 6 APR / Day 21 Projects 1

Topics

- Overview of the project process -- Question, Data, Analysis, Product
- Thinking carefully about and planning your research project.
- Locating sources for data analysis.
- Where to get help when something goes wrong

Week 12: PROJECTS 2

T 11 APR / Day 22: Project Overview

Topics

- Acquiring resources (e.g. Project Gutenburg)
- How much data do you need?

R 13 APR / Day 23: Data Acquisition

Topics

- Stoplists: How to choose them, how to customize them
- Training – Stability of output under parameter changes.

Week 13: PROJECTS 3

T 18 APR / Day 24

Topics

- Critically interpreting your project. What worked and what did not?

R 20 APR / Day 25

Topics

- Can you validate your model?

Week 14: PRESENTATIONS

T 25 APR / Day 26

R 27 APR / Day 27

Week 15: PRESENTATIONS

R 2 MAY / Day 28

Lost and Found

Dispersion plots

Clustering and small corpora

Loading XML files

Unsupervised clustering

Installing XML package

Using R to process XML

Metadata

Off the shelf software vs project specific programs (e.g. MALLET vs writing your own program).

Working on single files.

Removing proper names <?>

Parts of speech taggers <?>

Sentiment analysis <I think this would be a bit too much. What do you think?>

Looping through multiple files. How big should the database be?

SHINY visualization