Cultural and Literary Text Mining

Kent K. Chang
(MSc student in Digital Humanities, DIS, UCL)
<kai-hsiung.chang.17@ucl.ac.uk>

May 2, 2018 · Version 3.80

~~COURSE~~ OVERVIEW

Welcome to Cultural and Literary Text Mining. As suggested by its title, this study group is where we learn and explore how to apply computational methods to the study of literature and culture. We look at the methodological debate of this "quantitative turn", the tools and practices of which are becoming fundamental to the field now known as digital humanities. How does text mining inform literary and cultural studies? In a discipline that privileges close reading, what do people think about computational methods? What do we need to know to conduct data-driven research on literature and culture? What does a typical research project involve? We seek to find some answers to those questions here.

Although it's called a study group, you may very well think of it as a student-initiated (Kent being the student), student-led, 15-credit postgraduate taught MA + MSc module. We will proceed in form of:

- **mini-lecture and seminar (theory)**, where a participant (Kent by default) tries to highlight key arguments in readings for each session, and other participants may share their opinions on and questions for them (if they have read them in advance, which is not strictly required)
- **Python workshop (practical)**, in which we first learn the basics of Python by making connections with JavaScript/PHP, and then try some tasks typically involved in a text mining project.

I imagine we will go through three major phases (first two detailed below):

Phase I	emphasis on relevant theories + Python basics
Phase II	emphasis on practical text mining skills in Python + related academic projects
Phase III	advanced Python NLP in practice and case studies + some statistics

PREREQUISITE-ISH

This group is open primarily to students currently doing an MA/MSc in Digital Humanities, as well as to those in the Department of Information Studies, at University College London. Typical participants have a BA degree in a humanities discipline such as English, history, classics, etc., and/or are interested in learning about or applying computational methods in the humanities.

The following modules are informal (or, indeed, nothing is formal for an inherently informal study group) prerequisites:

- INSTG008 Digital Resources in the Humanities
- G018 Introduction to Programming and Scripting

G008 has given us some idea of the scope of today's digital humanities research, and to a certain extent, this study group takes some of its sessions related text analysis as point of departure. G018, on the other hand, has equipped its students with programming concepts that are language independent, as well as some hands-on experience of coding—both are useful when you are starting to learn a new programming language (Python, in our case).

Having said that, you are likely to familiarize yourself with the nature of DH/humanities research and develop intuition for programming along the way.

TIME AND PLACE

Details will be announced during the first meeting; ideally we will meet at least once a week, in an evening during the week.

Unless otherwise noted, our meetings should take place in Foster Court G31. Since this is not a computer cluster room, please bring your own laptops to the meetings, or borrow one at Science Library (UCL ID required).

WEBSITE AND FACEBOOK GROUP

Schedule and provisional agenda for each meeting will be made available at https://caltmig.kentchang.com. On this website you will also find links to lecture slides, reflective blog posts, IPython notebooks, etc., where relevant. If you are still using Facebook, you may want to join our group at https://www.facebook.com/groups/2079890148964442/.

TENTATIVE TOPICS

The following is an indicative list of topics and readings. Readings listed below will be covered and discussed during mini-lectures and seminars; a secondary reading list will appear in Kent's reflective blog posts after (presumably) each meeting.

I don't expect you to finish all the readings—let's face it: even if you have time you don't necessarily want to spend them on reading, however dedicated you are to the subject matter; and admittedly, this looks like a lot of work. Ergo I am using the following labels to help you as you prepare for each meeting:

- REQUIRED : essential to follow the mini-lectures and seminars
- !! : highly recommended
- ! : recommended

And please understand that this list is subject to change.

[SESSION 1.]

TOPIC 1 ■ **Orientation**

What is this group? What is this computational study of literature and culture?

- Jockers and Underwood, "Text Mining and the Humanities"
- !! Long and So, "Literary Pattern Recognition"
- Sculley and Pasanek, "Meaning and Mining"
- !! Underwood, "It looks like you're writing an argument against data in literary study . . ."
- Kent Chang, "Articulating the 'Love that Dare Not Speak Its Name' before and after 1895: Topic Modeling Works of Male Homosexual Desire, 1806–1922"

Practical. **Python Fundamentals (I)**—Setting up your own environment

- Background reading: Saskar pp. 51–65
  - Getting to Know Python
  - Installation and Setup

Phase I: Theoretical Foundation and Python Basics

[SESSIONS 2–5.]

TOPIC 2 ■ **Scale and the "Quantitative Turn": Rethinking the Part–Whole Dialectic in Literary Studies**

What is this quantitative approach to literary texts?

- **"Close reading", an overview**
  - New Criticism:
    - !! Eliot, "Tradition and the Individual Talent" (probably also his "Wasteland")
    - Richards, *Practical Criticism*: introductory

        * Empson, *Seven Types of Ambiguity*: preface and chapter 1
        * Wimsatt Jr. and Beardsley, "The Intentional Fallacy"
    - Stylistics: Spitzer, *Linguistics and Literary History*: chap. 1
    - Deconstruction: Derrida, *Limited Inc.*: "Signature Event Context"
    - Surface and Symptomatic Reading: Best and Marcus, "Surface Reading: An Introduction"
- **Reconsidering scales**
    - [REQUIRED] English and Underwood, "Shifting Scales"
    - Ramsay, *Reading Machines*: chap. 1
    - Hoover, "Quantitative Analysis and Literary Studies"
    - [!] Underwood, *Why Literary Periods Mattered*: chap. 6
    - Jay, ""Hey! What's the Big Idea?""
- **Some definitional articulations**
    - Distant reading
        * Moretti, *Distant Reading*: "Slaughter House of Literature", "Conjectures on World Literature"; *Graphs, Maps, Trees*: "Graphs, Maps, Trees", "Graphs"
        * [!!] Underwood, "A Genealogy of Distant Reading"
        * [!] Buurma and Heffernan, "Search and Replace"
        * Context of world literature— Casanova, *The World Republic of Letters*: preface and chap. 1
    - Cultural Analytics
        * [!!] Piper, "There Will Be Numbers"
        * Manovich, "Cultural Analytics: Visualising Cultural Patterns in the Era of "More Media""
        * Michel et al., "Quantitative Analysis of Culture Using Millions of Digitized Books"
    - Macroanalysis: Jockers, *Macroanalysis*: part 1
- **Some critical stances**
    - On the practices of DH
        * Golumbia, "Death of a Discipline"
        * Liu, "Where Is Cultural Criticism in the Digital Humanities?"
    - Between "close" and "distant" reading
        * Walkowitz, *Born Translated*: chap. 1
        * Bode, "The Equivalence of "Close" and "Distant" Reading; or, Toward a New Object for Data-Rich Literary History"
    - The nature and notion of "literary data"
        * Gitelman, *"Raw Data" is an Oxymoron*: introduction

        * Marche, "Literature Is Not Data"
    - Counterbalance— Piper, "Why are Non-data Driven Representations of Data-driven Research in the Humanities So Bad?"

Practical. **Python Fundamentals (II–V)**—Python Syntax and Essential Libraries

- Optional homework: Think about what you want to study and mine.
- Background reading: Sarkar (each bullet point ≈ one week)
    - chap. 4 (Python Syntax and Structure), pp. 66–84 (up to §Controlling Code Flow)
    - chap. 4, pp. 84–91 (Functional Programming)
    - chap. 4, pp. 91–105 (to the end of chapter)
- Dataquest: Python Basics (each bullet point ≈ one week)
    - Files and Loops; Booleans and If Statements
    - List Operations; Dictionaries
    - Functions
- Useful Libraries (handouts)
    - `numpy`
    - `panda`

[SESSION 6.]

TOPIC 3 ■ Formulating Research Questions

How do we actually address humanities inquiries through computational methods (i.e do something more than counting)?

- **Operationalization**
    - Moretti, ""Operationalizing""
    - Drucker, "Why Distant Reading Isn't"
    - So, ""All Models Are Wrong""
- **Literary modeling**
    - REQUIRED Piper, "Think Small"
    - McCarty, "Knowing . . .: Modeling in Literary Studies"
    - Wilkens, "Canons, Close Reading, and the Evolution of Method"

Practical. **Obtaining Texts**

- Optional homework: Prepare to share your project ideas with us.
- Downloading texts from Project Gutenberg using `Gutenberg`
- Web Scraping and `BeautifulSoup`

Phase II: Cultural/Literary Text Mining in Practice

[SESSIONS 7–9.]

TOPIC 4 ■ Text Classification

- **Authorship Attribution**: Jockers, *Macroanalysis*: chap. 6 (Style)
- **Cultural Capital and Literary Field**:
  - REQUIRED Theory— Bourdieu, *The Field of Cultural Production*: ideally pt. 1, at least chap. 1
  - !! Piper and Portelance, "How Cultural Capital Works"
  - Algee-Hewitt et al., "Canon/Archive: Large-scale Dynamics in the Literary Field"
  - Quist, "Laurelled Lives: the Swedish Academy's Praise for Its Prizewinners"
- **Gender**:
  - Theory—Butler, *Gender Trouble*: at least chap. 1 (Subjects of Sex/Gender/Desire)
  - Piper and So, "Women Write About Family, Men Write About War"
  - Argamon et al., "Gender, Genre, and Writing Style in Formal Written Texts"

Practical. **NLP Basics and Text Classification**

- NLP Basics—Sarkar: chap. 1 and 3 (Natural Language Basics; Processing and Understanding Text)
- Text Classification—Sarkar: chap. 4 (Text Classification)

[SESSION 10.]

TOPIC 5 ■ Social Network Analysis

- Background: Weingart, "Demystifying Networks"
- !! So and Long, "Network Analysis and the Sociology of Modernism"
- Schich et al., "A Network Framework of Cultural History"

Practical. **Network Analysis**—`NetworkX` in Python

- Al-Taie (not Sarkar!): chap. 2

[SESSIONS 11–12.]

TOPIC 6 ■ Topic Modeling

- Background— REQUIRED Underwood, "Topic modeling made just simple enough."; Blei, "Probablistic Topic Models" (alert: a bit mathy)
- **Themes in literature**
  - Jockers, *Macroanalysis*: chap. 8

- Jockers and Mimno, "Significant Themes in 19th-Century Literature"

- **Knowledge Structures**
  - Theory— Foucault, "The Discourse on Language"
  - Goldstone and Underwood, "The Quiet Transformations of Literary Studies"

  Practical. **Topic modeling**
    - Sarkar: chap. 5

TOPIC 7 ■ Commencement

Let's start from here.

- **More methods**
  - Sentiment analysis— Stanford Literary Lab, "Mapping London's Emotions"; *Optional practice*: see Sarkar chap. 7
  - Geospatial analysis— Wilkens, "The Geographic Imagination of Civil War-Era American Fiction"
  - Mixed methods— Himelboim, McCreery, and Smith, "Birds of a Feather Tweet Together"
- **Finale**: Concluding remarks

TENTATIVE PRIMARY READING LIST

Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser. "Canon/Archive: Large-scale Dynamics in the Literary Field." Stanford Literary Lab. January 2016. https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf.

Argamon, Shlomo, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. "Gender, Genre, and Writing Style in Formal Written Texts." *TEXT* 23 (2003): 321–46.

Best, Stephen, and Sharon Marcus. "Surface Reading: An Introduction." *Representations* 108, no. 1 (2009): 1–21.

Blei, Daivd M. "Probablistic Topic Models." *Communications of the ACM* 55, no. 4 (2012): 77–84.

Bode, Katherine. "The Equivalence of "Close" and "Distant" Reading; or, Toward a New Object for Data-Rich Literary History." *Modern Language Quarterly* 78, no. 1 (March 2017): 77–106. Accessed April 22, 2018.

Bourdieu, Pierre. *The Field of Cultural Production: Essays on Art and Literature*. Reprinted. Edited by Randal Johnson. Cambridge: Polity Press, 1993.

Butler, Judith. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge, 1990.

Buurma, Rachel Sagner, and Laura Heffernan. "Search and Replace: Josephine Miles and the Origins of Distant Reading." The Discipline. April 11, 2018. https://modernismmodernity.org/forums/posts/search-and-replace.

Casanova, Pascale. *The World Republic of Letters*. 1. Harvard University Press paperback edition. Translated by M. B. DeBevoise. Convergences inventories of the present. Cambridge, Massachusetts London, England: Harvard University Press, 2007.

Derrida, Jacques. *Limited Inc.* Evanston, IL: Northwestern University Press, 1988.

Drucker, Johanna. "Why Distant Reading Isn't." *PMLA* 132, no. 3 (May 2017): 628–35. Accessed April 23, 2018.

Eliot, T. S. "Tradition and the Individual Talent." In *The Norton anthology of theory and criticism*, 1st ed, edited by Vincent B. Leitch, 1092–98. New York: Norton, 2001.

Empson, William. *Seven Types of Ambiguity*. [3. ed.], 14. print. New Directions paperbook 204. New York: New Directions Publ, 1984.

English, James F., and Ted Underwood. "Shifting Scales: Between Literature and Social Science." *Modern Language Quarterly* 77, no. 3 (September 2016): 277–95. Accessed April 22, 2018.

Foucault, Michel. "The Discourse on Language." In *Truth*, edited by Jos Medina and David Wood, 315–35. Ames, Iowa, USA: Blackwell Publishing, January 1, 2005. Accessed April 23, 2018.

Gitelman, Lisa, ed. *"Raw Data" is an Oxymoron*. Infrastructures series. Cambridge, Massachusetts: The MIT Press, 2013.

Goldstone, Andrew, and Ted Underwood. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 4, no. 2 (2014): 359–84.

Golumbia, D. "Death of a Discipline." *differences* 25, no. 1 (2014): 156–76.

Himelboim, Itai, Stephen McCreery, and Marc Smith. "Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter." *Journal of Computer-Mediated Communication* 18, no. 2 (January 2013): 40–60.

Hoover, David L. "Quantitative Analysis and Literary Studies." In *A companion to digital literary studies*, edited by Raymond George Siemens and Susan Schreibman, 517–53. Blackwell companions to literature and culture 50. Malden, MA: Blackwell Pub, 2007.

Jay, Martin. ""Hey! What's the Big Idea?": Ruminations on the Question of Scale in Intellectual History." *New Literary History* 48, no. 4 (2017): 617–31.

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. OCLC: 829936927. Urbana, IL: University of Illinois Press, 2013.

Jockers, Matthew L., and David Mimno. "Significant Themes in 19th-Century Literature." *[pre-print]*, August 2012.

Jockers, Matthew L., and Ted Underwood. "Text Mining and the Humanities." In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 291–306. Chichester, UK: John Wiley & Sons, Ltd, November 27, 2015. Accessed April 5, 2018.

Liu, Alan. "Where Is Cultural Criticism in the Digital Humanities?" In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 490–510. University of Minnesota Press, January 1, 2012. Accessed April 23, 2018.

Long, Hoyt, and Richard Jean So. "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning." *Critical Inquiry* 42, no. 2 (January 2016): 235–67. Accessed April 5, 2018.

Manovich, Lev. "Cultural Analytics: Visualising Cultural Patterns in the Era of "More Media"." *Domus March*, March 2009.

Marche, Stephen. "Literature Is Not Data: Against Digital Humanities." Los Angeles Review of Books. October 28, 2008. https://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities/.

McCarty, Willard. "Knowing . . .: Modeling in Literary Studies." In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 391–401. Chichester, UK: John Wiley & Sons, Ltd, November 27, 2015. Accessed April 5, 2018.

Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, et al. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331, no. 6014 (January 14, 2011): 176–82. Accessed April 23, 2018.

Moretti, Franco. *Distant Reading*. London ; New York: Verso, 2013.

———. *Graphs, Maps, Trees: Abstract Models for a Literary Theory*. London: Verso, 2005.

———. ""Operationalizing": or, the function of measurement in modern literary theory." Stanford Literary Lab. December 2013. https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf.

Piper, Andrew. "There Will Be Numbers." *Cultural Analytics*, May 23, 2016.

———. "Think Small: On Literary Modeling." *PMLA* 132, no. 3 (May 2017): 651–58. Accessed April 23, 2018.

———. "Why are Non-data Driven Representations of Data-driven Research in the Humanities So Bad?" .txtlab. September 17, 2017. https://txtlab.org/2017/09/why-are-non-data-driven-representations-of-data-driven-research-in-the-humanities-so-bad/.

Piper, Andrew, and Eva Portelance. "How Cultural Capital Works: Prizewinning Novels, Bestsellers, and the Time of Reading." Post45. May 10, 2016. http://post45.research.yale.edu/2016/05/how-cultural-capital-works-prizewinning-novels-bestsellers-and-the-time-of-reading/.

Piper, Andrew, and Richard Jean So. "Women Write About Family, Men Write About War." The New Republic. April 8, 2016. https://newrepublic.com/article/132531/women-write-family-men-write-war.

Quist, Jennifer. "Laurelled Lives: the Swedish Academy's Praise for Its Prizewinners." *New Left Review* 104 (March–April 2017): 93–106.

Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism*. University of Illinois Press, 2011. Accessed April 23, 2018.

Richards, I. A. *Practical Criticism: a Study of Literary Judgment*. Place of publication not identified: publisher not identified, 1929.

Sarkar, Dipanjan. *Text Analytics with Python*. Berkeley, CA: Apress, 2016. Accessed April 5, 2018.

Schich, M., C. Song, Y.-Y. Ahn, A. Mirsky, M. Martino, A.-L. Barabasi, and D. Helbing. "A Network Framework of Cultural History." *Science* 345, no. 6196 (August 1, 2014): 558–62.

Sculley, D., and B. M. Pasanek. "Meaning and Mining: the impact of implicit assumptions in data mining for the humanities." *Literary and Linguistic Computing* 23, no. 4 (September 29, 2008): 409–24. Accessed April 5, 2018.

So, Richard Jean. ""All Models Are Wrong"." *PMLA* 132, no. 3 (May 2017): 668–73.

So, Richard Jean, and Hoyt Long. "Network Analysis and the Sociology of Modernism." *boundary 2* 40, no. 2 (June 1, 2013): 147–82. Accessed April 23, 2018.

Spitzer, Leo. *Linguistics and Literary History*. NJ: Princeton University Press, 1948.

Stanford Literary Lab. "Mapping London's Emotions." *New Left Review* 101 (September–October 2016): 63–91.

Al-Taie, Mohammed Zuhair. *Python for Graph and Network Analysis*. New York, NY: Springer Berlin Heidelberg, 2017.

Underwood, Ted. "A Genealogy of Distant Reading." *Digital Humanities Quarterly* 11, no. 2 (2017).

Underwood, Ted. "It looks like you're writing an argument against data in literary study . . ." September 21, 2017. https://tedunderwood.com/2017/09/21/it-looks-like-youre-writing-an-argument-against-data-in-literary-study/.

———. "Topic modeling made just simple enough." The Stone and the Shell. April 7, 2012. https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/.

———. *Why Literary Periods Mattered: Historical Contrast and the Prestige of English Studies*. California: Stanford UP, 2013.

Walkowitz, Rebecca L. *Born Translated: the Contemporary Novel in an Age of World Literature*. Literature now. New York: Columbia University Press, 2015.

Weingart, Scott B. "Demystifying Networks." The scottbot irregular. December 14, 2011. http://scottbot.net/lets-talk-about-networks/.

Wilkens, M. "The Geographic Imagination of Civil War-Era American Fiction." *American Literary History* 25, no. 4 (December 1, 2013): 803–40. Accessed April 23, 2018.

Wilkens, Matthew. "Canons, Close Reading, and the Evolution of Method." In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 249–58. University of Minnesota Press, January 1, 2012.

Wimsatt Jr., W. K., and M. C. Beardsley. "The Intentional Fallacy." *The Sewanee Review* 54, no. 3 (September 1946): 468–88.

ACKNOWLEDGEMENTS

This document is created based on Andrew Goldstone's LaTeX template at https://github.com/agoldst/tex/blob/master/syllabus/example/syllabus.tex.

In preparing for this document, I have referred extensively to syllabi publicly available online. Those made by Andrew Goldstone (350:509, 350:596), Andrew Piper (LLCU 614, LLCU 255), David Mimno (INFO 3350), Mark Algee-Hewitt (ENGLISH 184E, ENGLISH 253, ENGLISH 354), and Matthew Jockers (ENGLISH 4/898) have significantly informed my compilation.