Supporting Computational Text Analysis in East Asian Studies

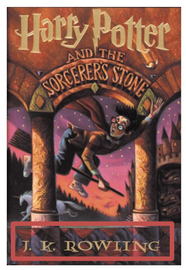
Xiang Li[1] and Yao Chen[2]
[1]University of Colorado Boulder, [2]Univeristy of Minnesota

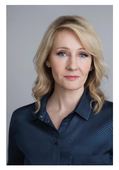
Introduction

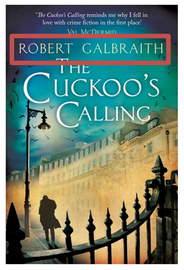
Text analysis (text mining) is the process of analyzing large collections of textual materials in order to discover new information (e.g., theme, relationship and trends).[1-4] Text analysis is more than information retrieval. It draws on "information retrieval, data mining, machine learning, statistics, and computational linguistics." [3]

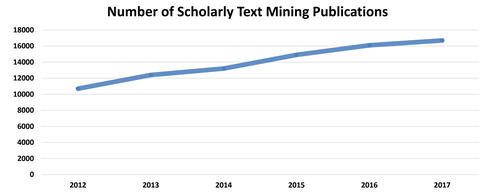
Computational text analysis has been utilized in a variety of humanities disciplines. "Text analysis looks at elements such as word frequencies, co-occurrence, and statistically generated 'topics' to perform 'distant reading' of texts. Humanists usually perform this analysis with the help of algorithms developed by computer scientists, statisticians, and linguists."[5]

Example: Did J. K. Rowling write the book *The Cuckoo's Calling*?









Computational text analysis is now rapidly developing in the field of East Asian Studies. "More scholars are becoming conversant in the variety of analytical possibilities these technological developments make available. We are slowly seeing more original research that applies digital analysis in dissertations and articles. Soon it will find an established place among more traditional modes of scholarly analysis." [6]

Examples of Computational Text Analysis Applications

Stylometry is the statistical analysis of variations in literary style. It is used primarily for authorship attribution studies and genre detection.

Sample projects:

- Vierthaler explored the complex stylistic relationships of texts in the late Mind and early Qing periods in China and found a gradient of style that ran from purely fictional works through historical romances (novels with historical content) and unofficial histories 野史 to official historical works.[7]
- Vierthaler used stylometric and machine learning analyses to explore the probable authorship of the late Ming dynasty novel the *Plum in the Golden Vase*金瓶梅.[8]

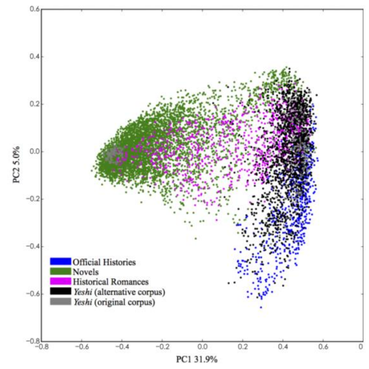
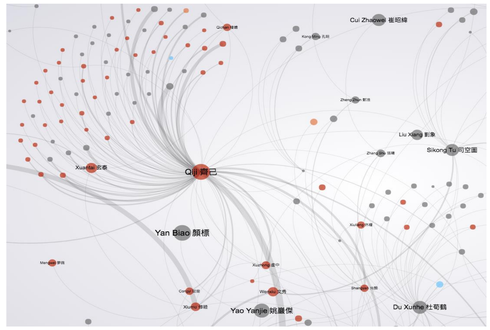


Figure 9. Principal component analysis of official histories, novels, historical romances, and texts taken from wenxian.famenli.com's Storied Records and Recorded Annals collection, which contains a significant number of yeshi, among other texts. This corpus is limited to texts written in the Ming and the Qing dynasties. The shape is similar to previous figures, but the overlap between yeshi and official histories is even more distinct, with significant intrusions into fictional space.

Social network analysis is the mapping and measuring of networked structures in terms of nodes (individual actors) and links (relationships or interactions) that connect them.

Sample projects:

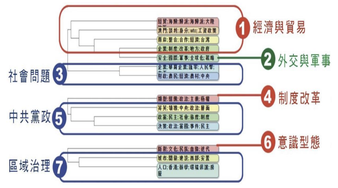
- Lee examined the 1917-1927 writer-periodical network in Korea and revealed the position of women writers as a prehistory to the formation of male-centered dongin.[9]
- So and Long analyzed the structural relations between poets in early 20 century United States, Japan and China and discussed how the poets' interactions help to constitute the field of modernist poetry as a whole.[10]
- Mazanec created a network map of poetry exchange in late medieval China (c. 860-960 CE). [11]

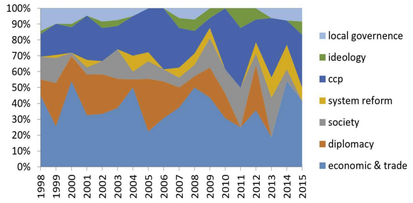


Topic modeling helps organize large collections of textual information to discover topics (themes) that occur in a collection of textual documents.

Sample projects:

- Shao, Huang and Tsai's study about how the Taiwanese do China Studies applied a topic modeling tool to analyze papers published in the *Mainland China Studies* during 1998-2015. Their results showed that the articles were clustered into seven salient topics. [12]
- Le, Lee and Lee analyzed multi-lingual customer comments about Starbucks in social network across US, Korea, Singapore, and Vietnam between 2011-2014. Top posted themes were summarized to examine Starbucks' marketing strategies. [13]





Sample Tools

- **R** and **Python** are programming languages that are widely applied to a variety of text analysis projects and are compatible with CJK scripts. There are ready-to-use packages that allow users to perform specific text analysis tasks with limited programming skills.
- **HathiTrust Research Center** contains a suit of tools that could help conduct text analysis projects.
- **Word Segmenter**
  - Stanford Word Segmenter, Rakuten MA, MeCab
- **Named Entity Recognizer**
  - MARKUS, NameLister
- **Network Analysis and Visualization Tools**
  - Cytoscape, Gephi, NetMiner
- **Topic Modeling Tools**
  - MALLET, Stanford Modeling Toolbox, Latent semantic analysis

How to Support?

- Know sources of text and available expertise
  - Free and paid data
  - Experts on campus or in communities
  - Communicate user needs to content providers and technology experts
- Perpetual access licensing
  - Understand different levels of access
  - Educate content providers
- Negotiate text mining rights
  - Advocate for full data access
- Secured storage and preservation
  - Work with preservation staff and data management staff
- Conduct systematic literature review
  - Know updates and new methods and tools
- Develop expertise
  - Join an existing library or community group
  - Attend workshops and webinars
- Outreach and training
  - Host events to increase awareness
  - Provide training to support new research

Contact Information

Xiang Li
< Xiang.Li@colorado.edu >

Yao Chen
< chen3200@umn.edu >

References

1. Bretonnel Cohen and Lawrence Hunter, "Getting started in text mining," *PLoS computational biology* 4, no. 1 (2008): e20.
2. Eileen Gardiner and Ronald Musto, *The Digital Humanities: A Primer for Students and Scholars* (Cambridge University Press, 2015).
3. Gabe Ignatow and Rada Mihalcea, *Text Mining: A Guidebook for the Social Sciences* (Sage Publications, 2016).
4. Yu Zhang, Mengdong Chen and Lianzhong Liu, "A Review on Text Mining," *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (IEEE, 2015): 681-685.
5. "Text Analysis Resources," Digital Humanities at Berkeley (Website), Accessed March 4, 2018, http://digitalhumanities.berkeley.edu/resources/text-analysis-resources.
6. Paul Vierthaler, "Imperial Chinese Studies and Trends in the Digital Humanities," *China Policy Institute: Analysis*, 2016.
7. Paul Vierthaler, "Fiction and History: Polarity and Stylistic Gradience in Late Imperial Chinese Literature," *Journal of Cultural Analytics*, 2016.
8. "Research," Paul Vierthaler (website), Accessed March 4, 2018, https://www.pvierth.com/.
9. JY Lee. "Before and After the 'Age of Literary Coteries': A Diachronic Analysis of Writers' Networks in Korea, 1917-1927." *Korea Journal* 57, no.2 (2017): 35-68.
10. Richard So and Hoyt Long, "Network Analysis and the Sociology of Modernism," *Boundary 2* 40, no. 2 (May 1, 2013): 147–82.
11. Tom Mazanec (website), Accessed March 4, 2018. http://tommazanec.com/projects/network-maps/.
12. Hsuan-Lei Shao, Sieh-Chuen Huang and Yun-Cheng Tsai, "How the Taiwanese Do China Studies: Applications of Text Mining," *arXiv preprint arXiv:1801.00912* (2018).
13. Hoanh-Su Le, Jong-Hwa Lee and Hyun-Kyu Lee, "Designing an Integrated Text Mining Framework for Evaluating Cross-cultural and Multi-lingual Customer Responses in Social Network Marketing," *한국경영학회 통합학술발표논문집* 2015 (2015): 1349-1361.