Distant Reading and Mapping Genre Space via Conjecture-based Distance Measures

Juola, Patrick
juola@mathcs.duq.edu
Duquesne University

One of the key problems facing digital humanities today is the increasing number and size of digital repositories and the relative lack of tools for studying them. A collection of a million books (Crane, 2006) is no more useful than a collection of ten thousand if you can't read more than a hundred of them in a realistic timeframe. Scholars like Moretti (2005) have proposed a new analysis method, termed "distant reading," to enable computer-aided large-scale analysis of such collections. In previous work (Juola and Bernola, 2009), we have proposed using a conjecture generator (Conjecturator, see also `http://www.twitter.com/conjecturator`) as another computer-aided analysis method.

Underlying the Conjecturator is the idea that the the computer can be deployed to autonomously generate "facts" about a given text repository. Like its predecessor and inspiration *Graffiti* (Fajtlowicz, 1988), the conjecturator generates template-based "conjectures" that might or might not be true about the repository and the texts in it. A sample conjecture might be something like:

- The concept of "archivist" appears more in mid-Victorian novels than in psychological realism novels or, more obviously,
- The concept of "femininity" appears more in feminist novels than in novels with gothic elements.

(Who would have thought, eh?)

As discussed in (Juola and Bernola, 2009), these simple conjectures can be easily and quickly tested to refute or confirm their validity. This enables the computer to quickly generate a pile of isolated "facts" about the text repository, but does not provide a useful framework for interpretation, explanation, or understanding (which still requires human expertise).

However, this "pile of facts" can provide useful source material for distant reading. In this paper, we demonstrate one way to extend this conjecture-based analysis to a large-scale "distant reading" and visualization of genre differences. Repeated generation of conjectures will create a large catalogue of potential differences between any particular category pair, some true/supported, and some false. The number of "true" differences, or alternatively, the percentage of true differences, can be viewed as a distance between the categories, a distance measuring the degree of difference between the concepts commonly written about in those genres. If, for instance, "epistolary novels" differ in 26 significant ways from "tragic novels", but only in one significant way from "fiction of manners," we can consider "epistolary novels" to be a closer genre in terms of expressed concepts to "fiction of manners" than to "tragic novels." This high-order analysis gives us a large-scale conceptual grouping of genre categories.

To aid in the study of such differences, we compile the differences into a matrix and apply multidimensional scaling (MDS) (Cox and Cox, 2001). This statistical technique takes a high-dimensional data set defined by interpoint distances and embeds/rescales it to fit a smaller number of dimensions (in this case, two) while minimizing distortion. The resulting two-dimensional coordinates can be plotted to give a visual "map" of the space of genres. We demonstrate this technique using an enlarged set of 136 novels representing 36 genres (including time period and authorial attributes as "genres") and approximately 10,000 validated conjectures (culled from approximately 85,000 conjectures in total).

The resulting images clearly indicate that this method is a new and viable way of performing large-scale distant reading. As can be seen in Figure 1, the resulting "map" passes many obvious tests for rationality; for for example, mid-Victorian novels represent an intermediate stage between early and late Victorian novels; similarly, "male authored novels" in general are an intermediate between "American male authored novels" and "English male authored novels," reflected exactly what intuition suggests. We leave it to genre

specialists to examine the map in detail and to see whether actual genres equally reflect our intuitions. It is easy and relatively efficient to apply and almost entirely document-agnostic; it can be applied as easily to journal articles (and map the space of scholarship) or to newspaper corpora (perhaps mapping the space of editorial policies and politics) as to novel genres.

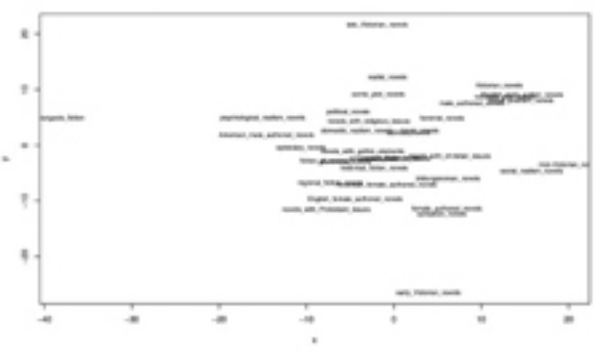

Figure 1

References

Cox, T.F., Cox, M.A.A. (2001). *Multidimensional Scaling*. Chapman and Hall.

Crane, Gregory (2006). 'What Do You Do With a Million Books?'. *D-Lib Magazine*. **12(3)**.

Fajtlowicz, Siemion (1988). 'On conjectures of Graffiti'. *Discrete Mathematics*. **72**.

Juola, Patrick (2009). 'Mapping Genre Space via Random Conjectures'. *Presented at DHCS-2009, IIT*. Chicago, IL.

Juola, Patrick, Bernola, Ashley (2009). 'Conjecture Generation in the Digital Humanities'. *Proc. DH-2009*. 2009.

Moretti, Franco (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.