MATH MATTERS

Apply It.

The math behind... Digital Humanities

Technical terms used:

n-grams, trees, culturomics, network theory, hidden Markov models

Uses and applications:

Researchers in the humanities have used math to decide disputed authorship claims based on statistical analysis of linguistic style, track the evolution of the English language quantitatively, design algorithms that can distinguish between the styles of different artists, and create large searchable databases of works of art (music, literature, etc.) to study patterns among them

How it works:

Tools from genomics, like networks (to study gene regulation), trees (for phylogenies), and n-grams (for gene sequence analysis), are applied to research in the humanities, earning the nickname "culturomics." Google released an "n-gram viewer" that allows users to track the instances of certain words or phrases in a large body of books over the last two centuries – without infringing on copyright by making the texts of the works available. This same idea of the n-gram (a length n sequence of letters) is used to train hidden Markov models that learn to recognize an author's or a composer's style with remarkable accuracy. Meanwhile, even basic math ideas are reshaping how some researchers work on the humanities. Novel theorist Frank Moretti at Stanford's Literary Lab pioneered the idea of "distant reading," analyzing masses of books across history statistically instead of close-reading works selected from the canon. The idea of a "literature lab" is itself ported from traditional sciences. Math has also enabled the creation of these large cultural databases in the first place, from optical character recognition that allows the digitization of manuscripts to musical transcription algorithms and MIDI encoding that convert audio files to a digital, symbolic representation for analysis.

Interesting fact:

The networks used to study literature so far have two main thrusts. One is quantitative: the large-scale statistical analysis of Shakespeare's plays to develop an algorithm for determining who is being addressed, for example. But just as useful to many digital humanities researchers was the idea of a network itself, and the new form of plot representation a network allowed. Instead of seeing plot as a linear unfolding in time, the interactions of characters could be seen all at once, and from that, qualitative patterns became visible—like the centrality of Horatio in the play, or the highly-interconnected world of the Court, where everything social is public.

References:

1) http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf

2) Computing in Musicology, Volume 13: CCARH and the MIT Press, 2004. Articles available online at http://www.ccarh.org/publications/books/cm/vol/13/contents.html

3) Michel, Aiden et al. Quantitative Analysis of Culture Using Millions of Digitized Books. Science 331:6014, Jan 2011.

Submitted by Katie Banks, Harvard College, third place *Math Matters, Apply It!* contest, January 2012.

Society for Industrial and Applied Mathematics
www.siam.org
www.siam.org/careers/matters.php
siam.