text mining workshop (friday morning) -- blog post

Taming the text-mining hydra

general pleas for more full text

http://www.dlib.org/dlib/march06/crane/03crane.html

there have been some very worthwhile introductions to text mining.

http://tada.mcmaster.ca/Main/WhatTA

fear mongering, from this one http://chronicle.com/article/Googles-Book-Search-A/48245/

It has become difficult to have a conversation about digital humanities without text mining Text mining can be done in a million different ways...it can work on enormous data sets; it can involve complex algorithmic analysis, things like normalized compression distances and machine learning.

no question that such techniques are fascinating and promising for the future of research. but just as they are at one of the cutting edges of digital humanities, can make even the idea of text mining, seem like a computer science hammer looking for a humanities nail.

it can certainly seem intimidating. topic modeling, machine learning, etc. but it doesn't have to be intimidating. there's no way to make it entirely non-technical, but it can be much more approachable. that is, remind those interested in text mining that it's more of a general approach to inquiry than a particular technique.

i want to make a case for a conversational approach that i've found quite useful. what do i mean by conversational? it can often mean swearing at your computer, but of course that's kind of one sided.

more profitable is to not require the computer to do all the discovery for you, or even find patterns more traditionally, you might want to get the computer to find what's interesting. there is some reticence to this approach--do we \_really\_ know what the computer is doing? how sure can we be that we've set it up to notice something worth noticing?

while text mining often means getting the computer to tell you what's going on with a set of texts, it doesn't have to. the machine's tolerance for mindless drudgery can be exploited with a much more guided approach as well. the idea is to find something you want to investigate, isolate it, actually read something--but very targeted reading. perhaps better than conversational is highly mediated text mining. it's an interative approach: repeating cycles of isolate and analyze.

2) simply reformatting to get a sense of something.

what if you'd like to know how to sin in the victorian period?

just searching for the word sin isn't going to be that revealing. <n-gram>

with some help from google, i got snippets from google books of 50 chars either side of wherever the word sin appears.

a bit of reformatting can line everything up. to do this, we can use PHP, a ubiquitous programming language on the web; easy to learn w great documentation adn examples.

in this case, i was starting with a nicely formatted data set.. you don't need that.

you could download a set of texts, creating your own research corpus, and extract relevant fragments from that.

could do it with "science of X" what can we learn about usage by looking at phrases that follow that.

what i really want to do is just see all the references to sin together. what can help with that is simple scripting tools.

conclusion

in a larger sense, the methodology reveals the importance of imagination and simple programming knowledge. doesn't need pretty or powerful tools.