Using Voyant and DREaM

This report sketches some ways that Voyant-tools and the DREaM corpus interact in order to help me (the analyst of early modern English literature) glean insight at a macro level. This report will demonstrate the some uses of Voyant and DREaM in a case study concerning the spiritual connotations of crime.

Voyant and DREaM are two distinct tools for managing and interpreting large quantities of documents. Together, they help idenitfy patterns and, therefore, exemptions in a corpus. This is helpful for micro analysis direct close readings of the documents. DREaM is a corpus creator and Voyant is a collection of tools for analysing a corpus and its individual documents.

The DREaM is digital humanities project which builds upon the EEBO database of Early Modern English literature. EEBO is an emormous corpus of texts with over 125,000 individual documents, 44,000 of which have made it into DREaM, fully transcribed and indexed.

The Selection Process

The selection process is crucial in such an analysis. And for this, Voyant and DREaM complement each other. Voyant helps identify quirks within a given corpus; and these quirks in turn refine the corpus selection process.

Voyant's limitations encourage a creative corpus selection: it can process many hundreds of texts but, as of now, it cannot process the entire database (thousands of documents). In this case the corpus must be narrow enough but inclusive enough to allow for the comparison of two sets of keywords. The relative frequency of keywords within the entire corpus is essential for the comparative potential of any two words. Dates are instrumental for a convincing analysis insofar as the researcher desires to generalize from the necessarily limited selection of documents.

Connotations of Crime

The two sets of words in this case comparison are characterized as either secular or religious transgressions: (1) poaching, wrecking, smuggling, coining, rioting, burglary and (2) atheism, heresy, witchcraft, sorcery, poisoning. These sets share in their illicit status but differ in the emphasis of spirituality; such a polarity warrants the use of automated literature processing. It is required to familiarize oneself with the vocabulary of Early Modern England in order to best formulate word lists; I did not expect coining to be such a ubiquitious crime, nor did I expect poisoning to be a decidedly devilish (spiritual) transgression. Discovering new words, and thereby fleshing out the word set is part of the initial explorations with Voyant.

The corpus I chose spans the entire database; I do this in order to maximize the generalizability of my findings. Otherwise, for period specific analysis, the DREaM corpus creator has an insightful tools which represents the relative frequencies of a keyword throughout the nearly three centuries under examination. Because spiritual crimes have such a higher frequency, I could not have made a corpus with such a time span. The corpus in question, however, has many instances of spiritual crimes (unsurprisingly, considering their high frequency).

I chose a set which spans the entire corpus because I was able to make the words constituting the corpus have a comparative frequency, around 40-80 documents with any given key word. One of my corpi was created with "shipwrecking*, shyppew*, shypew*, smugg*, poach*, burglarie, arson*" -- 302 unique documents -- with about 60 instances of either shipwrecking, smuggling, poaching, burglary, or arson. Selecting a comparatively useful corpus is a necessary skill when using Voyant and DREaM together.

Voyant Tools

Keywords in Context is one of the most useful tools available. With it, the researcher can easily get a sense of a how a word is used within the corpus.

Using Keywords in Context, certain documents with exceptional characteristics might be found. Like in the above case where document 8 uses poaching in contrast to the spiritual crime of atheism.

Collocations Visualizations is another useful (and aesthetically pleasing) tools. With it, the research can find interesting new keywords and map out the connections between various themes (or, mutual collocations). The connecting words, nodes between spiritual and secular crimes, are revealing. Below, criminal *roles* function as connectors (ie. "villian" and "cheater")

Collocations, besides the pleasing presentation, introduces a researcher perhaps less in tune with early modern diction to words which once enjoyed regular use: "schismaticks," "advowterer," and "atheistick" are among my favorite.

Although this case is not enough to make strong claims, it does suggest that secular and spiritual crimes were distinguished. Moreover, secular crimes insinuate heresies, but less so in the other direction. In order to extrapolate more about such a polarity of connotations, the DREaM corpus can be processed through a Python script. See, The Art of Literary Text Analysis for further instructions.

Sentiment Analysis

There are many Python libraries (stock code to help perform complex tasks with simple code) which aid textual analysis -- many of which can be reproduced in Voyant-tools without any fuss. Sentiment analysis libraries are helpful for studies which seek to compare such a polarity.

Polarities are the logical basis of sentiment analysis. Sentiment analysis is at its simplest a score of +1 for positive words and -1 for negative words; one such program can approach dizzying sophistication. For our purposes, the host of simple python tools will suffice. Below is a simple IPython notebook which reads the sentiment score of a single document within my corpus.

Textblob is a library for performed sentiment analysis on movie reviews; it is not enough for a critical essay on English Literature, but the simplicity helps illustrate how DREaM can integrate with a huge amount of tools.

```
import os
import nltk
import glob
from bs4 import BeautifulSoup
from textblob import TextBlob

#this is the function to string xml tags from files
#beautiful soup will parse the xml
def strip_tags(textFile):
    soup = BeautifulSoup(open(textFile), "lxml")
    stripped_text = soup.get_text()
    return stripped_text

#loop through the files using glob
textFiles = glob.glob("*.xml")

i = 0
for textFile in textFiles:
    cleanText = strip_tags(textFile)
    blob = TextBlob(cleanText)
    print i, ": ", blob.sentiment.polarity, " in ", textFile
    i += 1
```

Using sentiment analysis, the researcher is able to identify outliers within the corpus; this is inline with the pattern of macro analysis which works to direct readings towards an especially insightful text from an enormous corpus.

Some of the outliers I find in this corpus are (this is an example of what the above script prints):

```
136 : -0.0679569892473 in 1660 - unknown author - The
Rump ululant or Peni.xml
95 : 0.334507484768 in The yere of oure lorde MC - Watson
Henry fl 1500-1518 - Here endeth ye hystorye o.xml
9 : 0.284335838723 in 1586 - Ortâuñez de Calahorra Die -
Mirrour of Princely Deeds.xml
```

Interestingly, the most positive work is a spiritual work.

Conclusion

Differentiation is key for narrowing an analysis. In this case, the two sets of words overlap in meanings and my task as researcher is to show their differences. This means finding *documents* which go beyond the norm.

In this case, I found:

- There was an awareness of a difference between secular and religious crimes.
- Heresy was plainly criminal and/or stigmatized.
- Crimin*als* (cheaters, villians, atheists) are the common mutual collocation for these two types of crimes.

Although heresy was criminal, it was not the case that crimes were heresy (as I had suspected). Ultimately, 'crime' is a secular affair (with sin as the spiritual counterpart). That there is a connection at all, I think, is interesting. In fact, the reformation, of which the EEBO database is an interesting snippet, was the catalyst of this seperation of the spiritual and secular. That there is a connection of heresy and crime has likely more to do with cultural precedent, and this differentiation (and awareness thereof!) is novel.

April 2015, Fenimore Love