DHQ: Digital Humanities Quarterly
2016
Volume 10 Number 2

The Ancient World in Nineteenth-Century Fiction; or, Correlating Theme, Geography, and Sentiment in the Nineteenth Century Literary Imagination

Matthew L. Jockers <mjockers_at_unl_dot_edu>, University of Nebraska

Abstract

"The Ancient World in 19th-Century Fiction" is a lightly revised version of a lecture delivered at the first meeting of the Digital Classicists Association. The intent of the lecture, in accordance with the invitation to deliver it, was to introduce literary "macroanalysis" in the context of the ancient world and offer some exploration of how the ancient world is represented in the 19th-century literary imagination.

My primary objective in this paper is to introduce some methods that fall under the umbrella of what I call "macroanalysis." Despite my title, I'm not going to be drawing any grand conclusions about the ancient world. My hope is that by sharing these approaches I will stimulate some thought and give classicists a comfortable and familiar footing onto which I can introduce some foreign methodologies 1

A couple of years ago, a small group of us in the Stanford Literary Lab became interested in literary geography, and we began work on a research project to map the places that appear in a corpus of 3500 19th century novels.[1] Obviously this work cannot be done by hand, or, more accurately, it cannot be done quickly and easily by hand. It turns out that this is not a simple problem to solve computationally either. We faced two primary problems. 2

First was the problem of *identification*. How could we use computers to accurately identify place-names in novels? The obvious approach of using gazetteers proved problematic. In one gazetteer we consulted, *Providence* was a place and so was *Hope*: These may be places, yes, but they are also something else. Simply running a search for all places in the gazetteer that also appeared in the fiction would have yielded far too many false positives. 3

Second was the problem of *ambiguity*; place names are ambiguous. *Charlotte*, for example, is used only as a first name in our corpus and never as a city. *Florence* is almost always a character but occasionally a city in Italy. *Charlton*, *Denver*, *Albany*, *Hastings*, *Belmont*, *Gresham*, *Wilmington*, and *Woodstock* are all in our corpus, and all more commonly used as last names than as cities. A second problem of ambiguity is that place names are often reused in different locations: *Richmond* is both a town in southwest London and a city in Virginia. *Dartmouth* is a city in England, the U.S., Canada, and Australia. *Georgia* is both a U.S. state and a country in Eastern Europe. 4

To tackle the first problem of identification, we scrapped the gazetteer in favor of Named Entity Recognition (NER). NER is a Natural Language Processing (NLP) tool that identifies places using a trained statistical model that is sensitive to semantic and syntactic information in the text. NER is not perfect; it sometimes thinks that Florence the character is Florence the city, but we devised a way of muting that problem of place ambiguity using a technique called topic modeling. 5

Topic modeling is a complex process that cannot be explained in detail here. I have written a useful explanation on my blog.[2] For now, you can think of books like plates that you fill up at a buffet full of topics, and in this case the topics are places. Imagine that Jane Austen and Herman Melville wandered into an imaginary buffet, but instead of choosing to fill their plates with peas and fish they chose the settings for their novels: so, Melville might choose a helping of New 6

England and a bit of South Pacific, Austen a ladle of London and a smidgen of the English Midlands. The topic modeling process has a way of identifying and measuring these textual features as they exist within their individual plates and within the “buffet” as a whole.

In other work I had done for my book, I had applied this technique as a way to identify 500 themes in a corpus of 3500 novels [Jockers 2013]. Figures 1 and 2 offer example themes that the machine returned.

Figure 1.

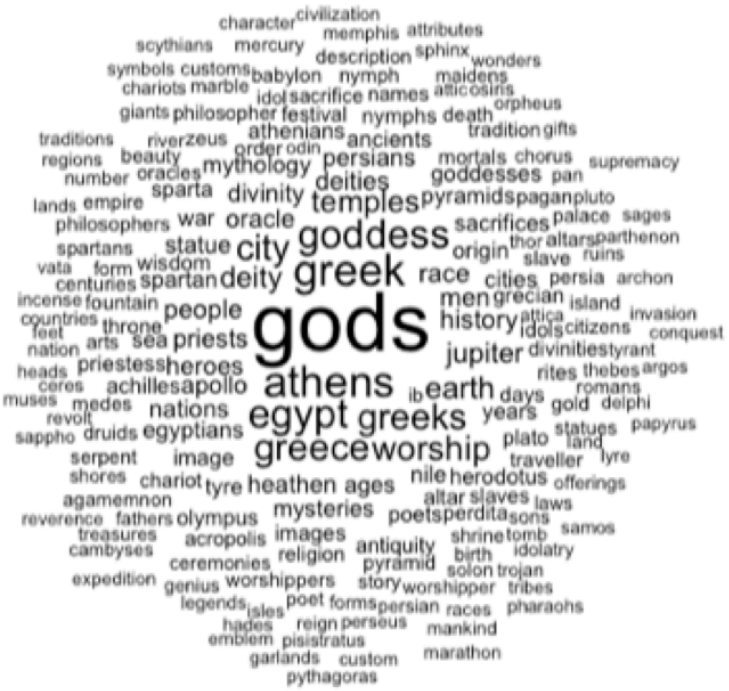

Figure 2.

The first of these is a cluster of words that might be *labeled* or *described* as *wilderness and frontier*. [3] The second we might call *ancient gods and goddesses*.[4] 8

It turns out that a similar approach can be used to effectively identify and disambiguate the place names in the corpus. This topic modeling method works because when a writer sets a book in a particular place that writer tends to mention other places that are geographically related. An author writing about Ireland, for example, might mention Dublin, Galway, and Belfast. Another author writing about Belfast, Maine, is not likely to also mention Dublin and Galway and so the machine learns that Belfast, Ireland is different from Belfast, Maine. 9

Figures 3 through 6 present four “place clusters” taken from 100 that I harvested in this research. 10


westminster bank
crotona sacred history
frederic cape trafalgar
cape de
cape corso cape bon
dinah
dockyard creek messina
monsieur
ithaca southern italy
winchester marseilles
channel
portsmouth
lunnun
mahomet
belle isle black sea
calabria
southern europe
dragon ryde
corsican
fanny
syracuse
or
va
bay
palermo goza sicilian
frenchman
sierra leone
mount etna del
tripoli non tel
piedmont
joanna ruth
europe
je pietro
bastia
la
toledo giovanni
port st .
straits
gulf
france
des
lower calabria
melita capri
malta
tunis
first
mediterranean
chancery
sicily
roma
catania
charity
upper province
st .
genoa
holland
gibraltar
santa rosalia
scylla
cagliari
elba ala valetta
la valette
san pietro
corsica
santa maura
turin
straits of gibraltar
canal greek
ma
cape st . sardinia
mediterranean sea
ajaccio
christendom
san fiorenzo
concord
fort st .
constantinople
north pole
straits of messina

Figure 3.

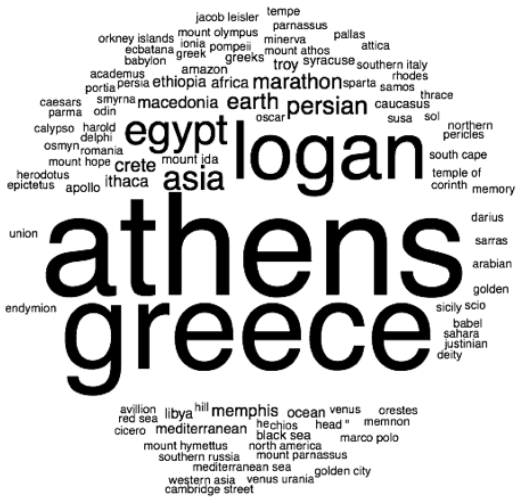
athens
greece
logan
egypt
asia
persian
earth
marathon
macedonia
ethiopia
africa
troy
syracuse
greeks
greek
pompeii
mount athos
minerva
parnassus
jacob leisler
tempe
pallas
attica
mount olympus
orkney islands
ecbatana
babylon
ionia
amazon
academus
persia
portia
caesars
parma
odin
smyrna
calypso
harold
delphi
osmyn
romania
mount hope
crete
mount ida
ithaca
apollo
herodotus
epictetus
oscar
caucasus
susa
sol
thrace
samos
sparta
rhodes
southern italy
northern
pericles
south cape
temple of
corinth
memory
darius
sarras
arabian
golden
scio
sicily
babel
sahara
justinian
deity
union
endymion
memphis
ocean
venus
orestes
libya
hill
red sea
avilion
cicero
mediterranean
hechios
head
memnon
black sea
marco polo
mount hymettus
north america
southern russia
mount parnassus
mediterranean sea
golden city
western asia
venus urania
cambridge street

Figure 4.


egypt
cairo
alexandria
baltimore
memphis
syria
minieh
suleiman
hydra
lilith
rose garden
stoneham
mohammedan
tripoli
marmorice bay
rhoda
apis
acre
perth
smyrna
pharaoh
arabian sea
yolande
africa
mount lebanon
valachia
babel
tibet
tyre
mitylene
omar
oatlands park
georgian
nilie
upper egypt
lebanon
stamboo
upper nile
mardin
rimnik
suez
syrian
helena
bosnia
venetian
mediterranean
persian
lockshan
mecca
arabic
arnaoots
on
grand cairo
alec
122 x
mekkah
crete
libya
baltic
gaza
sefton park
bayroot
nassau
hall
goshen
paros
derayeh
rodosto
ethiopia
trieste
this
upper
achmet
constantinople
lower egypt
mars
atlantic
aboukir bay
holywell
ared
greeks
mill hill
arabia
ispahan
prince of red sea
birmingham
damascus
horus
arabian
austria
king of egypt
christendom
turkish
romania
lucifer
cronstadt
york gate
pasha
southampton water

Figure 5.

Figure 6.

These clusters show that topic modeling can be used as an effective way to identify collocated place names and thereby aid in place-name disambiguation by giving us a general sense of the regions being talked about in the corpus. I found during this research that my conception of place needed to be a bit less geographic in nature and a bit less specific than what we had first imagined. That is to say, a bit less about Giza and more about Egypt, less about Crete and more about Greece. In retrospect this all seems very appropriate. In writing *Ulysses* James Joyce was not, after all, trying to write an atlas of Dublin. Joyce wanted to capture the essence of the city.[5] And so, as it happens, the places this method identified did not necessarily have to be places you could find on a map: heaven and hell were both represented in the corpus, as were Mars, Venus and Earth. More important than these outliers, the technique provided a tractable means of disambiguating different places with the same names: *Georgia*, for example, could appear inside a topic about the American South, as well as in a topic of countries in Eastern Europe (see figures 7 and 8).


belgrade .
queen charlotte
highness
inkermann
cyprus
irish sea
new publications
western
stisted
south sea islands
ashdale
scientific principles
sam
troy
mali
austria
guinea
odessa
bosnia
moscow
armenia
russia
arctic
of netherlands
amiens
pera
paris
georgia
queen victoria
veronica
turks
netherlands
by \ v .
peru
marmora
asia
st .
northern europe
avas
greece
france
santa casa
mont blanc
new timon
belgrade
kandilly
circassia
galata
western africa
northern sea
genius
turkey
seven hills
tripoli
westminster bank
aland islands
middle ages
monkey
stamboul
china
h . s .
sacred narrative
mount athos
julian
candia
mount olympus
buenos ayresj
babel
balaclava
europe
constantinople
mesopotamia
macedonia
gaza
earth
played
latin
san lazaro
scutari
illinois 823 v .
the netherlands
black sea
albania
tunis
caribbean sea
java
spain
constantine
caucasus
loretto
syria
moldavia
mackintosh
greek islands
bulgaria
ionian islands
mount ararat
cape coast
sebastopol
upper moesia

Figure 7.

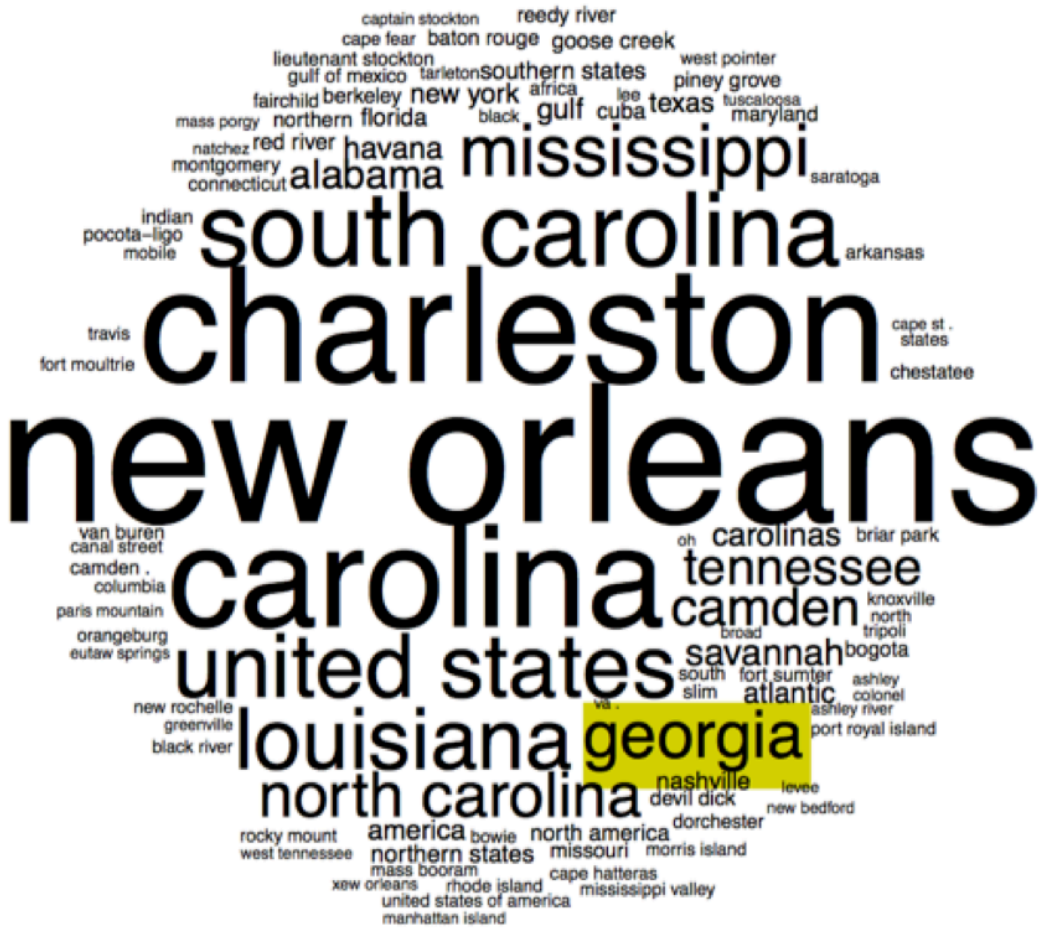

Figure 8.

Given the way the model works, we can be confident that instances of Georgia labeled by the topic model as occurring in the American South cluster are in fact mentions of the state, and those in the other topic are mentions of the country in Europe. 12

Occasionally character names slip into the geographic clusters. Note, for example, the occurrence of the name *Baltimore* in the Egypt cluster (Figure 6) or *Logan* in the Greece cluster (Figure 7). But in general these errors are not frequent enough across the entire corpus to corrupt the general sense of place conveyed in the cluster. In Figure 9, you will notice that *Illinois* shows up just below the “n” in “Ireland.” *Illinois* appears in this cluster because our text collection contains several hundred books that were originally digitized by the University of Illinois. 13

Figure 9.

These books contain bits of boilerplate metadata that we had not yet effectively extracted when I prepared these images. We have subsequently improved on this problem, but here again, the error is trivial in the larger context.[6] Which is to say that even with these obvious aberrations, the tool successfully identifies a cluster of words that captures the essence of the *place* that we know as Ireland. Irish cities and regions — Dublin, Munster, Cork, Tipperary, Ulster, and so on, all surround the headword, “Ireland.” Alongside these are smaller but obviously Irish localities: Galway Bay, Wicklow, etc. It is an imperfect method for identifying pure distinct places, but it is not at all ineffectual or unproductive in the identification of *place* as defined more broadly and thought of in terms of *literary representation of place* rather than pure geography. In that sense it is a very bad method for geographers and a very good method for literary scholars. 14

Not only does the topic model data help to give us a sense of the dominant places in the corpus as a whole, the model also provides us with a book-by-book measure of the proportion, or percentage, of each place or theme in each text. As we might expect, the model returned proportional data that showed that the Irish authors in the corpus were far more likely to write about Ireland than their English or American peers. This kind of data can be used to plot literary attention to place over time; figure 10 presents a picture of attention to Ireland over time broken out by author nationality. 15

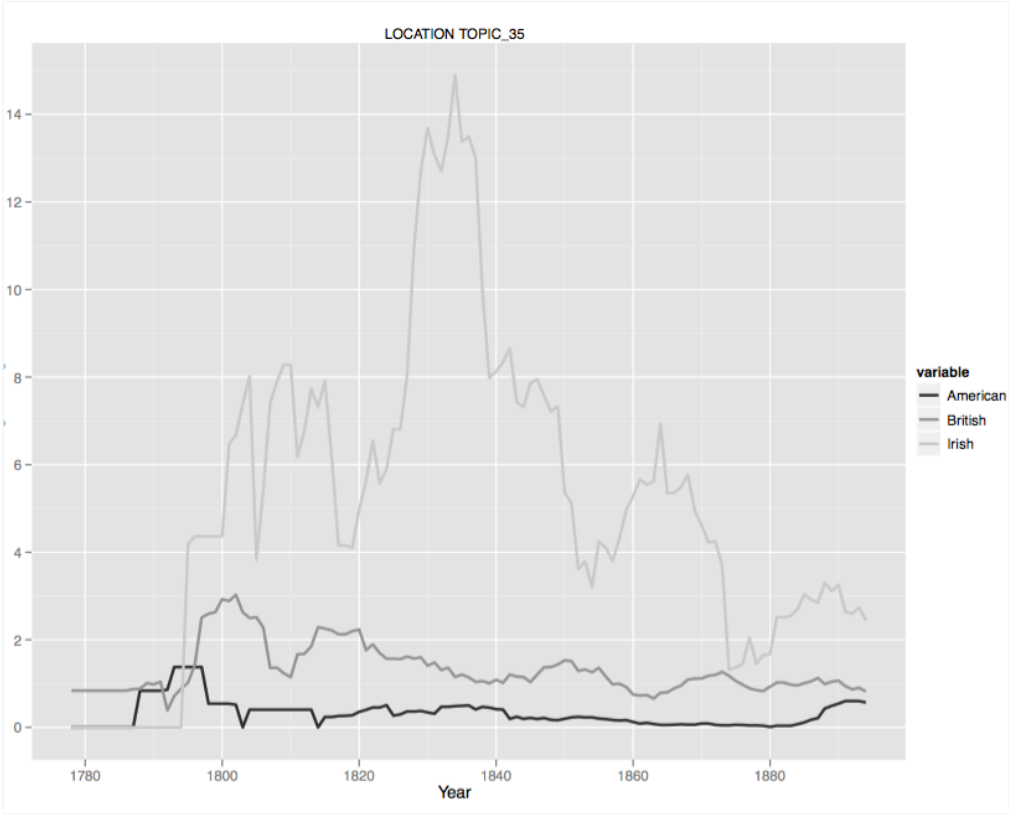

Figure 10.

The y-axis in Figure 10 is a measure of the yearly mean proportion of the Ireland cluster. Which is to say that in 1839, to pick a random year, about 14% of the places mentioned in Irish-authored books were Ireland. 16

To get the full sense of how Ireland is being portrayed within these particular books would seem to require closer inspection, and, at the scale of a few books, we could achieve this by close reading. As a method for gaining a fuller sense of how a place is represented across the entire corpus, however, close reading is just not plausible. 17

The method I'm describing here is effective if we really wish to study the representation of place on a broad scale. The question becomes not, *How is Ireland depicted by the Irish author Maria Edgeworth?*, but *How is Ireland depicted in the 19th century Irish novel?* Or, to use another example, when authors are writing about *slavery* in the *American South*, what words do they employ that express perspective, attitude, or sentiment; and even more importantly for literary history, how, if at all, do these representations change over time, across author genders, and across author nationalities (see Figure 11). 18

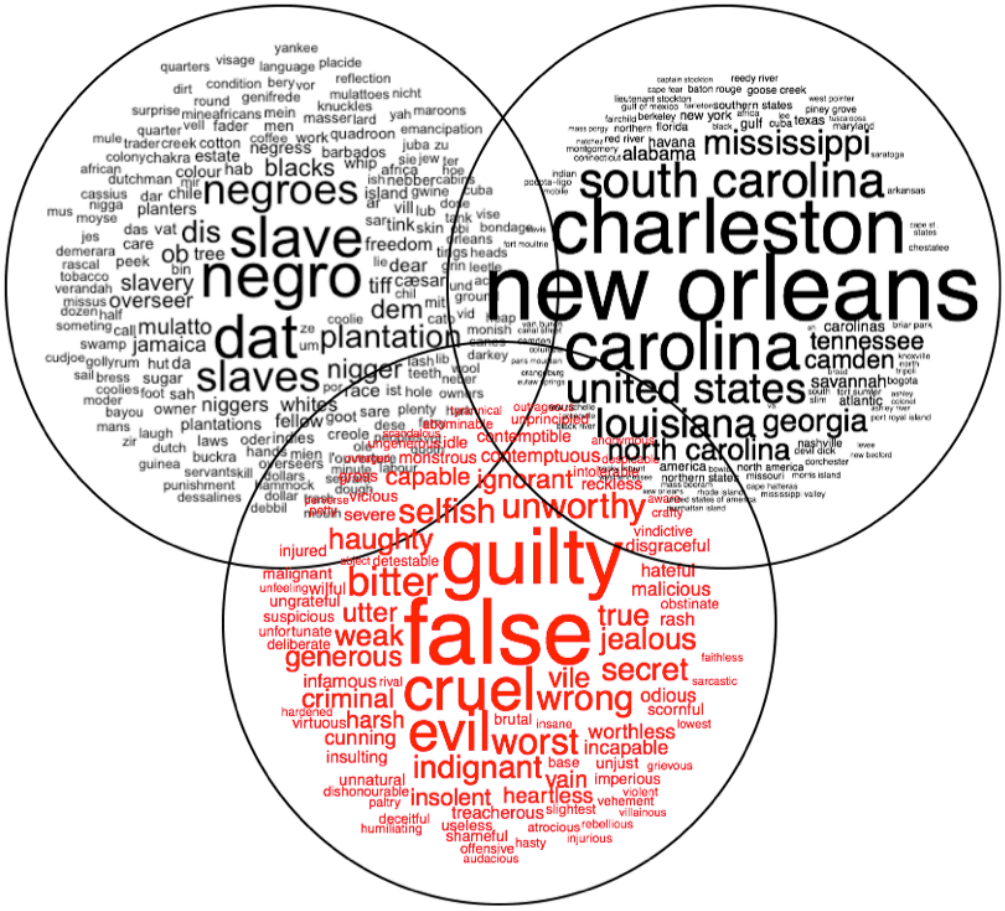

Figure 11.

Figure 11 introduces a third element of analysis: so far I have talked about place and theme, and here I introduce *affect* or what is more commonly referred to as *sentiment*. Using techniques similar to those described above, I generated a set of 25 sentiment clusters. You can see one of them in Figure 11 (emphasized in red). For this work, I drew on research in the fields of sentiment analysis and opinion mining in order to develop a method of scoring the topical clusters. In my system, each cluster is scored on a scale from -1 to +1 where -1 is a very negative perspective (colored in red) and +1 is positive sentiment that is colored green; 0 is neutral and yellow. The scoring here is based in large part upon a list of 6800 positive and negative opinion/sentiment words for English collected by Bing Liu of the University of Illinois at Chicago in his *opinion lexicon* [Minqing and Liu 2004]. [7] Space does not permit going into the details of the methodology here; I provide instead five more examples from the total of 25. Figure 11 shows a negative sentiment characterized by the large headwords of *guilt* and *cruelty*. Figure 12 shows five more clusters conveying (from top to bottom and then middle) that which is *beautiful,* that which is *amiable*, that which is *wretched,* that which is *dead*, and that which is *temperate*.

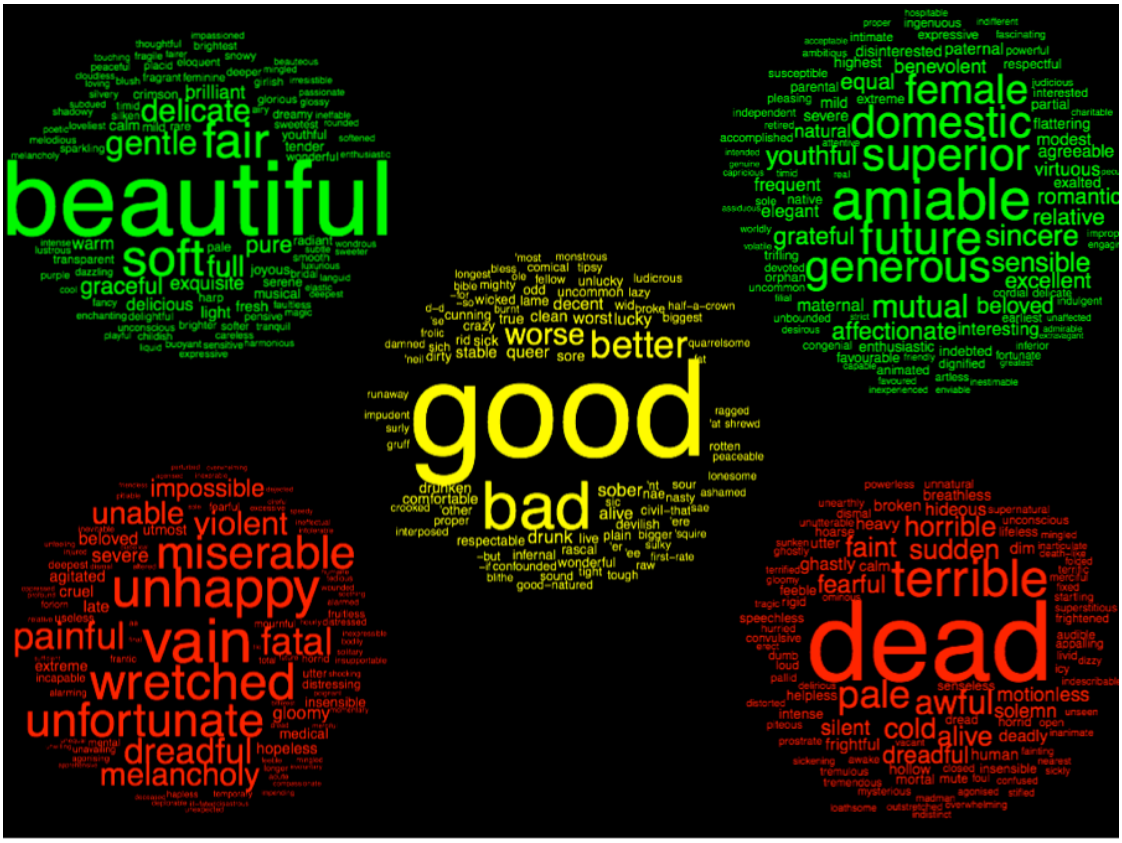

Figure 12.

Just like the thematic and geographic clusters previously discussd, the presence of these sentiments is measured across the corpus and across every book in the corpus. So, for example, the data suggest that the book in my collection with the most sustained expressions of *guilt* and *cruelty* is an Irish-authored novel appropriately titled *The Victims of Society*. 20

By simultaneously analyzing data about sentiment alongside the data for place and theme, we can begin to explore and identify areas of the corpus and specific books in the corpus where certain places are statistically correlated with particular themes and/or particular emotions. Unsurprisingly, the theme most closely associated with Ireland, is a theme I have labeled *Tenants and Landlords* (Figure 13). 21

Figure 13.

I say “unsurprising” because the history of Ireland in the 19th century is largely a history of tensions between the landed and tenant classes. What was interesting about these correlations was that they varied dramatically depending on author nationality. So, for example, in novels written by Irish authors, the tenant landlord theme was correlated with representations of Ireland. When only English authors were examined, the statistical correlation dropped to a level indicating no correlation whatsoever. What this suggests, then, is that it is the Irish authors and not the English authors who are fixating on the tenant landlord situation within a specifically Irish context. Put into plain terms: English authors, when talking about Ireland, tend to avoid mention of the Tenant and Landlord situation. 22

Other themes that correlated closely with Ireland included those in Table 1. I was rather disheartened to see *drinking* topping the list since it seems to confirm so many of the stereotypes we have about the Irish, but when I parsed the data, I discovered something surprising and unexpected. 23

Drinking: Liquor and Beer
Livestock and Produce
Peasant Dwellings
Family, Friends, and Neighbors
Villains and Traitors
Government and Rebellion
Police and Magistrates
America
Revenge and Vengeance
Personal Character
Crowds and Mobs
Habits and Customs
Hatred and Jealousy

Table 1.

24

At the scale of the entire corpus, *Ireland* and *drinking* were highly correlated. When I examined the data for Irish authors alone, however, I discovered that there was not such a high correlation. The Irish wrote about Ireland and the Irish wrote about drinking, but the two things were not generally written about simultaneously. Seeing this, I naturally expected to find that it was the English authors who were tending to write about the two things simultaneously. I was wrong. How then could *drinking* have been correlated with *Ireland* at the level of the whole corpus? It turns out we have the temperate Puritans to blame. It is the American authors who most closely associate *Ireland* with *drinking*. Indeed, in American books where *drinking* is a dominant theme, *Ireland* is the place most frequently mentioned. Also closely associated is the theme of *Sin and Repentance*! So, at least in terms of this corpus, it would seem that a good deal of the blame for the stereotype of the hard drinking Irish sits squarely on American shoulders.

25

I now return to the four place clusters and terrain that will be more familiar to classicists (seen together in Figure 14).

Figure 14.

I have culled these four out of the larger set of 100 because these four seem to be most closely associated with the ancient world; or at least to literary representations of the ancient world as it existed in the 19th-century literary imagination. For each of these four place clusters, it is possible to track literary attention over time, by author sex and by author nationality. Figures 15-17 graph a few of these dimensions for illustration.

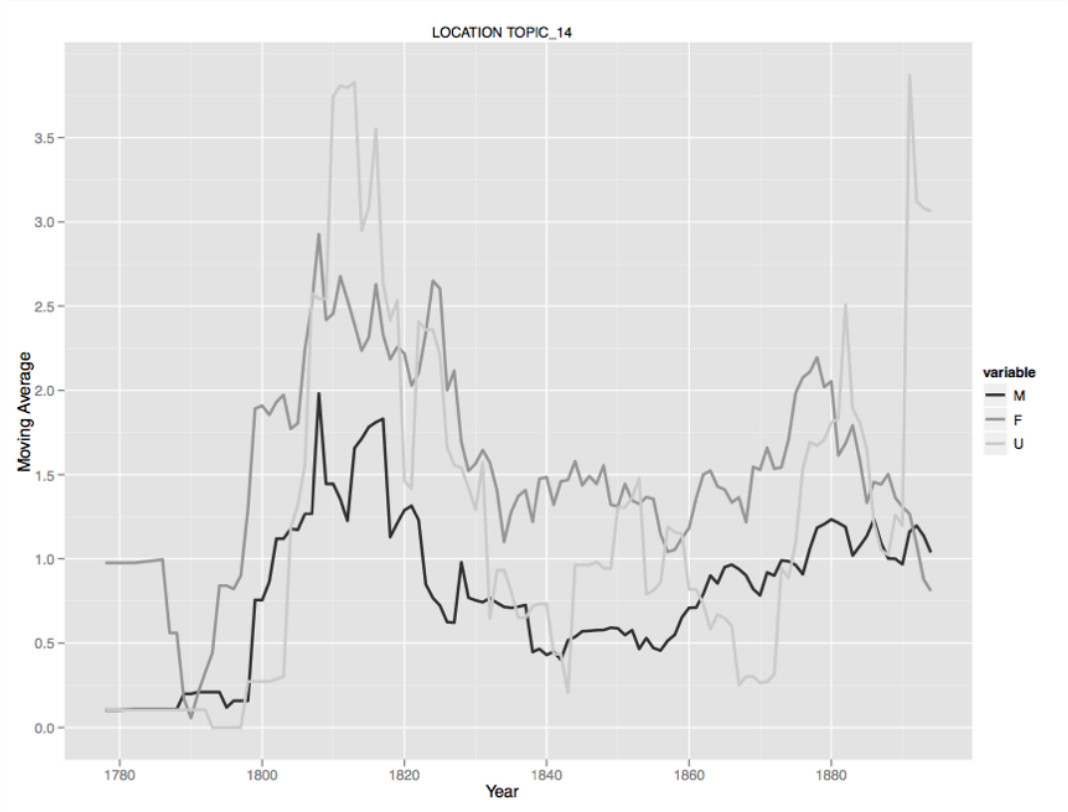

Figure 15. *Egypt* when explored in terms of *time* and *author gender*

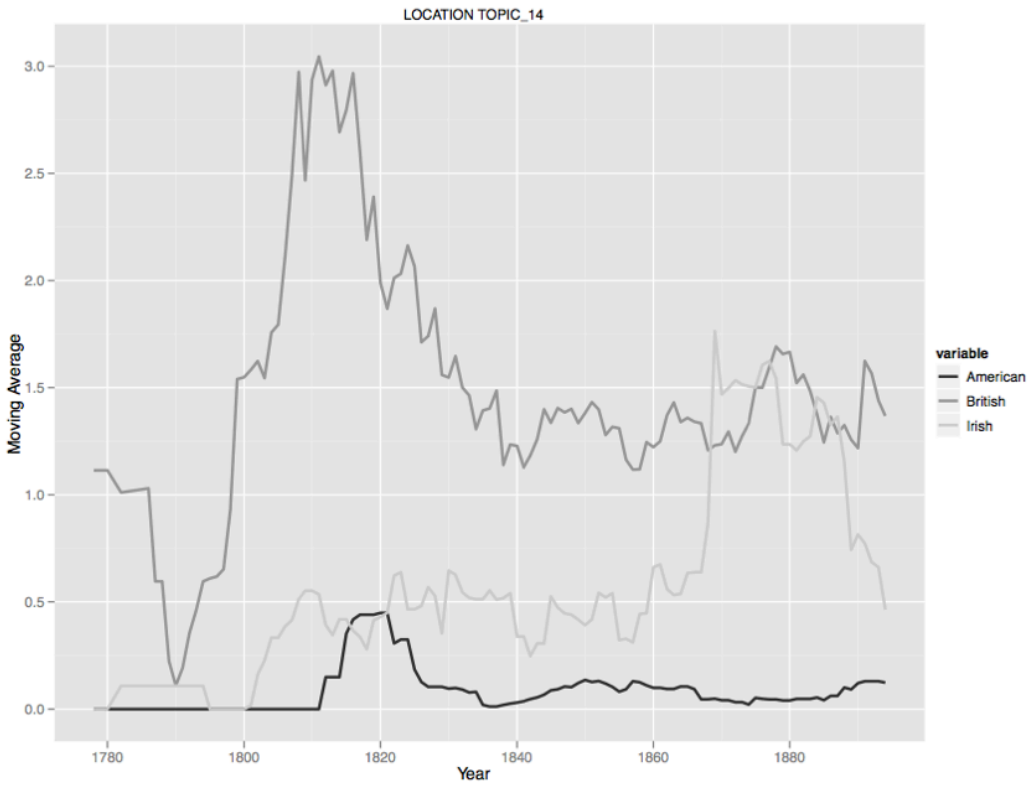

Figure 16. *Egypt* when explored in terms of *time* and *author nation*

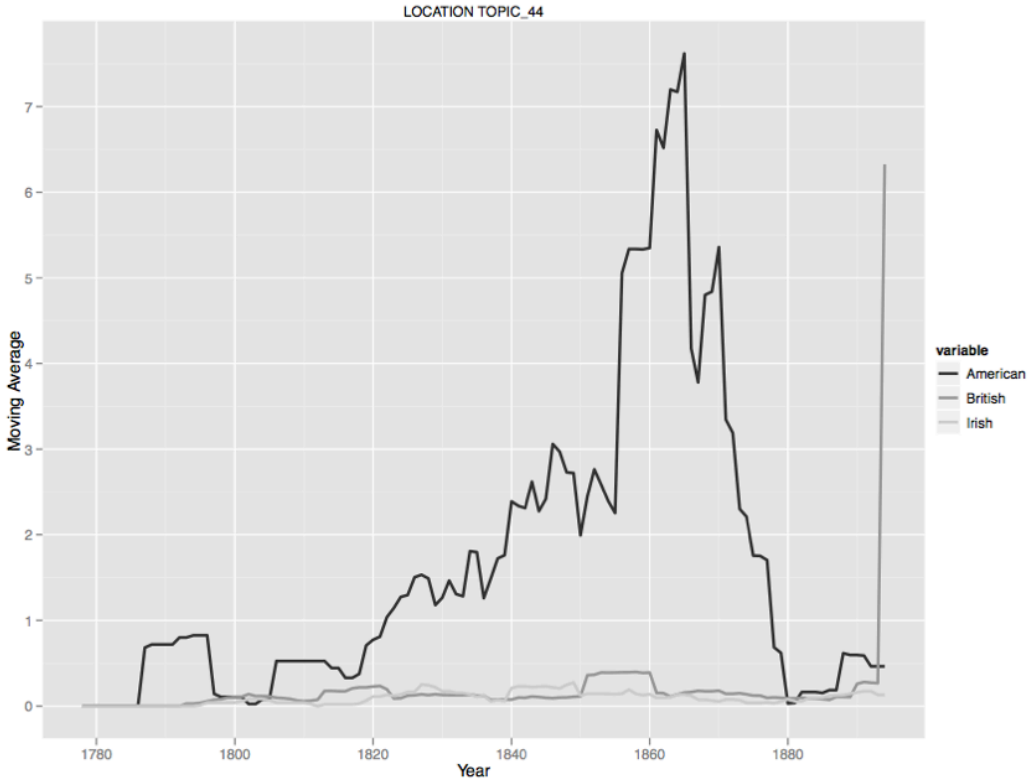

Figure 17. *Mediterranean* in terms of *time* and author *nation*

In addition to graphing, or tracking, literary attention to these places over time, I can also examine how the appearance of certain locales is correlated with certain themes and sentiments, just as in the previous example of *Ireland*. 27

When representations of *Egypt* are strong, the most highly correlated themes include *Persian slavery, Gods and Goddesses, Deserts, Art and Beauty, Earth and the Universe*. The most significant sentiments include that which is *Magnificent and Splendid,* that which is *Holy and Sacred,* that which is *Beautiful, and* that which is *Moral.* 28

When representation of *Greece* are strong, the most highly correlated themes include *Gods and Goddesses, Philosophy and Wisdom, Spirit and Soul, Art and Beauty, and Poetry*. The most significant sentiments include that which is *Magnificent and Splendid,* that which is *Holy and Sacred,* that which is *Beautiful,* and that which is *Moral*.[8] 29

When representation of *Jerusalem* are strong, the most highly correlated themes include *Jews, Land, Art and Beauty, Deserts, Processions and Spectacles, Sin and Salvation, Victory in War,* and *Heaven and the Soul*. The most significant sentiments include that which is *Holy and Sacred,* and that which is *Magnificent and Splendid.* 30

When representation of the *Mediterranean* are strong, the most highly correlated themes include *Ships and their Crews, Outlaws and Robbers, Natural Beauty, Ruins, Female Heroines, Sea Voyages, Men with Guns,* and *Servants.* This is a very different thematic profile, and the most significant sentiments include that which is *Fair and Mild,* that which is *Magnificent and Splendid,* and that which is *Unhappy and Wretched.* 31

It is tempting to move from these macroscale observations and correlations to deeper interpretations about the specific books in the corpus or the specific writers, and this movement between scales is precisely what I am advocating in my book, *Macroanalysis*. Instead of rehearsing those arguments here, I want to conclude with an observation of a more general nature. First a warning about the very sort of research I am engaged in: we need to keep in mind that this type of macroanalysis only reveals the larger, general tendencies in the corpus. The most we can say at this scale is that certain things tend to occur more often, or less often, than others; what we capture here are the general tendencies 32

within the corpus. Having said that, I would also note that it is equally problematic to speak only of particulars. In other words, it is just as dangerous to move from distant readings to specific conclusions as it to move from close reading to general theories of literary history. Just as we would not expect an economist to generate a sound theory about the economy based on the behavior of few workers in his or her neighborhood, I don't think we can generate sound theories of literary history by reading only a few books.

Notes

[1] Initial results were presented as a co-authored paper at the 37th Annual Meeting of the Social Science History Association: Allen, Ben, Cameron Blevins, Ryan Heuser and Matthew L. Jockers. "A Geography of Nineteenth-Century English and American Literature." Social Science History Association, 37th Annual Meeting, Vancouver, British Columbia, November 2, 2012.

[2] See http://www.matthewjockers.net/2011/09/29/the-lda-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/

[3] These word clouds are generated based on the weighted probabilistic data that the topic modeling process returns. Larger words are more central to the overall topic.

[4] You can find all 500 themes on my web site: http://www.matthewjockers.net/macroanalysisbook/macro-themes/

[5] This, despite that fact that Joyce once jokingly bragged that if Dublin were destroyed "it would be possible to rebuild the entire city, brick by brick, using *Ulysses*." (For details see http://www.irishleftreview.org/2010/06/18/dublin-psychogeographical-society-bloomsday-special-3/)

[6] To improve the process, we took two approaches to this problem. First, we removed from our consideration all places whose ambiguity we identified as especially pernicious. Second, we ranked the places that the tagger had identified from most to least frequent. We skimmed off the top 1000 places, which accounted for more than 85% of all mentions of all places that the NER tool had detected. We then divided these into four groups of 250 places and each of us, that is myself, Ben Allen, Cameron Blevins, Ryan Heuser, identified and read a random sample of 25 sentences in which each place occurred (we wrote a simple query script for this). If a given place occurred more than 12 times as a last name or other noun, or if the word changed the place to which it referred more than 12 times, we marked down the place in a blacklist. We were left with about 650 places we could trust.

[7] Also see http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[8] Note that these sentiments are identical to those for Egypt

Works Cited

Jockers 2013 Jockers, Matthew. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.

Minqing and Liu 2004 Minqing Hu and Bing Liu. "Mining Opinion Features in Customer Reviews." *Proceedings of Nineteenth National Conference on Artificial Intelligence* (AAAI-2004), San Jose, USA, July 2004.