Place Names and Geography in the Gutenberg Corpus

TEAM: Yellow Octopus
NAMES: Justin Clark, Ethan Davis, Hannah Tosi, Jackie Littlefield
DATE: 6 December 2019

Abstract

Reading literature is often equated to travel, allowing the reader to participate in the geographic imagination of a culture. Literature has the potential to create geographies of desire, migration, colonialism, and fantasy. This project explores geographies of imagination through mentions of places in English literature. The Gutenberg Corpus provides documents written in and translated into English. This analysis studies mentions of place-names in the documents over time and investigates the relationships between places mentioned. The specific research points addressed in this report are which places are mentioned in the corpus, with what frequency they are mentioned, and how that frequency changes over time.

The "geograpy" library in Python finds place names in literature. Using the VACC provided by the University of Vermont, this library was applied to every book in the Gutenberg Corpus, generating a data set containing every place name mentioned in every single book. This data was then combined with the Geonames dataset, an online collection of the name, country, and coordinates of every town and place in the world. Using this information, the results of the "geograpy" analysis were cross-referenced so that incorrectly identified names could be removed.

Time series plots were created from the data under different conditions such as truncation to certain time periods and selection of most-mentioned places. Examination of these plots can provide insight as to which countries were post well known at certain times. These plots have some noise due to false positives. However, overall, it is possible to draw historical conclusions from the data in the plots.

Introduction

The history of the English language in modernity is one of travel. With the advent of the printing press, evidence of the expanding scope of the British nation state and colonial empire was available to a growing literate readership. Books and writing allowed people to learn about what was happening in their own countries, as well as discovering new places they had not heard of before. As the number of readers grew, so did the metaphorical territory covered by writers. More and more places were mentioned in books as the world became more literate and more connected.

Some background into why this topic of place name evolution intrigued us, is the article "Quantitative patterns of stylistic influence in the evolution of literature", Hughes et al. (2012). In the article they do a statistical analysis of variations in literary styles between one writer, a genre, or across time periods. However, they do not reference how places mentioned in literature change over time, which also could really tie into variations in literary styles. For example, a certain place could be mentioned more times in a specific literary style and not so much in a different literary style.

The purpose of this project was to analyze the city names mentioned in English literature over time to understand how cultural knowledge of places evolved. This research was motivated by the idea that peoples' knowledge of other places could be mapped by taking a time series of the place-name mentions in literature. As people became more aware of other places, they mentioned them more in literature, so plotting these growing mentions helps show how knowledge has spread over time. When constructing a history of English-language literature, it is important to consider the changing boundaries of the British nation, the English language, and international travel. These socio-political changes can greatly affect cultural imagination and commercial desire. As the geographic boundaries of a culture change, the boundaries of its literary imagination are expected to change as well.

There has been considerable recent interest in the digital humanities, and the combination of literary and geographical data. Significant contributions to the expanding field have been concerned with topics ranging from literary style to historical public health. The utility of spatial analysis of literature is not limited to literary insight. For example, Porter et al. (2018) used a corpus of articles from the London weekly the *Era* to analyze public health in 19th century London. This study has implications in the computational analysis of historical public health.

Spatial analysis is also very useful for understanding geographic changes in human culture. For example, Franco Moretti's "Atlas of the European Novel" makes significant headway in the geographic analysis of English literature (Moretti 1999). Though innovative in its analysis, it is somewhat limited in its scope. Similarly, Flores et. al (2017), as well as Cooper et. al (2011), used travel writings combined with GIS to create spatial models of the journeys of several authors through the English Lake District. The authors employed Cost-Surface Analysis and Least-Cost-Path analysis to create likely travel paths from itineraries extracted from the literature. This study successfully employed location data and computational methods on a relatively small geographic scale to reconstruct otherwise concealed historical insight. Another article, branched off from this topic and looked at how different fields of study or times, like the medieval times, not just about different places/spaces mentioned but also how those places/spaces were understood, Flores et. al (2017).

Although there are many articles that dive into this idea of tying literature to geographical data, many like Flores et al (2019), find that it is only touching the service of what can be

done by combining these two worlds and that there is so much more that can be explored. As well as more insight to be gained from the analysis of such a combination.

For this project, we propose an analysis of geography though individual place-name mentions by document for all documents in a large corpus. This project is of sociological interest in that it could provide unique insight into the relationship between the geographic history of a culture and the written culture it generates, or between the geographic and cultural boundaries of a nation and language. The dataset used for this analysis was the Gutenberg Corpus. Named after the inventor of the printing press, the Gutenberg Corpus is an online collection of tens of thousands of books dating back to Homer and continuing through to current publications.

We will address the following questions in our results:

- How did mentions of place names in literature change over time?
- Do some place names seem to dominate?

Data and Materials

Data

The main source of data was the Gutenberg Corpus of online eBooks. This provided a nearly exhaustive collection of every literary work written or translated into English (about 60,000 works in all). The corpus is available online and is free for everyone. It was developed so that literature would be more accessible to people. Most of the books included have expired copyrights, so there is no need to pay to read them. Each book can be read in HTML format, and included with each book is publication information.

GeoNames is a database of geographical names with over 25 million entries, andis free under creative commons. (Geonames) We used a subset of the GeoNames database consisting of all countries, and cities with a population of over 15,000.

The statistical program R contains a specific library devoted to the Gutenberg Corpus, “gutenbergr”. Contained in this library is data about every book in the corpus, including information about the authors. This was used to create the file “authordata.csv”, which contains information on the name, birth date, death date, and Gutenberg identification number for every author who appears in the corpus. Because information about publication dates is not always accurate, especially for earlier works, identification is based on author birthdate. Books with authors without birth dates entries were excluded from the dataset.

Data Manipulation

The “geograpy” module in Python is a machine-learning named entity recognition application that searches through text for place names. This was run on every book from the Gutenberg corpus, generating a json file of data for each book. Contained in each file is the name of every word flagged as a city name, and every word flagged as a country name. We used the “strip_header” function of the “gutenberg” python library to exclude headers and footers.

“geograpy” returned many false positives, so we cross-referenced the results with the names in our subset of the GeoNames database, only keeping those that matched exactly. There was no reliable way to correctly match cities with the correct countries, so we performed analysis for cities and countries separately.

We merged the datasets for author birth year and place mentions or each book. We then created a table of counts for each place name per year. Counts were recorded for both the number of books mentioning a place name, and number of individual mentions of the place name. We binned the data to show counts for each decade rather than each individual year, and performed time series analysis on the data.

A pipeline to the UVM Vermont Advanced Computing Core (VACC) was created, allowing these computations to be completed quickly and in parallel.

Results

The plots below are visualizations of hypotheses we developed prior to examining the data. To answer the research questions, several methods were discussed and analyzed. It was decided that the most useful way to display conclusions about the data is to create time series plots from it. These plots are very flexible, as any number of countries and time periods can be included and compared. Because the data set is so large, there are an infinite number of possible plots that can be made. However, for the purposes of these research questions, only a few were useful and necessary. A few of the generated plots are shown below with their analyses.

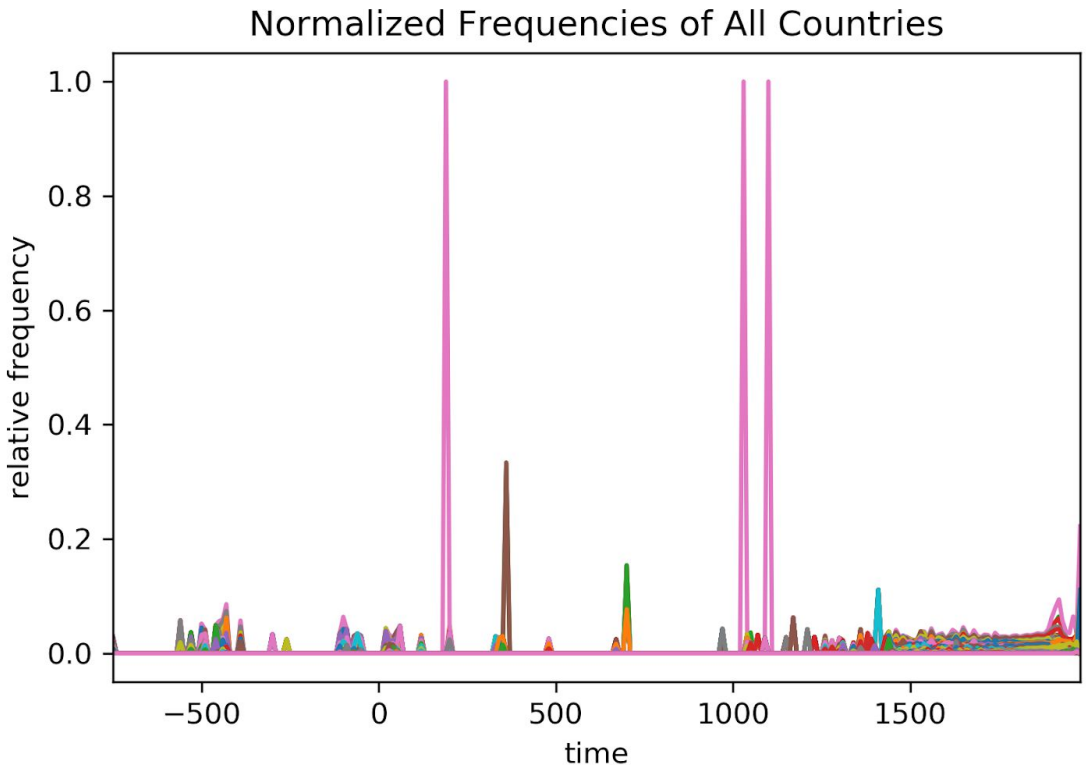


Figure 1. *This is a graph of the number of times each country was mentioned in a piece of literature beginning with Homer and continuing through the 20th century.*

The graph shows mentions of many countries before the year 500, which is unlikely to have occured in literature from that time. Some of the country names plotted did not exist, so there is clearly noise appearing in the plot that is not from the original data. Another feature likely caused by noise are the enormous peaks seen for a few countries beginning in about the year 250. For example, in this graph, there are three years in which the United States was the only country mentioned. All three of these years occurred before 1776, which indicates that there were inaccuracies in the machine-learning model that identified country names from the literature.

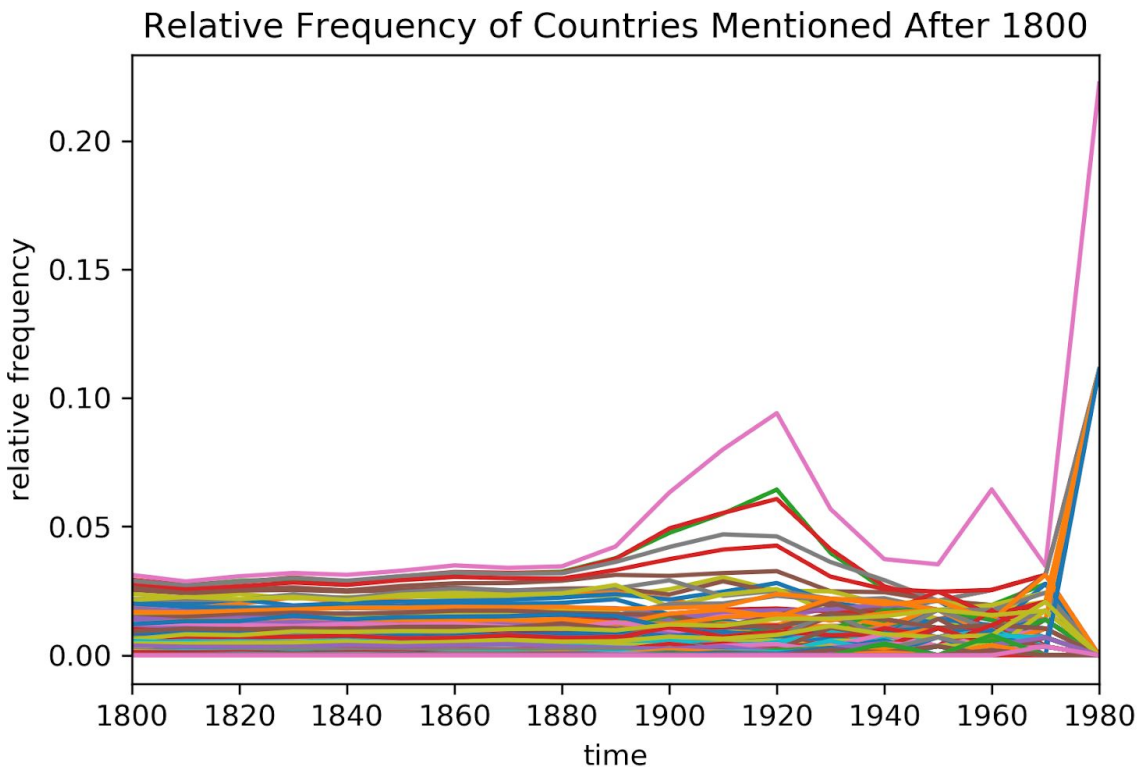


Figure 2. *This plot shows how often every country was mentioned in literature per year after 1800. The values are the relative frequency of each country's mention compared to all the others for each given year.*

Beginning with 1800, the United States is consistently the most-mentioned country. Discussion of the United States appears to peak three times: first between 1900 and 1920, then at about 1960, and then mentions skyrocket at 1970. It is possible that some of this is due to noise. However, these increases also correspond to important historical events. In the early 20th century, both Ford and personal refrigeration were invented in the United States, and World War I, the Dust Bowl, and the Great Depression all occurred. In the 1960s, the Green Revolution and the Civil Rights Movement got a lot of press. As the US became more powerful, it was mentioned in more and more pieces of literature, as can be seen in the increase in mentions starting around 1980.

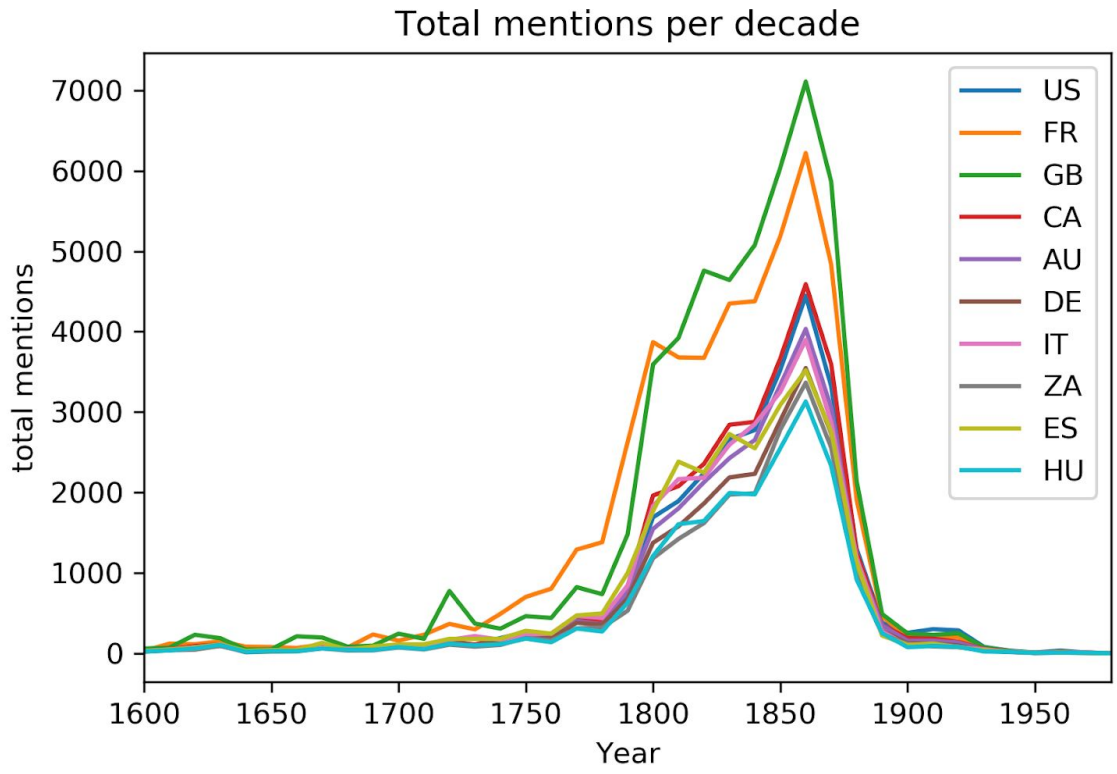


Figure 3. *This is a plot of the ten most-mentioned countries per decade beginning in 1600 and continuing through the 20th century.*

In Figure 1 and Figure 2, the United States appears to be the most-mentioned country overall. In Figure 3, Great Britain appears to be the most-mentioned until about 1880, at which point the United States becomes the most-mentioned for a short time. However, by 1950, all top-ten countries appear to be mentioned the same amount. In Figure 2, the United States appears to gain frequency starting in about 1880. This corresponds to what is shown in Figure 3. It is possible that because Britain was such an enormous colonial power, it appears most frequently until the United States after the end of the Civil War.

This graph (Figure 3) shows a different distribution to that in Figure 2. This is due to binning by decade instead of year. Here, all top ten appear to be mentioned the same amount from 1925 onwards. This likely means that these countries were all mentioned an equal amount over each decade, but some may have been mentioned more in certain years than others. There is a sharp dropoff in the number of mentions around 1900, corresponding with the approximate century limit for creative commons copyright accessibility.

Interestingly, South Africa (ZA) and Hungary (HU) are both mentioned many times, along with the more powerful countries shown.

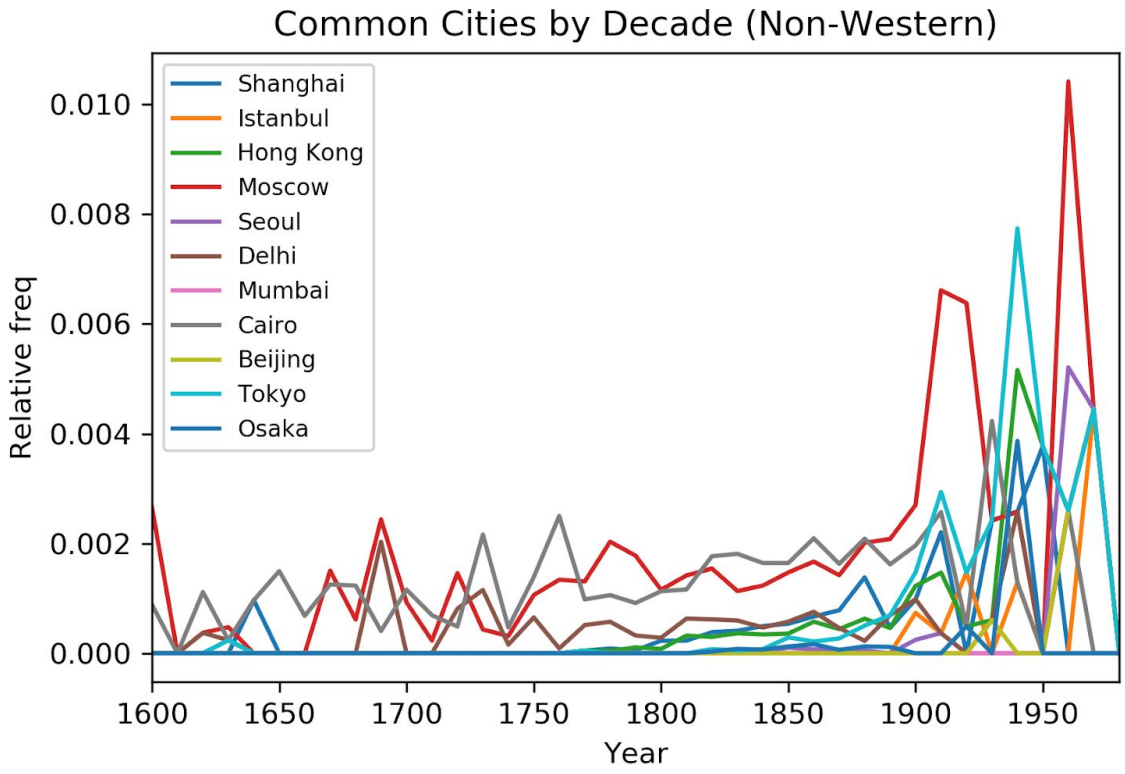


Figure 4. *This plot shows the top 11 most mentioned non-Western cities starting in 1600 and continuing through until present day.*

Because most English literature was written in the Western world, most cities mentioned are in the West. However, Eastern cultures played an enormous role in cultural development during these time periods, so it was important to examine them separately. Moscow appears to be the most mentioned for most of the time, It has a sharp peak during what may be the Cold War. Cairo is also frequently mentioned up through 1900. Tokyo becomes more frequently mentioned as time progresses.

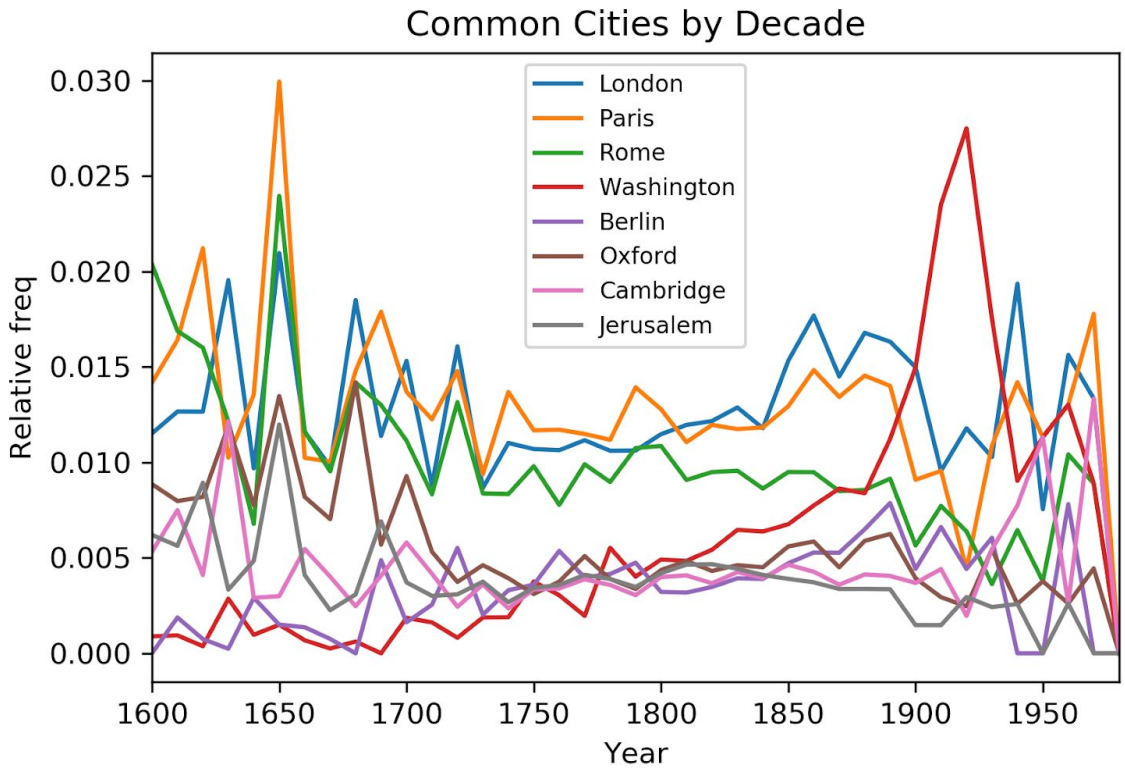


Figure 5. *This plot shows eight cities that were among the most-mentioned from 1600 through the present day.*

In Figure 4, some cities appear to gain frequency over time while others lose it. Jerusalem appears to be mentioned a moderate amount until about 1725, when it becomes one of the less-mentioned of the top eight. Conversely, Washington is not mentioned much in the 1600s. Its frequency increases gently until about 1880 when it becomes the most mentioned (and remains so through about 1925). This is likely due to events such as the Civil War and World War I, as discussed previously. However, Washington did not exist in 1600. It is possible that there were other cities by that name, but it is more likely that this is noise due to inaccuracies in the data.

Discussion and Conclusions

As previously mentioned, it is likely that some of the apparent patterns in the analysis are the results of noise. This in part exemplifies the difficulty of computational named entity recognition. Many cities and countries share names with common words or person names, and distinguishing between the two is an ongoing area of research in natural language processing. False positives may also have occurred due to occasional failures of the header

and footer stripping function, allowing copyright and digitization information with place names to be recorded.

Another place these errors could have occurred is during parsing. Every book in the Gutenberg corpus includes information about the author and the date and place of copyright. It is likely that this data was not trimmed properly for some of the books, so publication cities (and countries) may have been mixed in with names of places mentioned in the books.

Another source of error could have been the way the dates were recorded. As previously mentioned, "date" refers to the author's birth date. We did not expect this to affect large-scale analysis of trends over time because the data spans such a large time period. The results seen in graphs covering more than 200 years do not seem to have been affected by this. However, the results seen in the years 1800 and onward may have been. For example, in Figure 2, there are a few obvious peaks in the mention of specific countries (the US especially). These are addressed in the above discussion. We hypothesized that these are the result of specific historical events that occurred in the country at that time. However, this plot actually shows that authors born at those times tended to mention the US often. The books mentioning these countries at that frequency were likely written two or three decades after those peaks appear. This in itself could be an interesting result. It is possible that authors born during times of political change and unrest tended to write more about those events as they got older.

There is clearly room for more investigation. In future work it will be important to refine the data collection process to reduce false positives. It will also be important to find a way to be more lenient in filtering country names without significantly increasing false positives. This would allow the inclusion of alternate and historical country names (e.g. Gaul and the Holy Roman Empire, or America and the Indies) Some overarching trends have been identified, but it would be worthwhile to examine the results more closely. Book-by-book analysis would be of interest, especially if it were performed on a few books from each century. This would reveal more specific trends in how place-name mentions changed over time.

A related area of future research would include network analysis. A network could be created with place names as nodes with edges between places mentioned in the same book. Performed for different time periods, this could further illustrate the changing boundaries and connectedness of literary imagination over time. Plotting maps of all the locations mentioned in one book would be a great way to visualize this.

This project provided a novel, though broad, analysis of the gutenberg dataset, and illuminated several interesting areas for future work in the digital humanities.

References

Cooper, D. and Gregory, I. N. (2011), Mapping the English Lake District: a literary GIS. Transactions of the Institute of British Geographers, 36: 89-108. doi:10.1111/j.1475-5661.2010.00405.x

Hughes, J. M., Foti, N. J., Krakauer, D. C., & Rockmore, D. N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. Proceedings of the National Academy of Sciences, 109(20), 7682–7686. DOI: 10.1073/pnas.1115407109

Murrieta-flores, P., & Howell, N. (2017) Towards the Spatial Analysis of Vague and Imaginary Place and Space: Evolving the Spatial Humanities through Medieval Romance, Journal of Map & Geography Libraries, 13:1, 29-57, DOI: 10.1080/15420353.2017.1307302

Murrieta-Flores, P., and Martins B. (2019) The geospatial humanities: past, present and future, International Journal of Geographical Information Science, 33:12, 2424-2429, DOI: 10.1080/13658816.2019.1645336

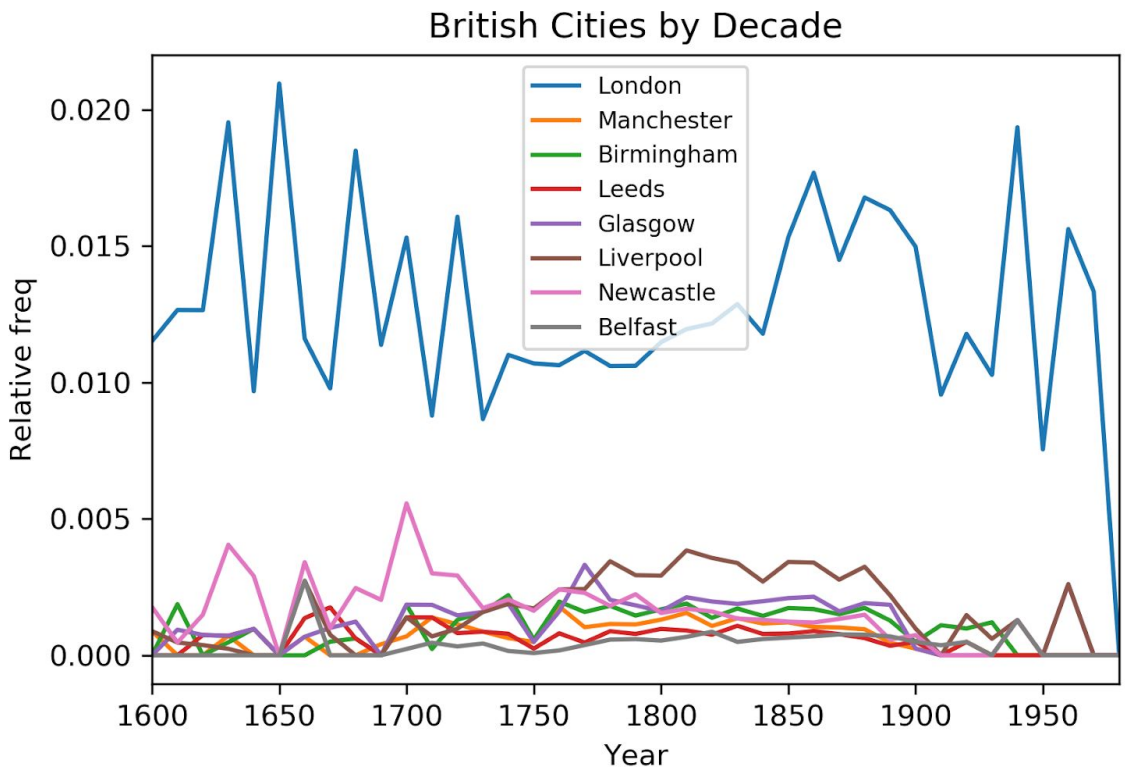
Murrieta-Flores, P., Donaldson, C., & Gregory, I.N. (2017). GIS and Literary History: Advancing Digital Humanities research through the Spatial Analysis of historical travel writing and topographical literature. *Digital Humanities Quarterly, 11*.

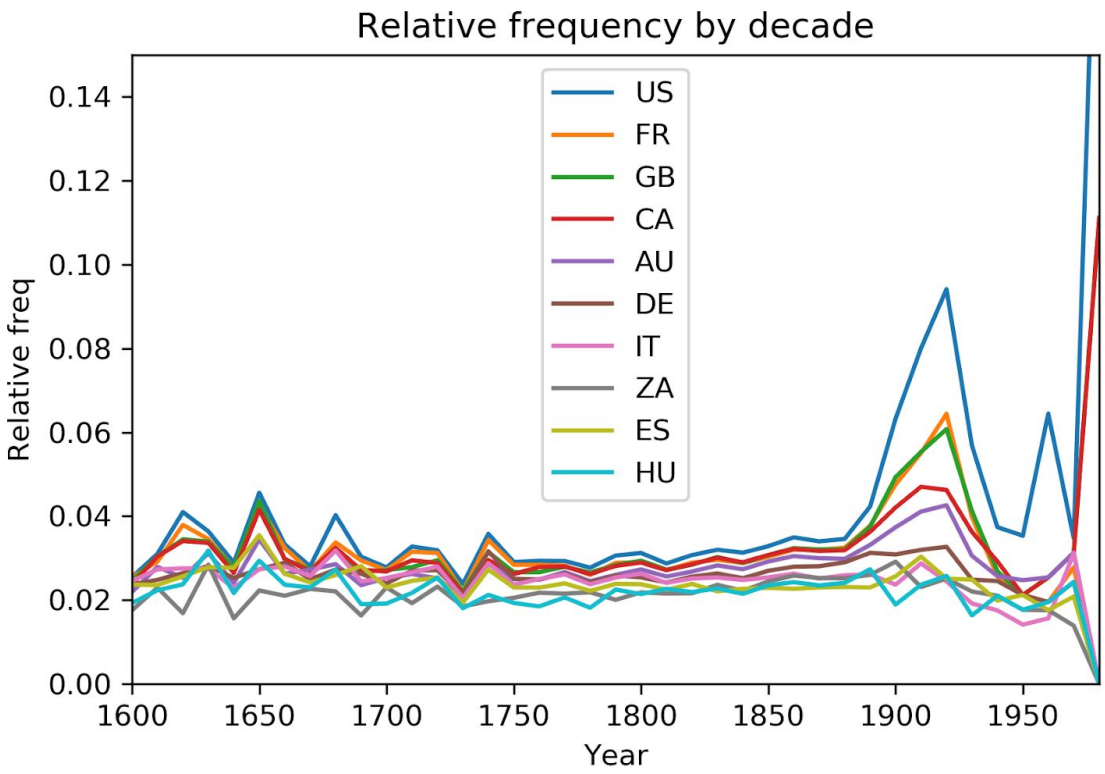
Porter, C., Atkinson, P., and Gregory, I.N. (2018) Space and Time in 100 Million Words: Health and Disease in a Nineteenth-century NewspaperInternational Journal of Humanities and Arts Computing 2018 12:2, 196-216

Moretti, F. (1999). *Atlas of the European Novel: 1800-1900*. New York: Verso.

Appendix

There are a countless number of plots that can be created from this data. It is not possible to include them all here, or even make them all in the time frame of this project. However, a few more plots were created that were not as essential in our analysis. They are included below.


British Cities by Decade
London
Manchester
Birmingham
Leeds
Glasgow
Liverpool
Newcastle
Belfast
Relative freq
0.020
0.015
0.010
0.005
0.000
1600
1650
1700
1750
1800
1850
1900
1950
Year


Relative frequency by decade
US
FR
GB
CA
AU
DE
IT
ZA
ES
HU
Relative freq
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
1600
1650
1700
1750
1800
1850
1900
1950
Year