Real and Imagined Geography at City-Scale: Sentiment Analysis of Chicago's "One Book" Program

Ana Lucic
alucic@depaul.edu
DePaul University, United States of America

John Shanahan
jshanah1@depaul.edu
DePaul University, United States of America

Since fall 2001, the Chicago Public Library (CPL) has chosen fiction and nonfiction around which to organize city-wide public events, book discussions, and other creative programming. This "One Book One Chicago" (OBOC) program has been a successful ongoing civic initiative with great public visibility, with participants ranging from the Mayor of Chicago to countless book group volunteers across the city. Our "Reading Chicago Reading" project—supported by the National Endowment for the Humanities Office of Digital Humanities and Microsoft—works to discover how text characteristics, library branch demographics, and promotional activities are linked variables that can be used to predict patron response to future OBOC titles. The OBOC program acts as a recurring experiment in data capture, for each chosen work represents a probe into library usage and, by extension, a window onto the elective reading behavior of the diverse patrons of a major American city.

This paper will report comparative circulation data for three recent OBOC choices that are Chicago-centered and three that are not Chicago-centered. The three Chicago-centered books are:

1) *The Adventures of Augie March* (1953) by Saul Bellow
2) *The Warmth of the Other Suns* (2011) by Isabel Wilkerson
3) *The Third Coast* (2014), by Thomas Dyja

The three recent non-Chicago OBOC choices are:

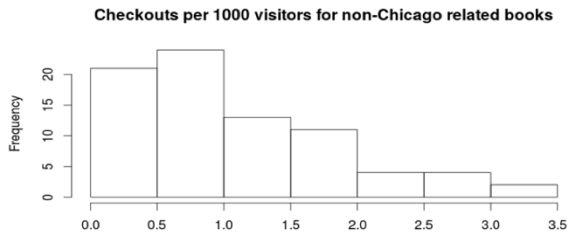
1) *The Book Thief* (2007) by Markus Zusak
2) *The Amazing Adventures of Kavalier and Clay* (2000) by Michael Chabon
3) *Gold Boy, Emerald Girl* (2010) by Yiyun Li

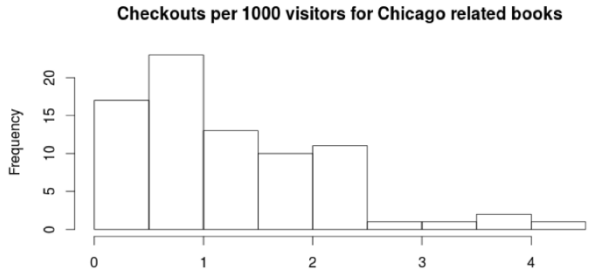
We are keen to answer whether a Chicago setting and, more particularly, particular measures of linguistic sentiment about Chicago people and places, have measurable influence on the popularity of books across Chicago. Specifically, we are interested in examining the question of whether "Chicago" books, fictional and nonfictional, checked out in greater numbers when they feature characters, events, and places situated close to the readers' own neighborhood library branch? (We use CPL library branch as a proxy for patron home address, which we cannot know from the library system's anonymized checkout data.)

The results of this analysis have the potential to provide empirical answers to long-standing questions in digital humanities research: in his Atlas of the European Novel 1800-1900, for example, Franco Moretti speculated that perhaps "fictional spaces are particularly suited to happy endings," but did not have hard numbers to judge one way or the other at the time (18 n.6). More recently, in The Bestseller Code, Jodie Archer and Matthew Jockers argue that "while it does matter whether an author chooses a city or the wilderness, the specific city does not matter all that much when it comes to bestselling" (227). Our sentiment analysis findings will contribute to open research questions: maybe in fact the city matters when readers are in that city, and when the places and people in that same city are written about in particular linguistic registers. If literary form and real geography do have detectable ties to one another, our project ought to be able to capture the effect.

To compare the circulation pattern of non-Chicago and Chicago-related OBOC books, we used one year of city-wide circulation data for each book, starting with the date of the title's public announcement as the OBOC choice. This data was normalized by dividing the circulation raw numbers with the total number of visitors for that year and multiplying the result by 1000. Normalizing by the number of circulated copies was sometimes difficult because some branches did not have any copies (but could borrow them from other branches). Given that libraries oftentimes allocate books based on the size of the library and based on the number of visitors, we decided to

normalize by the overall number of visitors. Distribution patterns for Chicago and non-Chicago related sets of books are represented through the histograms and QQ plots in Figure 1. As this analysis indicates, the circulation distribution across 79 Chicago library branches does not follow bell-shape distribution and is positively skewed. The Wilcoxon signed rank test was used to test to what degree the difference in the distribution for these two sets can be attributed to chance.





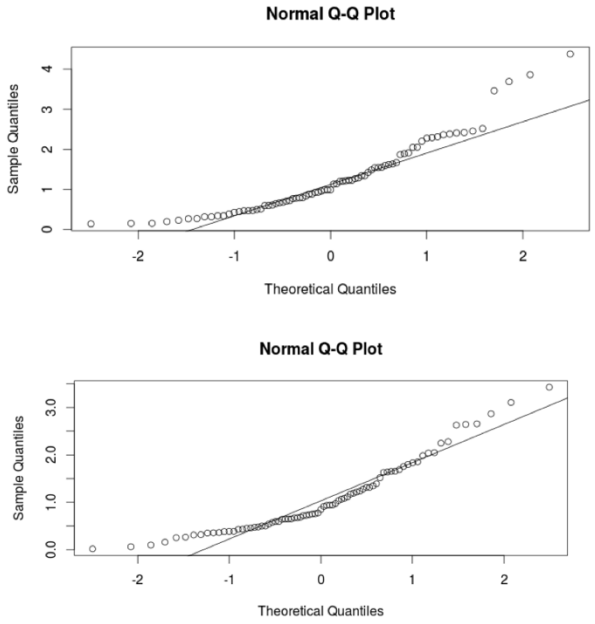


Figure 1. On top, histograms representing checkouts per 1000 visitors for non-Chicago related and Chicago related books. Below, the QQ plots for non-Chicago and Chicago related books checkouts.

Results indicate that the probability that the difference in the circulation distribution across 79 Chicago library branches for the two (paired) sets of books can be attributed to chance is very low ($p < .01$).

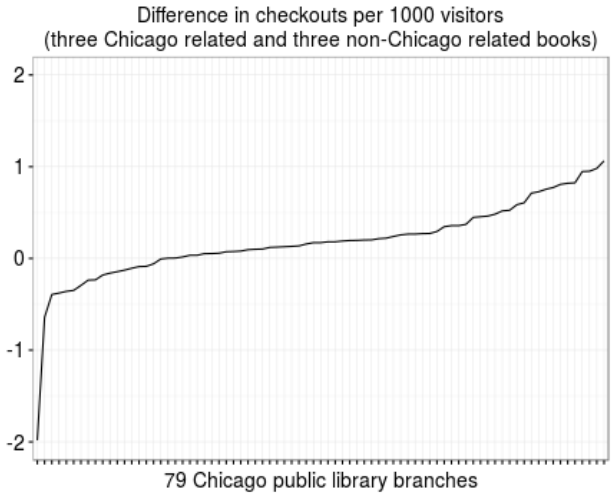


Figure 2. The y-axis indicates the difference in the checkouts (1-year of circulation data) for three Chicago related and three non-Chicago related books.

The y-axis in Figure 2 represents the difference between the checkouts per 1000 visitors for Chicago related and non-Chicago related books. Figure 2 indicates that the three non-Chicago related books

circulated more than the set of Chicago related books in some library branches in the Chicago area (where the line drops into negative difference). In some branches, however, the difference is almost minimal. The plot also indicates that, in some branches, the OBOC Chicago-related choices had, in fact, more checkouts than the non-Chicago OBOC choices (where the difference is positive). Although it is difficult to establish which factors contribute to this difference in circulation and although we cannot attribute this difference between the two distributions to the mere fact that one set contains references to Chicago whereas the other does not, we plan to represent the library branches that are associated with a larger number of checkouts for Chicago non-related books and those that are associated with a larger number of related books on the Chicago map and analyze them against the sociodemographic and socioeconomic characteristics of different branches (obtained from the American Community Survey data). In the future, we plan to add more Chicago related books to the analysis and observe how this may affect this observed pattern.

A further question of interest to us is, do the sentiment measures for these texts map in consistent ways for different neighborhoods? To examine these questions, we rely on Stanford CoreNLP natural language processing capabilities (Manning et al., 2014, http://stanfordnlp.github.io/CoreNLP/). Given that the identification of places and locations is important for our analysis, we use a tool that has consistently achieved good rankings and, in general, boasts superior accuracy rates when compared to other named entity recognizers (Rodriquez et al., 2012; Atdağ & Labatut, 2013): the Stanford Named Entity Recognizer, a part of the CoreNLP suite of tools. Before running the named entity recognizer, the text is first tokenized into sentences using the NLTK sentence tokenizer (http://www.nltk.org/). The CoreNLP program then tokenizes sentences into words, identifies lemma for each individual word, uses the Penn Treebank part of speech information (Toutanova & Manning, 2000), and also notes Persons, Locations, Time Reference, and Numbers in the sentences (Finkel et al., 2005). We are specifically interested in locations as this category would not only identify Chicago as a location but also its streets and landmark buildings. For sentiment analysis, we are using the Stanford sentiment analysis tool (Socher et al., 2013)—also part of the Stanford CoreNLP—to annotate each sentence with the sentiment score on the following scale: Very Positive, Positive Neutral, Negative, Very Negative.

Preliminary analysis on the sentiment associated with sentences that contain the word Chicago in the three Chicago-related books is indicated in Figure 3:

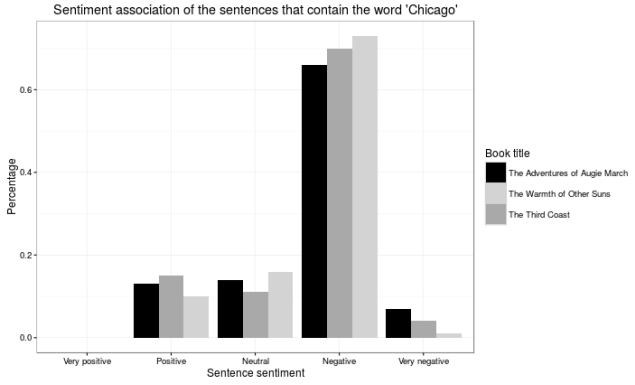


Figure 3 Sentiment distribution of sentences that contain the word Chicago in Augie March, The Warmth of Other Suns, and The Third Coast

The raw count of sentences with their sentiment ratings was normalized by the total number of sentences that contain the word Chicago. Noticeable in Figure 3 is that although these three books differ according to genre, and although they differ in terms of topical coverage and date of publication, we see a rather similar sentiment score pattern with respect to sentences that contain the word Chicago. We suspect that this overall similarity pattern will start to change as we dig deeper into the location data: we must note here that this initial analysis above does not yet take into account local references such as Pizzeria Uno, Pullman, the South Side, Monroe Street, and the like, but do not also use the word Chicago in the sentence. We plan to obtain a set of place names associated with Chicago through resources such as Open Street Maps and GeoNames and search for all the occurrences of Chicago place names. Additionally, we plan to use the indexes in the back of some of the books as trusted sources of local place names.

Bibliography

Archer, J. and Jockers, M. (2016). *The Bestseller Code: Anatomy of the Blockbuster Novel*. New York: St. Martin's Press.

Atdağ, S., & Labatut, V. (2013). A Comparison of Named Entity Recognition Tools Applied to Biographical Texts. In *2nd international conference on systems and computer science* (pp. 228–233).

https://arxiv.org/ftp/arxiv/papers/1308/1308.0661.pdf.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. Paper presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the Annual Meeting of the Association for Computational Linguistic*s.

Moretti, F. (1998). *Atlas of the European Novel 1800-1900*. London: Verso.

Rodriquez, K. J., Bryant, M., Blanke, T., & Luszczynska, M. (2012). Comparison of named entity recognition tools for raw OCR text. In *KONVENS* (pp. 410-414). http://www.oegai.at/konvens2012/proceedings/60_rodriquez12w/60_rodriquez12w.pdf

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Paper presented at the EMNLP.

Toutanova, K., and Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Paper presented at the Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13, Hong Kong.