Cultural Commentary on Wikipedia

Nathan Drezner[1], Andrew Piper[2], Richard Jean So[3]

[1]B.A. English & Computer Science; [2]Director @ .txtLab; [3]Professor in Cultural Analytics & Digital Humanities

Introduction

This project was developed to study and understand the dynamics and semantics of debate on Wikipedia, specifically by studying how language on Wikipedia pages about cultural media shifts over time. The various revisions to a page can clarify the salient language of disagreement and consensus building on a forum of collaborating volunteers, answering more complex questions prompted by the network of activity across genres. In this project we investigate the way genre, period, and medium of different cultural objects influence and inform language use and user behavior on a crowd-sourced information platform.

Abstract

Wikipedia offers a massive data set, ready to be analyzed to uncover how individuals use language and better understand questions of representation, conflict, and community standards. In this project we analyzed a set of articles based on Wikipedia lists of American media: "canons" of literature, film, and television curated by an online community. The collected data provides insight into how change occurs on Wikipedia, specifically based on how users cluster around certain cultural objects and the language that distinguishes discussions of those objects.

Understanding the Data Set

- *Data set:* 25,355 Wikipedia articles on films, 10,523 Wikipedia articles on novels, and 2,324 Wikipedia articles on TV shows, including the text and metadata for every edit made to each article

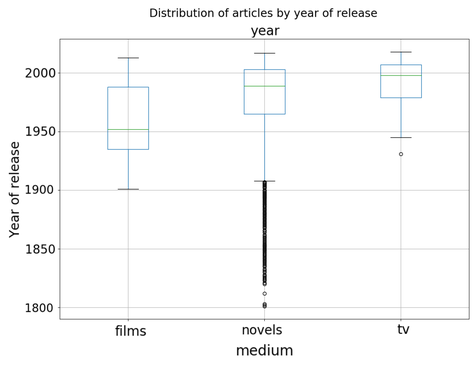

Figure 1: The median release year for novels was 1989, the median release year for films was 1950, and the median release year for TV shows was 1999.

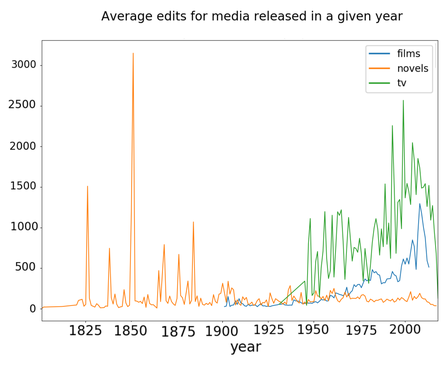

Figure 2: Articles on films had an average of 235 edits, articles on novels had an average of 122 edits, and articles on television shows had an average of 1216 edits.

- *Release year distribution:* Articles on novels and television tend to be skewed towards more recent releases, while articles on films tend to be weighted more evenly across the span of release dates represented on the lists.
- *Quantity of edits:* Articles about more recently released films, television shows, and novels all tend to have fewer edits than articles on media released in the mid-2000s. This suggests editors tend to focus on articles about the recent past rather than the immediate past.

Acknowledgements

This project was funded by a 2019 McGill ARIA award.

Contact Information

- nathan.drezner@mail.mcgill.ca
- www.txtlab.org

Building a Model

- *Operationalizing our goals:* We studied how users tend to cluster around and write about three distinguishing features of the articles: genre, period, and medium. Our question is the extent to which users focus their commentary on single domains or fluidly cross between them. How segmented is editorial behaviour on Wikipedia?

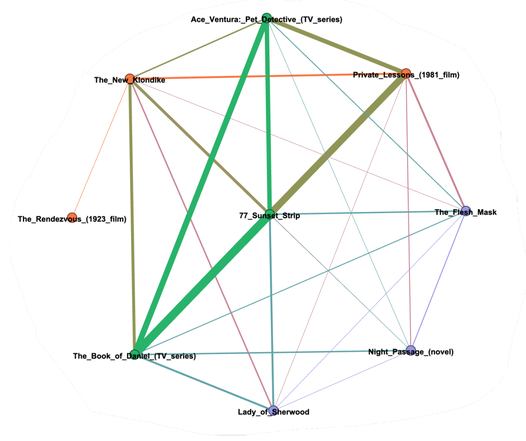

Figure 3: A sample network representing 9 random articles, colored by medium. Edges are weighted by the number of editors who edited both connected nodes. The networks used for analysis each represented 300 random articles.

- *Modeling editing patterns between articles:* We generated a set of 150 social networks representing articles users tend to edit in tandem. We used Louvain community detection to split the networks into sets of nodes and studied the coherence of each set by testing its purity by distinguishing feature (genre, period, or medium).
- *Quantifying semantic difference:* We used logistic regression to compare the language of different articles based on a predictive model. We generated distinguishing features for each class and close-read the most predictive language of each article.

Results

- *User segmentation:* There was a very strong division of user behavior based on medium: users tend to edit only articles about a single medium (92.5% avg. purity). There was a fairly strong tendency for users to edit only either early- or late-period articles on films (78.6% avg. purity) and television (70.8% avg. purity). For novels, where there was very little period-based segmentation (57.9% avg. purity). There was no division of user behavior based on genre: users tended to edit both articles on comedies and dramas rather than specializing by genre. Purity scores were equal to the distribution of articles for both films (57.5% avg. purity) and television (76.4% avg. purity).
- *Semantic segmentation:* The classifier was 95% accurate for identifying articles by medium and 88.7% accurate for genre, indicating very strong semantic differences between articles by medium and genre. The classifier was also very accurate for classifying early- and late-period television (82.0% accurate) and films (86.0% accurate. It was less accurate for novels (76.3%), similar to the pattern of user segmented behavior across period.
- *Distinguishing semantics:* Early-period novels, films, and television shows tend to be identified in writing distinctly by genre (e.g. "silents", "westerners", "melodramatic"), whereas late-period media tends to be written about in more general terms.

Future Work

- *Expanding scope:* We can study and compare other domains Wikipedia—including politics, sports, and science—using these methods.
- *Changes over time:* We could compare writing on Wikipedia by edit time to identify normalization in writing over time on Wikipedia.