Case studies

1. **Inside the Creativity Boom**

(Researchers: Samuel Franklin and Peter Organisciak)

This project investigated the usage of the words “creative” and “creativity” in twentieth-century text. It is understood that the words “creative” and “creativity” found increased usage in the English lexicon in the twentieth century, with creativity rising sharply in use after World War II. This project set out to understand more about this “creativity boom”. In what contexts were the words used over time? And what about their meanings can be deduced from how they were used?

The project began by creating a so-called *creativity corpus*. The corpus consisted of data from volumes in HathiTrust containing the search term creativ*. To address duplicates in the HathiTrust collection, the project team de-duplicated identical volumes or editions, but retained different editions of the same work, totally 2.7 million volumes. They wanted to account for influence and prominence in cases where a work had been re-published over time, but not where identical versions of one title had been scanned by multiple contributors to HathiTrust. For each of the remaining volumes, the research team downloaded the corresponding HTRC Extracted Features file. The HTRC’s Extracted Features dataset includes words and word counts on a page-level basis for each volume in HathiTrust, along with other data and metadata for each volume. In the Extracted Features files, the words and counts are aggregated at the page level, and word order is not maintained. Because the project set-out to learn more about discourse around “creative” and “creativity”, the project team narrowed the dataset to only the parts of the Extracted Features files that corresponded to pages where creativ* was located and discarded (i.e. deleted from the dataset) data corresponding to pages where it was not found. They were only interested in knowing what concepts were talked about around creative, so they wanted to retain only text in close proximity in the volume. They further restricted their dataset by removing words based on part of speech, discarding pronouns and conjunctions but keeping the more “meaningful” words such as nouns and adjectives.

The corpus was analyzed using topic modeling techniques that allowed the team to model the topical themes prevalent around creativ* in the corpus. Topic modeling is a method of using statistical models for discovering the abstract "topics" that occur in a collection of textual documents. (In this project, page-level words and word counts would represent a document.) The computer treats every document as a bag of words. It looks at every word in every document, and then using machine learning, makes a prediction about the words that are likely to co-occur in a document. These co-occurrence predictions are the topics that are modeled. The logic in place in topic modeling is that words that frequently co-occur are likely about the same thing.

In one common topic modeling technique called Latent Dirichlet Allocation (LDA), the order in which texts are sampled is either chronological (starting from the beginning of the list) or randomized to control for time. For this project, however, the team wanted to use time as a factor in order to identify topics specific to their particular eras. They also did not want older topics to be drowned out by the massive number of works published in the latter-years of the period of study. The team developed a topic modeling workflow that temporally weighted the training sample (chronologically, by decade, randomizing within each decade) in order to soften the temporal bias without entirely removing it.

The project resulted in a set of topics that could be graphed to show their prevalence in the creativity corpus over time. Topics that decreased in prevalence from 1940 to 2000 include: 1) god, christ, jesus, creation, word; 2) species, animals, natural, plants, soil; 3) nature, mind, creative, world, human; and 4) invention, power, creative, own, ideas. Topics that increased in prevalence from 1940 to 2000 include: 1) advertising, media, marketing, sales, television; 2) economic, development, capital, economy, production; 3) poetry, language, poet, poets, poems; 4) social, creative, study, development, behavior. From these results, the project team argued that the usage of "creative" changed over the course of the twentieth century become less related to words like "generative" or "productive" and more related to art or imagination.

Link to full paper/report:
https://wiki.htrc.illinois.edu/display/COM/Advanced+Collaborative+Support+%28ACS%29+Awards?preview=/31588360/48595100/Franklin_ACS_End-Report.pdf

Further reading:

- Peter Organisciak and Samuel Franklin, Modeling Creativity: Tracking Long-term Lexical Change," *Digital Humanities 2017*, https://dh2017.adho.org/abstracts/563/563.pdf

2. **How Capitalism Changed American Literature**

(Researcher: Dan Sinykin)

This article describes a research project to investigate the impacts of the consolidation of the publishing industry on American literature. Starting in the 1960s, American publishing became increasingly concentrated in a few major publishing houses. Concurrently, small, independent presses proliferated and began to claim themselves as more literary than the major publishers. This researcher asks whether the rise of publishing conglomerates led to differences in what was written, or how it was written. Their research looked at novels published in the United States by the large publishing houses and a subset of independent presses from 1980 through 2007. It explores whether there are distinguishing features recognizable by a machine between these two types of novels that a human reader would find difficult to identify, either because they are at the level of word patterns or because their reading would be colored by preexisting ideas about the various publishers and the novels.

The researcher's dataset consisted of novels gathered from HathiTrust identified to have been published by a list of publishers within the researcher's time period of interest. They worked with their data in an HTRC Data Capsule, which is a virtual machine environment for analyzing text data from HathiTrust.

This project utilized machine learning methods for classification. The researcher built a computer model and tested whether it could differentiate between novels published by the conglomerate publishers and those published by nonprofits. The features they fed into the model were primarily diction and syntax, in other words, the words and parts of speech that were used and how often they were used. Syntax was tracked using parts-of-speech bigrams to determine patterns for how parts-of-speech followed one another in the text. The researcher found that their model was able to accurately predict whether a publisher was a conglomerate or a nonprofit 70% of the time.

They identified the diction that is most likely to distinguish the conglomerate and nonprofit fiction. For example, the author argues that the distinguishing words in nonprofit novels relate to embodiment, such as body parts, bodily actions, and perception. They also argue that the words that differentiate conglomerate novels grouped into three categories which they call law and power, bureaucracy, and dispositions. The researcher points to previous research that identifies one primary characteristic of fiction as its engagement with perception to claim that their results suggest that nonprofit fiction, with its emphasis on embodiment, is more literary than conglomerate fiction.

Link to full paper: https://www.publicbooks.org/how-capitalism-changed-american-literature/

Further reading:
Dan N. Sinykin, "The Conglomerate Era: Publishing, Authorship, and Literary Form, 1965–2007," *Contemporary Lit.* Winter 2017 vol. 58 no. 4 462-491, doi:10.3368/cl.58.4.462.

3. **The Transformation of Gender in English-Language Fiction**

(Researchers: Ted Underwood, David Bamman, and Sabrina Lee)

In this paper, the authors explore gender in relation to English-language literature, including examining the depictions of male and female characters over time in English-language fiction and the rates of authorship by male and female writers.

Their dataset was composed of 104,000 works of fiction published between 1703 and 2009 drawn mostly from HathiTrust, though the majority of the works, and the focus of their paper, is on the volumes published 1780-2007. In order to build their fiction corpus, the research team used an algorithmic, machine learning process to distinguish fiction works in HathiTrust from works of non-fiction. The authors worked with the text data for these volumes in an HTRC Data Capsule, which is a virtual machine environment for analyzing text data from HathiTrust.

They conducted their study using an analysis pipeline called BookNLP, where NLP stands for Natural Language Processing. Natural Language Processing encompasses wide ranging methods for parsing, understanding, and generating human-language text. BookNLP brings together NLP tasks from the Stanford NLP software and MaltParser, along with other functionality, and optimizes their use for book-length documents. In the basic BookNLP pipeline, first, the text is tokenized, which means it is split into words that the machine will be able to group, for example, recognizing that “wizard” and “wizard” are the same. Then, sentence boundaries, or the beginnings and endings of sentences, are identified—which is harder than it may seem! Next, the tokens are tagged with part of speech using standard codes from the Penn Tree Bank. After that, it does dependency parsing to identify which words are associated with other words based on grammar (i.e. the subject and its verb) rather than proximity. Then it uses entity recognition to identify which tokens are names of people. And finally, it performs entity mapping to group the names and pronouns, and their associated tokens, that refer to the same character.

This project used BookNLP to identify the names of characters, cluster them (e.g. linking “Harry” and “Harry Potter” as one character), and assign each character a gender. It then identified words associated with each character: what they do, what is done to them, how they are described, and the things associated with them. The gender assignations used by BookNLP for this research are male, female, or other/unknown, which the project team argues are sufficient for their study of conventional roles in the period of study.

The paper describes two main results from this project. First, the researchers found that the rate of female authorship declined by half in novels from 1850 to 1950, at the same time that depictions of women in fiction also declined. The researchers were surprised by these results and compared them both to a similar corpus of novels from the University of Chicago Text Lab and yearly bestsellers lists from *Publisher’s Weekly*. They found similar results using the comparison data. While they recognize that their results may be heavily influenced by the source of their data—academic research libraries—they argue that the results align with existing socio-historical observations about the publishing industry, the rising status of novelists, and the sharp rise in female authorship overall. Using the genders identified by BookNLP, they also found that concurrently, the rate of female characters in fiction declined. They claim that male authors disproportionately write about male characters, whereas female authors have tended to write about both male and female characters.

The researchers also were able to use their results to make an argument that depictions of male and female characters have become less distinct over time. First, they showed a (computer) model data about a subset of characters, their identified gender, and the adjectives BookNLP associated with them. With that input data, using machine-learning techniques for classification, the model learned what words were associated with male and female characters. Then, the researchers showed the model characters and their associated adjectives without the gender identification, and asked the model to predict the gender. The researchers supposed that if the model performed well across time, based on publication date, then that would show consistency in male and female depictions. But if it became less accurate over time, then that would show divergence. What they found was that it became harder for the model to differentiate male and female-identified characters the later through the twentieth century.

Link to full paper: http://culturalanalytics.org/2018/02/the-transformation-of-gender-in-english-language-fiction/

Interactive visualization: http://ec2-35-165-215-214.us-west-2.compute.amazonaws.com/dataviz/genderviz

Further reading:

- Andrew Piper, “How can we understand characters using data?,”.*txtlab* (blog), January 22, 2019, https://txtlab.org/2019/01/how-can-we-understand-characters-using-data/.
- David Bamman, Ted Underwood and Noah Smith, "A Bayesian Mixed Effects Model of Literary Character," ACL 2014, http://acl2014.org/acl2014/P14-1/pdf/P14-1035.pdf.
- BookNLP GitHub: https://github.com/dbamman/book-nlp