Eleanor Funkhouser

Looking for patterns in the shapes of stories

Project Overview

I used Project Gutenberg (well, mostly mirror sites) to pull 8 novels written during 1850-1860. Four of the novels were written by white Americans and four by African Americans. Inspired by "Exploring the shapes of stories using Python and sentiment APIs" on the Indico blog, I wanted to examine story 'shapes' (trials and tribulations for the main character(s)) to see if black or white storytellers tended to use common plot arcs. I examined these data using the Pattern sentiment analysis API on the book texts.

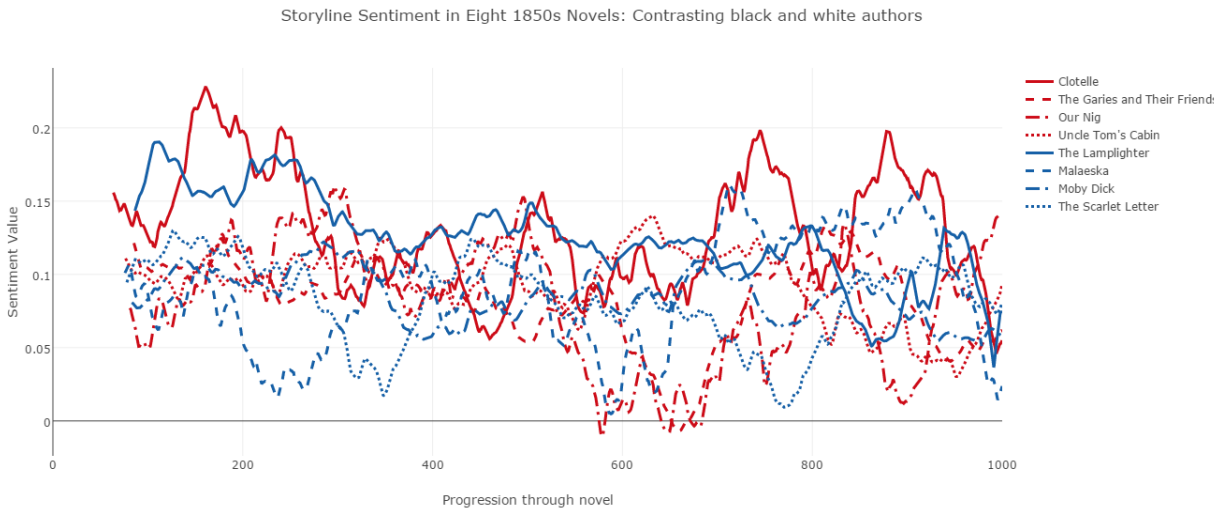
Novels: *Clotelle* by William Wells Brown, *The Garies and Their Friends* by Frank J. Webb, *Our Nig* by Harriet E. Wilson, *Uncle Tom's Cabin* by Harriet Beecher Stowe, *The Lamplighter* by Maria Susanna Cummins, *Malaeska* by Ann S. Stephens, *Moby Dick* by Herman Melville, and *The Scarlet Letter* by Nathaniel Hawthorne.

Implementation

My implementation divides into roughly four parts: data pulling, trimming, piecewise analysis, and plotting. I pulled book text files from the Gutenberg mirrors using urllib and wrote them to files in my documents with pickle. Unfortunately, Gutenberg includes lengthy documentation at the beginning and end of each book, so I wrote functions that would trim off any text before 'Chapter 1 (or I)' and after the end of the novel (by finding the word 'Gutenberg' and cutting off after that). Once the novel was trimmed down to only its content, the more complicated work of doing sentiment analysis came in to play. I chose to do this by first splitting the novel into a list of sentences, then writing a sliding-window function (again inspired by the work in the Indico blog post) to measure the sentiment value of multiple sentences at a time, then move forward one sentence and do it again and so on – so that there was overlap and a gradual, reasonably cohesive change in sentiment to hopefully avoid discontinuities. In plotting, I found that the sentiment data still had a fair amount of noise, so I imported Pandas and plotted the rolling-mean of sentiment to smooth out the noise. I also scaled the x-range of each dataset to be the same in order to look at all of the books' plot arcs on the same domain. With data finally ready, I used Plotly to graph all 8 books.

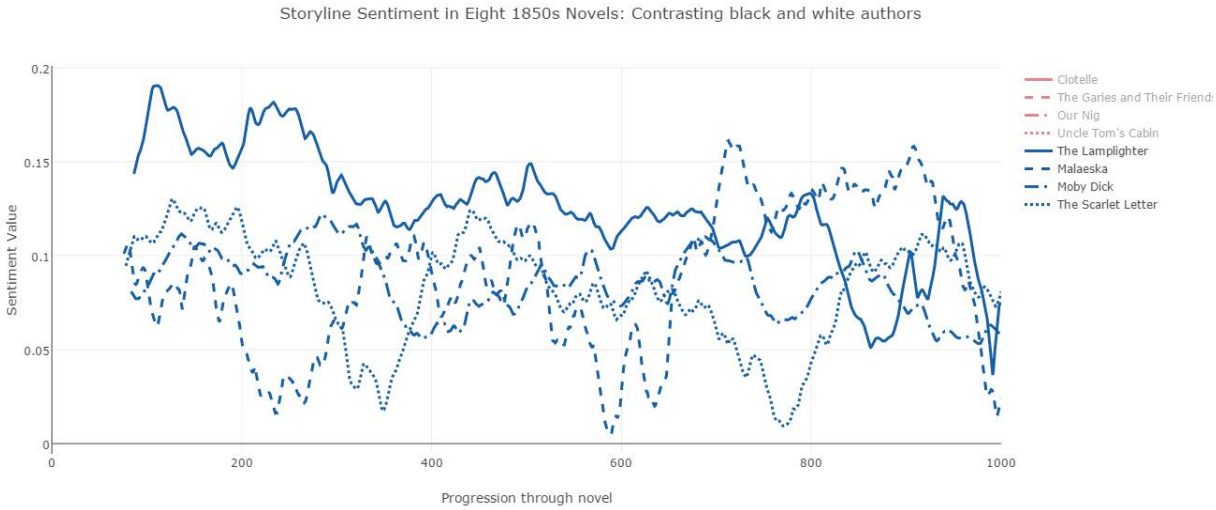
The use and parameter-determination of a sliding window in measuring sentiment were really challenging for me. A small window meant that data points would have huge jumps in sentiment every step, which meant ridiculously noisy data that was difficult to find pattern in. Because I don't have a lot of experience with complicated filtering methods, I didn't feel up to using this data and so I increased the window size to 8 sentences. With every sentiment measurement, the program would step forward one sentence. I think that the high ratio of window size to step, while it smoothed my data nicely, also muted it to some degree – as is visible in the figure below, the sentiment scores of these 8-sentence windows stay mostly between 0 and 0.2 (out of a possible -1 to +1 range). Perhaps this is an inherent flaw in the plan of using a sentiment analysis tool (intended for 'product advertisements') to capture literary complexity, or perhaps I should have taken steps larger than 1 sentence. I made my choices and now I must live with them for the rest of my life. If I did this project again, I would probably fiddle around with those parameters more and do more validation by performing analysis on a book I know really well (as opposed to comparing the curve I got to Wikipedia's summary of *Moby Dick*).

Results

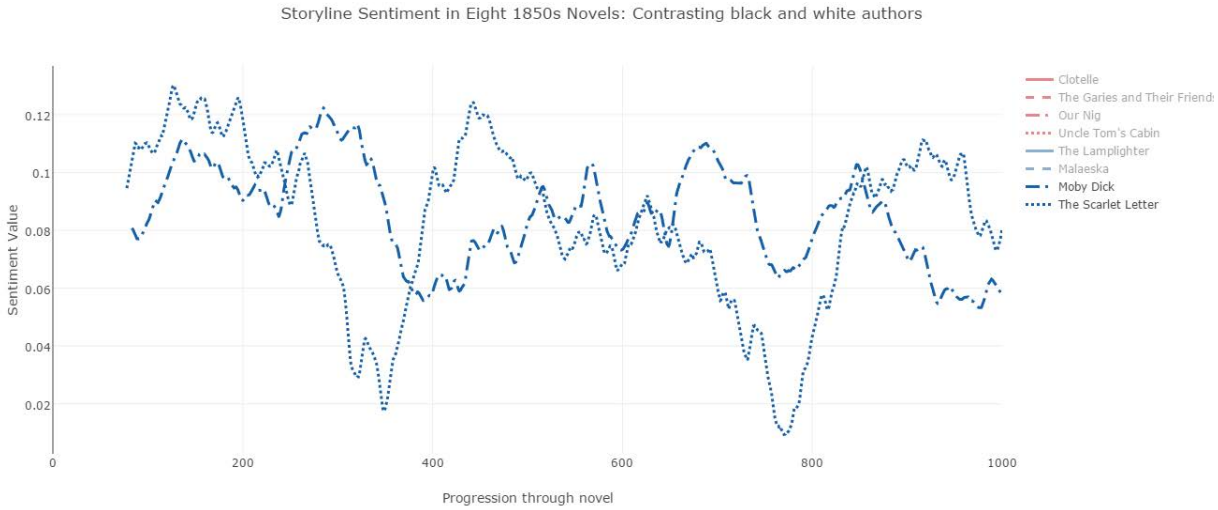


What a mess! This is all 8 books, plotted together. While there may be parts of stories that follow similar patterns, overall it is fairly safe to conclude that there is no neat way to tie these results up into a "Wow, they all ____!" format. Which is great – it indicates that even in an era when novels were just barely starting to become popular (including the invention of the 'dime novel'), it is impressive that they vary so widely in plotline.

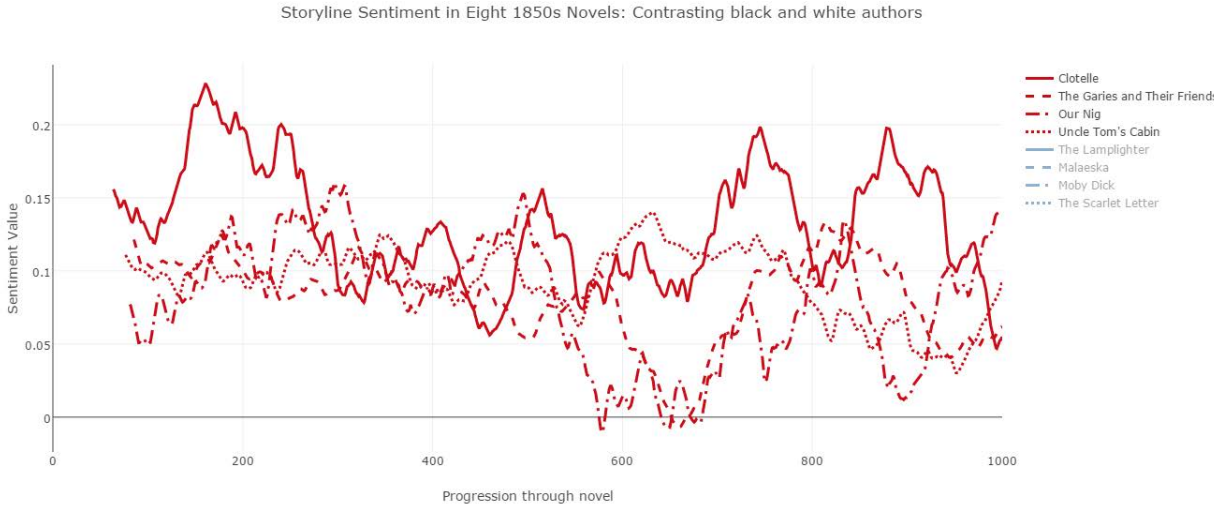
Let's break it down:



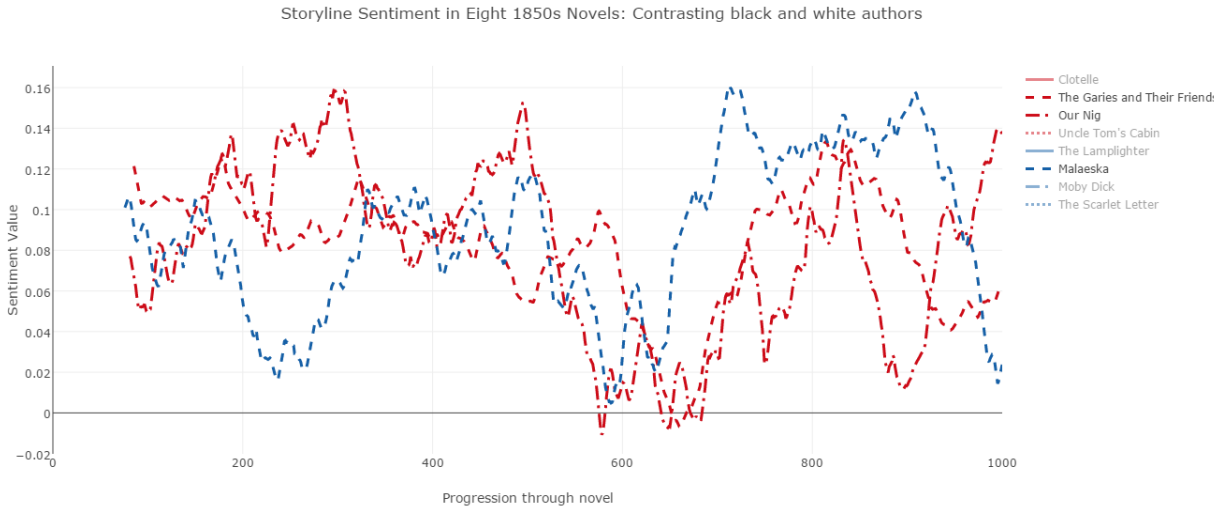
There isn't much consistency or 'unified story' among the white writers. All storylines seem to have a dip by 2/5ths into the story, and all seem to become markedly more negative at the end (including the 'pop novels', Malaeska and The Lamplighter, which one might expect to provide an escapist pick-me-up).



Looking at just Moby Dick and The Scarlet Letter, there are some similarities: a dip early on, some improvement in life, a dip around 4/5ths into the book, some more improvement, then a downer ending. Ahab is dragged away by the whale and Hester's lover dies. Thanks, literary classics!



The black writers also demonstrate a huge amount of variation in plot styles, with no 'typical story arc'. It is notable that they seem to stay fairly positive until some calamity in the middle of the book, and all except *Clotelle* have happy (or at least things-are-getting-better) endings. *The Garies and Their Friends* and *Our Nig* have particularly comparable plotlines, and bear some similarity to *Malaeska*, demonstrating the calamity-in-the-middle pattern I pointed out:



Reflection

As cool as it would have been to stumble upon story arcs that hewed to cultural traditions, it is even cooler that there were few real patterns in this data, and that there is no reason to assume someone will tell a story a certain way based on their culture. Each author brings a distinct style to the table.
Of course, this is still a small sample, not controlled for factors like the author's geography, gender, or enslavement status/past. There is also the possibility that every story had exactly the same 'shape' and I grossly misjudged what would be reasonable sampling parameters to measure the sentiment. Or that my sampling algorithm was fine and the fault was in using a sentiment analysis API not remotely designed for this subtlety of language (not to mention the nuances of capturing plot – this was really just a measure of whether the author was using positive words or negative words at any given point).
Overall, there were a lot of failure modes here; probably the biggest and most preventable one is that I did not use the final plotting algorithm on a novel-length story whose arc I knew well, and so didn't do the work to validate the model. The project was appropriately scoped but I did not scale my time commitment to account for how I would struggle with the new plotting tools or how I would need to fiddle with design parameters in parallel with running a validation case. I think these things are pretty easily fixable moving forward, and overall I'm not dissatisfied with the work done here considering it was my first experience with text analysis and I had a week to do it. Hurray, mini-projects!