Coleman Ellis
25 February 2016
SoftDes MiniProject #3

Sentiment Analysis of Jane Austen's *Pride and Prejudice*

Overview

Using Pattern's sentiment analysis function, I plotted the average sentiment polarity of each sentence that mentions "Mr. Darcy" over all 61 chapters of Jane Austen's *Pride and Prejudice*, hypothesizing that the average sentiment towards Darcy would start fairly low and increase over the course of the book.

Implementation

My project has three major parts:

1. Retrieve the text
2. Clean the text
3. Analyze the text.

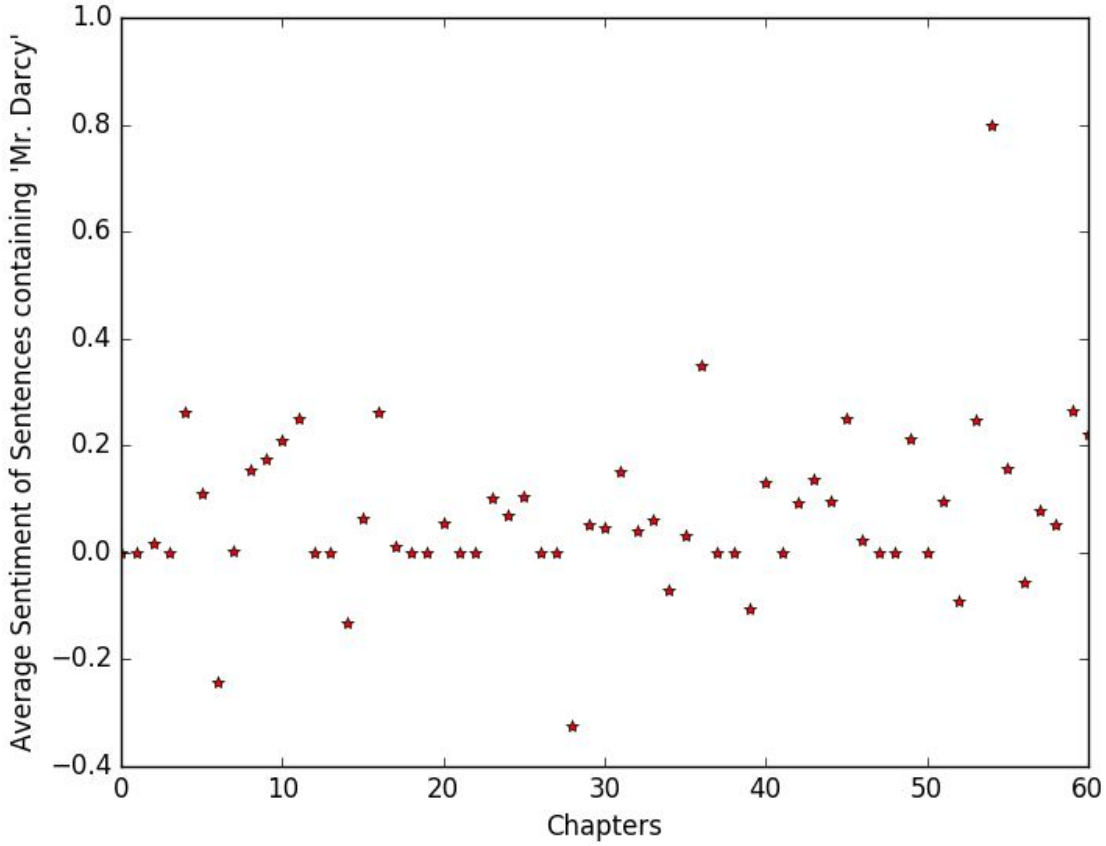
Retrieval of text was fairly straightforward, simply use pattern to download the text file of *Pride and Prejudice* from Project Gutenberg (or, more often, a mirror thereof) and save it.

Cleaning the text was more involved: my first pass at cleaning the text involved breaking the whole book down into a list of chapters, where each chapter was a list of case-corrected words without punctuation. Later on, I decided that a list of lists of words was a little too processed; a list of lists of sentences would be more useful for analyzing sentiment, as it would allow me to check which full sentences mention Darcy. I thus changed to using the Python Natural Language Toolkit's tokenizing function to split the chapters into lists of sentences. This nested list was pickled and saved for easy future access.

Analyzing the text was again fairly simple. For each chapter, I did two sentiment analyses: the average sentiment of each sentence containing a mention of "Mr. Darcy", and the average sentiment of each sentence. These were stored as two separate lists, each containing the average sentiments for each chapter. These lists were not pickled-- I spent a lot of time tweaking the parameters of analysis and experimenting with sentiment analysis, and saving a file just for it to get saved over again before being accessed seemed like a waste. Plus, the full sentiment analysis of the text only took a couple of seconds at most, so pickling the lists wouldn't've saved a significant amount of time.

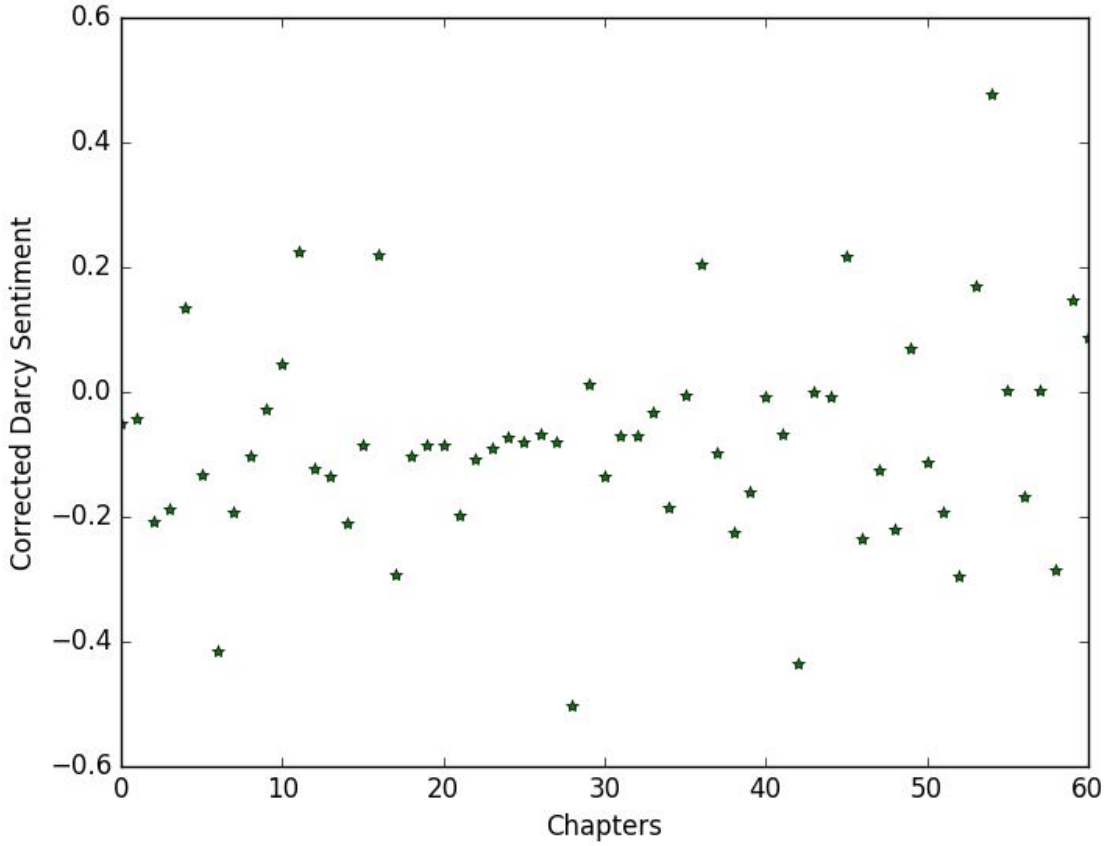
Results

My results, unfortunately, were not particularly compelling. This is the graph of the average sentiment of sentences containing “Mr. Darcy”:



As you can see, there’s not much correlation between book progress and sentiment toward Darcy-- the graph maintains a fairly constant average sentiment of approximately 0.1. There are a few spikes of positive or negative sentiment, but beyond the fact that the largest positive spike is near the end of the book they don’t appear particularly meaningful. This is doubly true with the knowledge that the large positive spike at the end was caused by this sentence: “Not a word passed between the sisters concerning Bingley; but Elizabeth went to bed in the happy belief that all must speedily be concluded, unless Mr. Darcy returned within the stated time,” a sentence which is barely relevant to Mr. Darcy (and, in fact, implies that Mr. Darcy could ruin Elizabeth’s happy belief), which Pattern gave a sentiment of 0.8 (and is, in fact, the only sentence in that chapter that mentions Darcy).

My next step was to compare the average sentiment toward Mr. Darcy to the overall average sentiment of the chapter and see if that made a difference. Here's that graph:



It is, if anything, worse. All that it shows, at best, is that the average sentiment towards Mr. Darcy is slightly lower than the average sentiment of the book as a whole.

These were not the only two tests I ran; since my function that found the sentiment of sentences containing "Mr. Darcy" could test any word or name, I tested every variation of "Darcy" and "William" (his first name), "Mr. Wickham" (another character, but who starts off in high esteem and is eventually despised), and several other characters, as well as only sentences containing dialogue about those characters. For the average sentiment, I tried only grabbing the sentiment sentences that contained common words, averaging sentiment over several chapters, and only sentences containing dialogue. None of these options resulted in compelling graphs, and I've opted not to include any of them as I feel the two above adequately summarize the fairly meaningless nature of my results.

Reflection

Process-wise, I feel my project went quite well: it was appropriately scoped, I wrote effective code and documented it while writing it (which is not something I usually do), and other than a few written functions that ended up not being used, most of the coding I did and time I spent on the project resulted in either gained experience or useful results. It was nice to learn

how to open and manipulate files on a fairly high level, and to learn about pickle (which is something I'd never even thought existed before this project started). I do wish, however, that I'd known what the outcome of this project would be before I'd started; while I don't regret spending the time I did pursuing this project, I do wish that I'd pursued a project that had a more interesting and compelling end product. If I hadn't been swamped with other homework as well this week I would have changed gears and investigated some other topic, but there's not much I can do about that. In any case, I'm happy with what I accomplished.