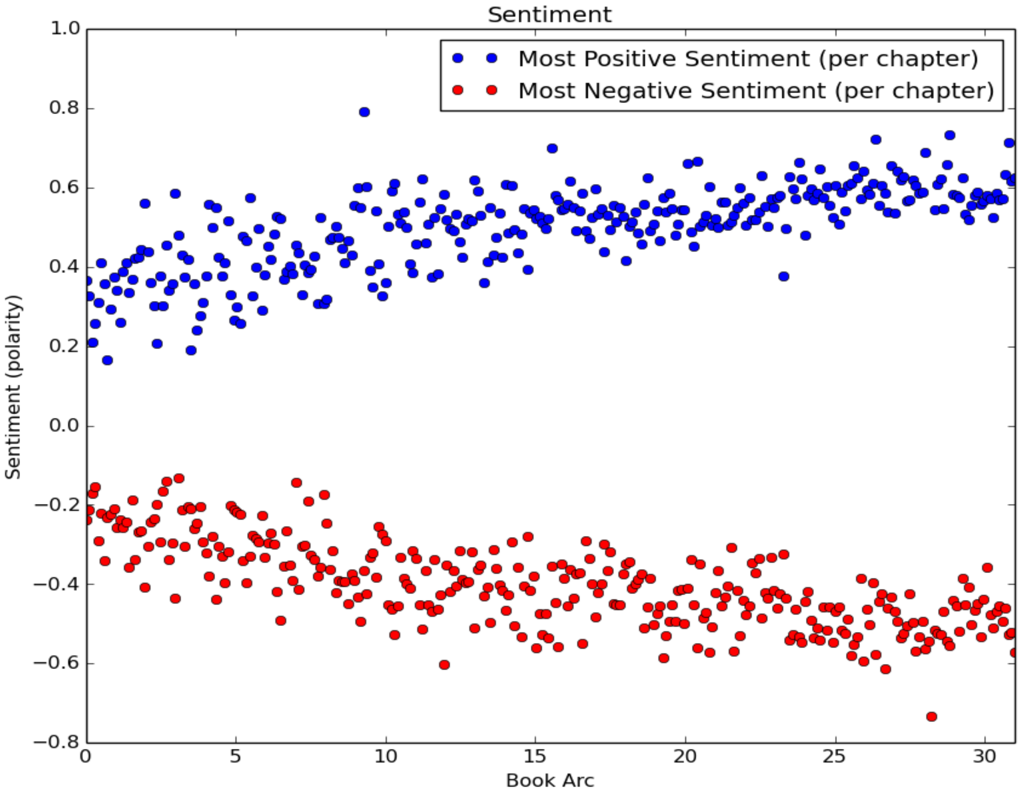
Web Scraping and Sentiment Analysis of Worm
by Sean Carter

For my project, I decided to use pattern.en to do a sentiment analysis of one of my favorite stories, a 1.75 million word web serial posted over the course of two and a half years, after scraping it from the web. It's a complicated story, with long character arcs and dramatic change in the way the main character approaches different situations. Since it went mostly unedited, the writing style evolved as well, meaning that there was a very good chance of some sort of gradual change in the way that the text is written, either from the writing or the story.

To get the text, I downloaded the Table of Contents page, and looked for all of the URLs formated like a chapter's URL, so that I knew what pages to call for the rest of the story. Once I did that, I downloaded each chapter into a separate text file, cut out all of the text that wasn't the story (mostly reader comments, but also some HTML that didn't get parsed right.) I chose to save to separate files so that if I wanted to, I could easily experiment with a single chapter, and not mess anything up. It also made it easier to re-download only some of the documents, when a formating issue resulted in incorrectly parsed files. Finally, they were text files, so that I could look at them to find what form of standardized formatting the chapters used, as well as telling whether I messed them up.

Then, I used the pattern sentiment analysis on each sentence of each chapter individually, and averaged the top 30 most positive, and top 30 most negative sentiments, to get a feel for how emotionally powerful each chapter was. Finally, I graphed both the positive and negative data point for each chapter in the first figure. Yes, each point is actually a whole chapter: it's REALLY long story. I went with an average, because returning just the most extreme sentence wouldn't really give an emotional overview of the whole chapter. And the reason that I graph the extremes is that even in an emotionally charged situation, a majority of the sentences are descriptions of places or people, without a lot of real emotional weight, and they drag the average towards zero. The scatter-plot is so that someone reading the graph won't be biased to see things in the data by the presence of a line already drawn there.

As it turns out, the results were fairly interesting. The earlier chapters averaged a much smaller emotional extreme, both in the positive an negative direction, as you can see in the graph.



As the story moves on, both the positive *and* negative sentiment extremes tend to become more so. This suggests that the writing and/or story become more emotionally charged as they move on. I was a little bit surprised by this – I expected a flatter plateau of positive sentiment, followed by very little later in the book. I also thought that the positive and negative sentiments would simply be the same function, offset from each other. It does make sense, though. Reading the book, the main character grew into a more empowered person, but faced progressively more awful situations. I think that the strongly positive moments in the beginning that I remembered might be those outlying dots that are well above the average, as opposed to the general trend.

Finally, I do think this was a reasonably scoped project. I'm still getting used to the idea of finding new python libraries, and looking through their documentation to get the one or two functions I need to complete a project. Scraping the web for this book was a monumental task, and briefly spiraled down into the deepest pits of unicode hell: the formating for URLS was entirely human-maintained, and changes several times on the table of contents page. Especially dealing with the '½' character. The symbols indicating where the text begins and ends were equally bad. My analysis was more modest than it might have been, but I learned a lot, and still got it all done. If I wanted to do more in the future, I think I might track sentiment related to specific characters throughout the book, or compare this story to the second one that the author is currently writing.