Text Mining Mini Project

Linnea Laux

February 24 2015

1 Project Overview

I used the text of the anonymous Therapy mailing as my data source for this project. I downloaded the text from the list archives for the first 12 months that Therapy was used and for the past year, and analyzed the ways in which Olin's use of Therapy has changed.

2 Implementation

My first, and most time-intensive, task was to clean the raw text. In the text file from the archives, each email contained a header with the date, the subject line and other information. I needed to get rid of all of this nonsense and get to a string of words with spaces in between them. First, I used s.translate with a dictionary of None to get rid of all the formatting characters that I knew I didn't want. Next, I used s.replace to replace all the common header words and a few characters (/,-,.,:) with spaces. At the end of that work, I used s.replace again to make any long strings of spaces into one space. Once I was happy with the quality of what was printing, I split the string on spaces to create a long list of words. I used these methods on both the old Therapy emails and the new ones. I used the histogram() and mostfrequent() functions from reading journal seven, as well as sentiment from pattern to analyze my lists of words. I created a list called oldwords with the most common 200 words in oldtherapy and did the same with newwords and newtherapy. For both of these lists, I started at the first element that was a noun. It was a bit arbitrary, but it seemed like the quickest way to ignore the most common "little words". I then created two

more lists called uniquenew and uniqueold and used two for loops to fill them with the words from newwords that did not appear in oldwords and conversely. I found that these lists of unique words themselves were pretty interesting pieces of data, but I decided to do sentiment analysis on them. I used pattern's sentiment() function to do sentiment analyses on oldwords, newwords, uniqueold, and uniquenew.

3 Results

I don't have any graphs, but I will provide the lists of words and the sentiment analyses, and my analysis of both. The sentiment and polarity for the first 200 words of each list were similar enough that I don't feel comfortable drawing conclusions from them. They were both slightly more positive than neutral and halfway between subjective and objective. The lists of unique words are much more interesting. It's pretty clear just from looking at them that people talked about different things during these 12 month time frames. There is no swearing, for example, in the first one, and lots of words that suggest productive high-level discussion such as "anonymity" and "Honor" "Board". In the second one, there seems to be a lot of discussion that is negative and Olin-focused. UOCD and scope were mentioned, as were "design" and "professors". The sentiment analyses for the unique word lists support my initial assumptions in some ways, but not in others. I expected the polarity of uniqueold to be more positive than uniquenew, but they were both very close to zero and uniqueold was actually more negative. The subjectivity scores were much closer to what I expected. The uniqueold score was halfway between neutral and objective, and the uniquenew score was on the more subjective side. It's pretty easy to see that the words in uniqueold are less subjective than those in uniquenew, so that makes sense.

```
Sentiment for first 200 old therapy:(0.11453136810279667, 0.4089096749811034)
Sentiment for first 200 new therapy:(0.12992857142857145, 0.550484126984127)
['Honor', 'list', 'Board', 'next', 'case', 'type', 'Skipped', 'HB', 'anonymity',
 'Tuesday', 'broken', 'anonymous', 'i', 'honor', 'A', 'BOARD', 'HONOR', 'may', '
board', 'he', 'anyone', 'members', 'find', 'No', 'little', 'Code', 'last', 'off'
, 'down', 'such', 'Hall', 'theyre', 'Ball', 'Snow', 'though', 'might', 'never',
'come', 'policy', 'lists', 'back', 'two', 'his', 'Man', 'saying', 'her', 'THE']
Sentiment for unique old therapy:(-0.10615079365079365, 0.2507936507936508)
['about', 'not', 'are', 'or', 'shirt', 'fuck', 'scope', 'seniors', 'olin', 'frie
nds', 'classes', 'many', 'shit', 'Also', 'campus', 'college', 'use', 'money', 'T
hat', 'doing', 'Audit', 'Why', 'design', 'information', 'sure', 'situation', 'di
fferent', 'theres', 'reason', 'makes', 'bad', 'got', 'arent', 'idea', 'day', 'an
other', 'professors', 'non', 'having', 'UOCD', 'stuff', 'great', 'few', 'else',
'same', 'hate', 'own']
Sentiment for unique new therapy:(0.009090909090909104, 0.6300505050505051)
```

Figure 1: The output of my program in the terminal.

4 Reflection

With more time and effort, I could have done a lot more with this data. I suppose that means my project wasn't appropriately scoped, but I also don't know how the scope could possibly have been smaller. I suspect that my life isn't appropriately scoped. I don't really feel like I can reflect appropriately on this project because I did't feel like I could fully "do" the project. This is no one's fault (a person could argue that it was mine, but that person probably doesn't know my life), and I think having a partner for the next project will help me prioritize SoftDes and get more out of the project.