Amherst History Report

Andrea Boskovic, Oliver Baldwin Edwards

October 4, 2020

Contents

Executive Summary (for Biddy)

By using text analysis techniques on *History of Amherst College during its first half century, 1821-1871* by William Seymour Tyler, we can draw conclusions about Amherst's history during this time period. Based on sentiment analysis, which summarizes positivity, negativity, and other emotions conveyed in a body of text, Tyler's perspective on Amherst seemed to be relatively neutral. Sentiment analysis techniques categorize words into sentiments according to sentiment dictionaries. We also investigated other emotions, not exclusively positive or negative, and the distribution of words in different categories varies largely between chapters. In the chapter about the Civil War, for example, most words are associated with fear, anticipation and anger, but in the Preface, the most frequent words are relatively evenly distributed across the six sentiment categories: fear, anger, anticipation, joy, surprise, and trust. Below is the word cloud with sentiments for the chapter about the Civil War.

One particularly interesting finding comes from Chapter 5, which discusses efforts to align Amherst and Williams. This chapter contains more negative words than positive words overall. The most common words in the chapter include "controversy," "difficult," and "doubt," which could represent the tension between the colleges that is now manifested through their rivalry.

We also investigate the differences in word count and sentiment for the duration of tenure for two Amherst presidents during the time period: Dr. Hitchcock (chapter seventeen) and Dr. Stearns (chapter twenty). Dr. Stearns is referred to largely in a positive light, with the most common positive words being "modern" and "success," which appear more often than any negative word. The most common negative words about Dr. Stearns are "lost" and "scarcely." These words, however, do not capture the context of the chapter, which is a critical limitation of this kind of analysis. In particular, "lost" in this chapter almost always refers to the North building, which burned down at the time. The word "scarcely," on the other hand, does not refer to the president in most cases and is used in conjunction with other words.

Chapter seventeen, which discusses the presidency of Dr. Hitchcock, contains more negative words. As with Dr. Stearns, though, most of these are taken out of context. "Debt," the most common word, has a negative connotation, yet in the context of the chapter, we see that Dr. Hitchcock actually relieved the college of its outstanding debts, meaning he had an overall positive impact. The next most common words are "doubtless" and "faith," capturing the president's attitude towards issues in the college.

When we try to split chapters into two topics, we don't see much difference in the words represented. That is to say that Tyler is usually consistent in what he discusses in each chapter. The only exception to that is in some chapters, where one can discern a financial component in the main discussion of the chapter. In these cases, we see that one topic contains words such as "trustees," "fund," and "dollars," likely indicating that the chapter in part discussed the college's finances.

Although this project certainly conveyed the limitations of text analysis, specifically with sentiment analysis, our findings with these techniques also illuminated their value. There wasn't much difference in topics discussed within any chapter, but this simply tells us that the author's discussion in each chapter was focused on the thesis of the chapter and didn't significantly deviate from it. Likewise, sentiment analysis showed us the general tone of each chapter, but we needed to put those findings into context because they could be misleading.

Wrangling

Process Chapters

A Function to Remove Hyphens

The main part of our text processing involved handling the hyphens at the end of lines in each chapter. If we ignore words at the beginning and end of lines with hyphens, then we can lose valuable information, especially when many lines in the text end with hyphens. Below, we show the function to remove the hyphens from the ends of lines and then combine the truncated words. It is important to note that this function won't work if the chapter passed in ends in a hyphen, but this shouldn't be an issue because each chapter is complete.

```
# A function to remove dashes from the ends of line (and combine the truncated words)
remove_dashes <- function(lines) {

  # Iterate through all lines
  for (line in 1:(length(lines) - 1)) {

    # If the current line we're on is blank then skip it
    if(lines[line] == " " | lines[line] == "") {
      next
    }

    # Otherwise, we need to check for a dash at the end of the line
    # Keep track of the last word in the current line
    last_word <- word(lines[line], start = -3)

    # If the line that  we're on ends in a dash, then we want to remove it
    if(substr(last_word, nchar(last_word), nchar(last_word)) == "-") {

      # If the next line is blank, skip it and check next line for the first word
      if(lines[line + 1] == " " | lines[line + 1] == "") {
        first_word_next_line <- word(lines[line + 2], start = 1)
      }

      # Otherwise, the first word is on the next line
      else{
        first_word_next_line <- word(lines[line + 1], start = 1)
      }

      # Keep track of the beginning of the line (not including truncated word at end)
      beginning_of_line <- word(lines[line], end = -4)
      # Remove dash from the last (truncated) word
      word_without_dash <- gsub("-", "", last_word)
      # Update current line so to remove dash and add word from following line
      lines[line] = paste0(beginning_of_line, " ", word_without_dash, first_word_next_line)

      # Remove first part of word from line below - check again that next line is blank
      if(lines[line + 1] == " " | lines[line + 1] == "") {
        # If so, update the line two lines down
        lines[line + 2] <- word(lines[line + 2],
                                start = 2:nchar(lines[line + 2]),
                                end = -1)[1]
      }
```

```
      # Otherwise we just update the line one down
      else {
        lines[line + 1] <- word(lines[line + 1],
                                start = 2:nchar(lines[line + 1]),
                                end = -1)[1]
      }

    } # End if statement that checks for a dash
  } # End for loop

  # Return our updated lines removed dashes
  lines
} # End function
```

Example of Removing Hyphens

Below is an example of removing hyphens from a chapter. We first load chapter 1 and refer to this as `sample_chapter`. We give an example of lines 185-195 of chapter 1. Clearly, there are some lines with hyphens at the end because the entire word couldn't fit on one line. To fix this problem and avoid losing information, we run our `remove_dashes()` function shown above on the selected lines of `sample_chapter`. After running this, we see that the hyphens are indeed removed, the words are combined, and information is preserved.

```
# Read in sample chapter
sample_chapter <- readLines("chapter_files/chapter01.txt")

# Lines 185-195 before removing dashes
sample_chapter[185:195]
```

```
##  [1] "The beauty of New England villages is universally recognized,  "
##  [2] "whether by visitors from other sections, or travelers from foreign  "
##  [3] "lands. Dr. Dwight finds this beauty in its highest perfection in  "
##  [4] "the towns on or near the Connecticut River, and expatiates with  "
##  [5] "much satisfaction on the plan of the villages, as it is there car-  "
##  [6] " "
##  [7] "ried out, and the excellence of the social, intellectual, and moral  "
##  [8] "results as they are there realized. The selection of the site, not  "
##  [9] "like a village or large town in the Middle States, where trade,  "
## [10] "commerce or manufactures demand, but wherever beauty or con-  "
## [11] "venience, pleasure or moral uses may invite the bringing of the  "
```

```
# Remove dashes from lines 185-195
remove_dashes(sample_chapter[185:195] )
```

```
##  [1] "The beauty of New England villages is universally recognized,  "
##  [2] "whether by visitors from other sections, or travelers from foreign  "
##  [3] "lands. Dr. Dwight finds this beauty in its highest perfection in  "
##  [4] "the towns on or near the Connecticut River, and expatiates with  "
##  [5] "much satisfaction on the plan of the villages, as it is there carried"
##  [6] " "
##  [7] "out, and the excellence of the social, intellectual, and moral  "
##  [8] "results as they are there realized. The selection of the site, not  "
##  [9] "like a village or large town in the Middle States, where trade,  "
## [10] "commerce or manufactures demand, but wherever beauty or convenience,"
```

```
## [11] "pleasure or moral uses may invite the bringing of the  "
```

Clean Up All Chapters

We repeat the process shown above for all lines of each chapter. Using the `map()` function, we read in the text datasets for each chapter. We then run the `remove_dashes()` function on each chapter.

```
# Read in chapter files
location <- "chapter_files/"
chapter_names <- list.files(location)

# Read in all chapter.txt files
all_chapters <- map(chapter_names, function(x) readLines(paste0(location, x)))

# Remove dashes from each chapter
for(chapter in 1:length(all_chapters)) {
  all_chapters[[chapter]] <- remove_dashes(all_chapters[[chapter]])
}
```

Analysis

The main limitation of our sentiment analysis is that it does not take the context of the sentence into account. This problem was highlighted in our investigation of Amherst presidents. In comparing chapter seventeen and chapter twenty, which outline the presidencies of Dr. Hitchcock and Dr. Stearns, respectively, we initially thought that Dr. Hitchcock was a much worse president that Dr. Stearn because there were many more negative words associated with his presidency than Dr. Stearns'. In particular, "debt" was the most common word in the Bing lexicon describing Dr. Hitchcock. A reasonable initial assumption of "debt" being the highest frequency word is that Dr. Hitchcock created debt for the college. Given the context of the chapter, however, we realized that Dr. Hitchcock relieved the college of its debts. Moreover, looking across all the lexicons (Bing, NRC, and AFINN), most of the negative words describing Dr. Hitchcock arise for the same reason. Sentiment analysis, namely with the Bing lexicon, doesn't capture Dr. Hitchcock's overall positive impact on the college. Words such as "doubtless" and "faith," some of the other most frequent words in chapter seventeen, on the other hand, may more effectively capture the President's approach towards Amherst's issues.

Although it's crucial to be aware of the drawbacks of sentiment analysis, there are some cases in which it proves useful. In chapter twenty-six, which discusses Amherst in the time of the Civil War, the most common words reflect fear, anticipation, and anger. "Battle," "war," "rebellion," and "sick," all words one would expect to see in a chapter of this nature, are encapsulated in these areas. In general, few strictly positive or negative words appear in the word cloud for NRC over the course of the chapter because most of these words are evocative and thus tied to a more specific emotion.

There was no clear drawback with topic modeling as a technique, but it didn't prove useful in our analysis. Topic modeling uses the Latent-Dirichlet Allocation algorithm to find the mixture of words associated with each topic along with the mixture of topics in each chapter. One reason for its failure in this case could be that Tyler's chapters were relatively narrow in scope and thus didn't have many clearly separable subtopics. In some cases, we could see a separation in discussion of Amherst's financial situation, with one topic containing words such as "trustees," "fund," and "dollars."

Discussion

Sentiment analysis and topic modeling are exciting and informative, but we learned that these techniques don't always work well. Our data was a book on the history of Amherst for its first fifty years, separated by chapter. In this case, topic modeling, for instance, didn't work well because there aren't distinct topics within a chapter in general. This creates overlap between the two topics because it is difficult for the Latent-Dirichlet Allocation algorithm to create two classes when there is so much similarity between words in the chapter. Sentiment analysis, though it did capture some interesting patterns, especially in chapter twenty-six, which described Amherst during the Civil War, wasn't useful in most cases. In general, sentiment analysis is misleading because it takes words out of context. In the future, it might be useful to consider n-grams, which can take n-tuples of words into account, but even this may not be appropriate here. Topic modeling likely works better with larger works with more variety, and sentiment analysis is probably more informative in evocative writing, unlike a history book.

Although we often strive to include as much information as possible into plots and model summaries, it may be easier to interpret a pattern without so much noise. To account for this, we added an option to view a word cloud for the selected chapter at the selected lines without information added from different sentiment analysis lexicons. In order for users to better understand the text and avoid the issues raised by sentiment analysis described previously, the dynamic text output changes so that the user can read parts of the text they selected, giving them a better impression of the sentiment. In the future, it would be interesting to consider something like the `word2vec` algorithm, as this method could be informative with this type of text.

To interact with the data yourself, you can view our Shiny App here: https://r.amherst.edu/apps/obaldwinedwards21/AmherstHistory/.

Technical Appendix

In our analysis, we used a few techniques that went above and beyond what was discussed in chapter 19 of MDSR. Specifically, we checked for multiple edge cases when removing dashes from the provided text and examined topic modeling (https://www.tidytextmining.com/topicmodeling.html).

Our `remove_dashes` function is robust in the sense that it accounts for multiple edge cases. It accounts for situations where there are multiple, subsequent lines of text that end in dashes as well as when a page ends with a dash (meaning that there is a blank line in between both sides of a word cut off with a dash). This meant that we had to write code that wasn't specific to any one task of removing dashes and was instead focused on being scalable and applicable to multiple edge cases.

We also used topic modeling (included as a tab in our Shiny app). Topic modeling is an unsupervised machine learning technique similar to clustering that groups words into a specified number of different topics based on how related the words are to one another (using word groups and phrases to decide groupings). More specifically, topic modeling uses the Latent-Dirichlet Allocation algorithm to find the mixture of words associated with each topic along with the mixture of topics in each chapter. Topic modeling was not mentioned in chapter 19 of MDSR and was an exciting technique to explore (even if it didn't provide much interesting insight into our data).