Amherst History Report - Revisions
Group 2 - Nicole Frontero & Sabir Meah

Format:

- Question (written by Nick or classmates)
    - Response (written by us)

Feedback from Nick:

- I would prefer a function to do the processing that you have within the for loop. But you can keep what you have.
    - Noted for future assignments. We've kept what we have as you said was permissible.
- I couldn't confirm that you glued together things like col- lege → college with your code. Can you please confirm?
    - We did indeed glue together these types of cutoff words, doing so by removing the "- " then merging together lines. Table 1 showing the first 250 characters of each chapter shows these results (none of these cutoff words are present there).
- Table 1 was nice! Perhaps extend to 250 characters (would need to be a long table).
    - We have extended Table 1 to include the first 250 characters of each chapter. We utilized longtable=TRUE to make this happen.
- I might suggest removing "college" and "amherst" after you display table 2 then add a new table 3 that lists the top 10.
    - We decided to display table 2 with the top 10 most frequent words to appear in the book. We think that it is valuable to see that "college" and "amherst" appear so frequently, but we recognize that by only including the top 6 most frequent words and still including "college" and "amherst", we only see 4 additional words. Including the top 10 words allows for readers to see that "college" and "amherst" appeared most frequently, but they can also see 8 other words that appeared most frequently.
    - Table 3 however, showing the top 3 most frequent words in each chapter, was recreated to not include "college" and "amherst".
- I would suggest report top 3 most frequent words per chapter e.g. history (n = 13), XX (n = 8), XX (n = 6), etc.
    - Table 3 now has a row for each chapter that includes the most frequent, second most frequent, and third most frequent words within each chapter. For this analysis, we did not include "amherst" nor "college." Also, we chose to not include the number of times each word appeared because we believed that it

would make the table unnecessarily confusing and would not add anything to the interpretation.

- Score for tf-idf (also suggest not displaying tf, idf)
    - We removed the tf and idf columns that previously appeared in table 4. Now only the chapter number, word, number of occurrences, and tf-idf are present in the table.
- Suggestions in executive summary
    - We addressed all of the comments given here, with the one exception of the suggestion to add a second word cloud of a specific chapter. We didn't think that we should go too much into detail about any chapter-specific things, because the executive summary is a high-level overview.

Feedback from classmates:

- I would clean up the data for chapter 28 during the analysis of the most common word per chapter. Although not a big deal for the big picture, 00 doesn't have any information in itself.
    - We agree that 00 does not add any information. We have decided to remove 00 from the strings in our wrangling section. We left other numbers however because we felt like they were still meaningful.
- I would adjust the analysis of the first 130 characters. Generally, this just grabs the title of every chapter. I think it would have been better as an analysis of chapter titles. The first sentence doesn't add much and looks inconsistent because of the hard cutoff at 130 characters.
    - Addressed in Nick's comment, we expanded to 250 characters which includes more than just the title. Also note that the point of this table was to just display the results of our wrangling, not to perform any analysis (hence why it was in the wrangling and not analysis section).
- I think removing Amherst and college would be a good way to gleam more meaning from the text, especially if the chapter table includes only 1 word per chapter.
    - Addressed in Nick's comment, we removed "amherst" and "college" from Table 3.
- More data wrangling could have been useful, especially since Chapter 28 has 00 in the top spot for tf-idfs. This seems to me like a problem with how a book was scanned and not a result of the author using "00" as a separate part of sentences.
    - Addressed in a previous classmate's comment. We removed any instance of "00" in our data. We left other numbers however because we felt like they were still meaningful.
- Yield titles for each chapter, instead of first 130 characters

    - Addressed in Nick's comment, we expanded to 250 characters which includes more than just the title. Also note that the point of this table was to just display the results of our wrangling, not to perform any analysis (hence why it was in the wrangling and not analysis section).
- Word frequencies do not really give much useful information since words like "Amherst", "College", "President" are expected to be very common in the text to begin with. Could remove those words beforehand to perhaps get more interesting results.
    - Addressed in responses to Nick: we removed "amherst" and "college" from table 3. Also, we included more words in table 2 so that we could still show that "amherst" and "college" were frequent words, but we could still show other frequent words.
- For loops are a frequent occurrence in your wrangling; have you considered a more elegant approach in the Tidyverse style?
    - Addressed in Nick's comment. We admit that there could be a more elegant solution, but Nick has told us it is permissible to keep it as it is.
- Figure 1 (faceted top 5 tf-idfs by chapter) seems needlessly dense; would a Shiny app be a better way of visualizing what you're trying to convey?
    - We admit that a Shiny app may be a more compact way to visualize this data, but in all it shows the same information. The way we currently have it as well also offers the advantage of being able to be displayed in more traditional contexts than a dynamic Shiny app (e.g. in a published paper). Our figure may seem a little overwhelming to readers but they are absolutely free to just skim or only look at certain chapters instead of every single faceted graph.
- Is there a more compact way of conveying the information you're trying to show in Table 3?
    - In addressing Nick's comments, we haven't made this table more compact, but we did make it more information-dense by adding additional columns for the top 3 words per chapter and not just the top 1.
- define tf-idf's in the executive summary
    - Addressed in Nick's executive summary comments - we defined them in our revised submission.
- less words in the word cloud
    - We did consider this, but considering that this visualization (or at least the first word cloud) is about the entire book we wanted it to be as detailed as possible. Also, we believe the most common words are still easy to pick out on account of their size.