2

Amherst History Report

Nicole Frontero & Sabir Meah

October 3, 2020

Contents

nice job! some thoughts:

1) I would prefer a function to do the processing that you have within the for loop. But you can keep what you have

2) ~~you weren't able to fix~~ I couldn't confirm that you glued together things like col- lege → college with your code. can you please confirm?

3) Table 1 was nice! perhaps extend to 250 characters (would need to be a long table

4) I might suggest removing "college" and "amherst" after you display table 2 then add a new Table 3 that lists the top 10

5) I would suggest report top 3 most frequent words per chapter O history (n=13), XX (n=8), XX (n=6), etc.

6) Same for tf-idf (also suggest not displaying tf idf

add author

Executive summary

We set out to see what we could learn about the book *History of Amherst College During Its First Half Century, 1821-1871* through text analysis. We wanted to gain a big picture understanding of the entire book, including a sense for the structure of the book, and also gain insights about individual chapters. To accomplish this goal, we used word frequencies, ~~term frequency-inverse~~ document frequencies (~~tf-idf~~'s), and word clouds.

jargon: need to define + motivate

Word frequencies allowed us to see the most frequent words that appeared throughout the entire book and in each chapter. We used tf-idf's to give us a sense of certain words that may be more common in specific chapters, and to help us identify certain themes in the chapters. Finally, we also created word clouds to display the top 100 most frequent words in each chapter and in the book as a whole in a compact and colorful way.

We wanted to provide a visualization that would serve as an overview of the book as a whole in this summary. We have reproduced figure 2 (found in the analysis section), a word cloud for the 100 most frequent words to appear in the entire book, below.

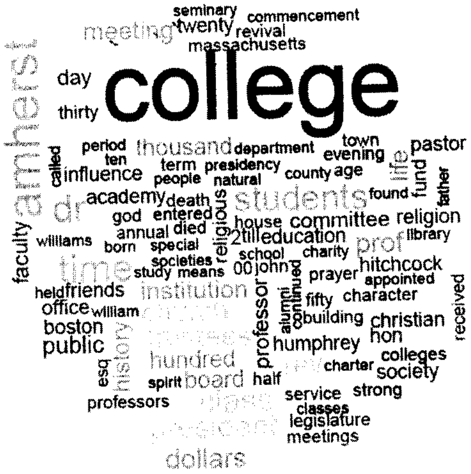

Figure 2: Word cloud of the entire book

add word cloud for a specific chapter here (to compare)?

When we look at figure 2, we see many words that we would still characterize as being pertinent to Amherst College today. For example, "department," "trustees," "students," "dollars," and "fund" are all words that color our current conversations and experiences as Amherst College students, even though this book tells the story of Amherst from over a century ago. However, there are also words in figure 2 that are less relevant to the present day Amherst experience, such as "legislature," "christian," "revival," "academy," and "seminary." While these words were more common and integral to Amherst in the 1800s, they are less pertinent to Amherst College today, albeit not completely absent. The difference between the words that characterize Amherst in this book, and the words that we may use when discussing Amherst highlights the ways in which Amherst has changed in the past century and a half.

We should note that in addition to trying to learn about the book as a whole, we also tried to learn more about each of the 29 chapters. We performed every analysis method on both the book as a whole and on the individual chapters (except for tf-idf's, which we performed only on individual chapters). Performing the analyses on the individual chapters allowed us to gain insights into the themes and topics of the chapters themselves, and also allowed us to get a sense for the structure of the book.

Performing text analysis on this book allowed us to extract and distill the key information from 722 pages of Amherst history in an efficient and concise manner. While reading a book obviously is one way to gather information, we have learned that text analysis can provide important summary measures for a text and can allow for identifying patterns that may not be easy to recognize when reading.

Are there any specific insights?

the first 50 years of