Anna Buchele
Softdes Spring 2016

Text Mining Project Overview

For my text mining project, I used several books from Project Gutenburg. I hoped to analyze the words used most often in books over a period of time, to see if the more freqently words changed in any recognizable way. To do this, I used the most popular book published in each decade (according to Goodreads.com) from 1900 to 1970 and compared the 100 most frequently used words of each book to see which frequently used words the books had in common.

Implementation

I implemented this by first making a program to edit each book to remove punctuation, weird/unusable characters, capitalizations, and excess empty space. The program then turned each book into a list of words, and calculated the frequency of the words used and returned a dictionary of words used and the number of times they appeared, sorted from most used to least used. The second program was used to import all the texts, run them through the first program, and return a chronological list of dictionaries for each book. The final program ran the previous programs, and added the 100 most frequent words of each book to a nested chronological list. It then calculated the number of books studied for which each word was in the 100 most frequently used words. Then, it sorted those words and numbers into several lists: words used very frequently in: every book, every book but one, 4-6 books, 2-3 books, and only one book. Then, it found which decades' books contained (or did not contain, for the words used frequently in every book but one) frequent usage of each word, and which decades had which groupings of frequent words in comparison to other books.

The major components of my program were: first, importing the texts so they can be read; second, turning the texts into lists of words and removing unusable portions (punctuation, etc); third, counting the frequencies; and lastly analysis. For the analysis, I transformed the word frequency dictionary into a list of the top X words to be studied using a reversed sort by value operation, and then using a for loop to put only the words into a list. I did this because I decided that once a word was in the top X words used in a book, the exact number of times it was used didn't really matter. Since books have different lengths, it might have been better to provide some sort of analysis by book length, but I decided that even short books would likely have the same sort of word frequency as longer books. Shorter books would have more skewed data regardless, and I assumed that the shorter books would have less common words in their list of most frequently used words.

After that, I compared the other books' list to each book's list (making sure that each book is

only compared to the ones after it in the list, so a word isn't counted double or more based on the book's placement in the list) and counted the number of books each word appeared in. I did this because the list of books' frequent words was ordered according to decade, so I was able to just search the books' list and make a nested list of the decades it appeared most frequently in. I also counted which subset of words each decade's words fell into, as I thought that might make some interesting analysis as well.

Results

I found that the kind of words one would expect to be very frequent in every book were, in fact, the ones found very frequently in each of the 8 books studied. These words were articles, conjunctions, pronouns, question words, and prepositions. The words that were very frequent in every book but one seemed mostly random, both in the word and the decade they were missing from, and I couldn't see a pattern in them. The words that were very frequent in fewer books gave a clearer pattern: the word 'eyes' was on the list for the books from 1910-1950, but not after that. The word 'go' was on the list from 1900-1960. The words 'any', 'little', and 'will' were similarly common in the first half of the decades studied, but not after that. Many words were common in 1900-1910, and then not common again until 1970-1980. Interestingly, the word 'Mrs.' was very frequent in 1900 and 1910, while 'Miss' was common in 1930 and 1960. The words that were only frequent in one book were often character names, items or adjectives important to the story, or sometimes seemingly random words.

I've summarized my data in the table below, grouping it into the similarities I saw:

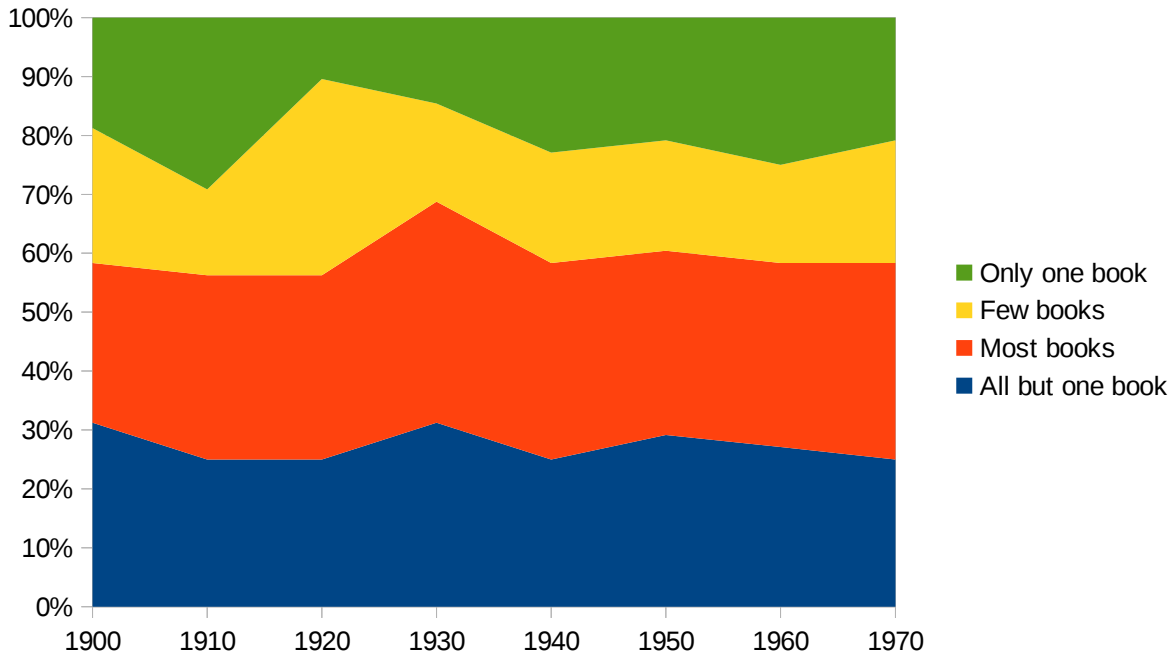
Words In 100 Most Frequent Words of:	
Every book studied	all', 'its', 'had', 'to', 'then', 'them', 'his', 'know', 'they', 'not', 'him', 'like', 'back', 'out', 'what', 'said', 'for', 'be', 'by', 'on', 'about', 'of', 'into', 'one', 'down', 'from', 'there', 'was', 'that', 'but', 'with', 'he', 'this', 'up', 'were', 'and', 'do', 'is', 'it', 'an', 'as', 'at', 'have', 'in', 'if', 'no', 'you', 'who', 'a', 'i', 'so', 'the'
Every book but one (missing decade included)	'me': 1940, 'or': 1960, 'just': 1940, 'her': 1970, 'would': 1920, 'over': 1960, 'when': 1970, 'been': 1950, 'your': 1910, 'she': 1970, 'time': 1910, 'now': 1910, 'my': 1940, 'could': 1920, 'are': 1920
Books from 1900-1940 only	'any', 'little', 'will', 'Mrs', 'before', 'even', 'much', 'too', 'seemed', 'after', 'than', 'came'
Books from 1930-1970 only	'only', 'old', 'right', 'Miss', 'again', 'other', 'face',
Books from 1900-1930 AND 1960-1970	'because', 'say', 'moment', 'going', 'oh', 'went', 'didn't', 'think',

I don't know what insight to draw from this. Some words found mostly in the earlier decades do 'sound' older than the words found in the later decades, but not in any measurable way. I do think it is interesting though that patterns seem to emerge.

This table shows the number of frequent words of each decade that fell into each category:

Decade:	1900	1910	1920	1930	1940	1950	1960	1970
Frequent words found in every book	52	52	52	52	52	52	52	52
All but one (books with word counted)	15	12	12	15	12	14	13	12
All but one (missing books counted)	0	3	3	0	3	1	2	3
Most books (4-6)	13	15	15	18	16	15	15	16
Few books (2-3)	11	7	16	8	9	9	8	10
One book only	9	14	5	7	11	10	12	10

This chart shows the percentage of frequent words (that were not in every book) for each book of each decade that fell into the categories shown.



Although we can see some interesting patterns here, I only used one book for each decade. If I used multiple books for each decade, I would be able to be much more certain of any patterns I noticed. Any differences shown here could be easily explained by differences in writing styles or book genres.

Reflection

I wanted to do more with this project, as I think this stuff is really interesting, but I have become very behind on work since I got sick and I didn't want to add this project to my long list of things that are going to be late. I think my project was appropriately scoped for the amount of time I had to spend on it and my skill in python. I understand lists, tuples, and dictionaries much better than I did before I started. There were many parts of my code that I ended up writing and rewriting before realizing I could just scrap the whole thing and replace it with a line or two, but I learned a lot from doing that and I'm glad I've learned to write more simple code. For unit testing, I didn't have a great plan, but it worked. I mostly just ran each program with the books I was planning on using, and printed the outputs to see if they were coming out like they should. While this worked for my project, I realized while doing this writeup that I was supposed to use doctests and other types of unit tests. However, adding them in now would feel dishonest, since I already know my code works as it should. I picked this project because I had been having a lot of trouble with dictionaries, lists, and tuples. I learned a lot about those from this project, which I'm really happy about. I think I understand how they work now and what they should be used for, which should make my life a lot easier going forward.