ASSIGNMENT 2 Write Up

Susanna D’souza & Eoghan Neely

October 21, 2020

Project Overview

For this project, we focused on analyzing ten different books written by four different genre authors using the Gutenberg Project. We did four techniques: 20 most common and 20 most uniquely common words in each text excluding stopwords with the goal to compare the differences. We then used natural language toolkit technique for each text to determine the overall tone of the book. Finally, we did clustering to compare two texts and test our hypothesis we created using the NLTK process.

Implementation

The 10 texts we chose were: *Sense and Sensibility*, *Emma*, *Pride and Prejudice,* and *Mansfield Park* all by Jane Austen, *Little Women* and *Little Men* both by Louisa May Alcott, *Poirot Investigates* and *The Man in the Brown Suit* both by Agatha Christie, and *The Raven* and *The Cask of the Amontillado* by Edgar Allen Poe. The first step we took was to clean each text by removing the header and footer “Start of Project …” and “End of Project…” details that did not pertain to the actual text.

Our four questions we answered in this project:

1. Excluding stopwords, what are the 20 most common words used in each text?
2. Excluding stopwords, what are the 20 most common words unique to each text?
3. What is the overall tone of each text?
4. Using text clustering and text similarities, given all texts but one, which text is this closest to? Then display it on a chart using text clustering.

For the first and second question, our implementation idea was to first create dictionaries of solely the words for each text in order to streamline the project and coding. The similar coding for question one was also used for question two as the questions were similar, however, since question two targeted unique words - we needed to create another dictionary as a first step before implementing question two’s main function. This second dictionary would be for 9 out of the 10 texts, in order to compare the 10th text to this dictionary to find words that are unique to the text. There might be a method that uses less steps to answer this question, but this was the easiest conceptually for us to answer the question.

A design decision that we made was with question 3. In order to attack this question, we first considered using NLTK to analyze each sentence that had an exclamation point. Then, we would average the result for each sentence in each text to get the overall tone. This would eventually retrieve the overall tone of the text - the answer, but it would be much too complicated and might end up not giving a fully accurate answer by our implementation technique. Instead, with Professor Zhi's help, we were able to gather the overall tone by a simple code from NLTK that analyzes the *entire* text in one line and gives the overall tone results. In the beginning, we had this function return the score[compound] but after realizing that this normalized the score and most of the answers we received were 1.0, it wasn't helpful or enough information to analyze, so we changed the function to give the entire score breakdown.

For text clustering, we first had to calculate the text similarity scores. We used this link(https://dev.to/coderasha/compare-documents-similarity-using-python-nlp-4odp) to initially help us create the code. However, the code we created did not utilize all the texts in it's list and rather, by our conceptual understanding, broke down a text by each sentence. With Professor Zhi's help and code, we were able to fix this issue and test it on *The Man in the Brown Suit* by Agatha Christie. As we continued this thought process and worked on our function for text clustering, we realized that for the array part of function, we need the similarity list for each book so we created multiple variables to get the similarity for each text, not just the Agatha Christie novel. Our hope was that when the array printed we could then compare to see where the book was closer to. However, the implementation of that final function did not work out.

Results

Our results for question 1 gathered some interesting insight. A lot of the most frequent 20 words in each text showcased a lot of names and titles such as "mr" or "miss" or "mrs". The text with the highest word count for the most frequent word was *Little Women*, with the most said word being "Jo" for 1250 times. Jo is the main character in the book and this makes sense for it to be the most frequently used. The results also gave insight to the shortest novels, both written by Edgar Allen Poe as both his books didn't break 50 times for the most frequent word. In Poe's *Cask of Amontillado,* the most frequent word "said" occurred 24 times. In fact, the word "said" is found as one of the most common words in 9 out of the 10 texts, the only book it is not most common in, is *The Raven*, another Edgar Allen Poe novel. Another important result we noticed in *Poirot Investigates* and in *The Cask of Amontillado*, both written by different authors, is that the code picked up on multiple spaces in certain spots and counted it as a word which came on the most frequent list. If we were to continue to work on this project and make changes, this would definitely be an issue we would address and fix.

The results for question two were quite interesting. A lot of the most frequent words that we saw when we compared to the results from question one, were more of names and variation to names, such as “amy’s.” Another point noticed is that the number of frequencies decreased to less than 100 compared to when we had done question one, suggesting that there were a lot of common words between the texts. In the novel, *Emma*, we noticed that for question 2, the word “surprized” was considered unique, perhaps due to its unique spelling of the word. In addition, Jo, the main character in *Little Women,* is also the main character in *Little Men,* so her name does not show up as most uniquely common for either books when doing question 2.

Emma Results 1:
The most common words are:

mr	1154
emma	787
mrs	701
miss	602
will	573
must	571
much	486
said	484
one	448
every	435
harriet	415
well	403
thing	398
weston	389
think	384
little	361
good	359
never	358
knightley	356
know	337

Emma 2:
The most common unique words are:

emma	787
weston	389
knightley	356
elton	320
woodhouse	278
fairfax	210
churchill	193
hartfield	159
highbury	125
harriet's	91
randalls	90
emma's	79
perry	74
elton's	67
isabella	56
weston's	51
donwell	49
dixon	39
woodhouse's	37
surprize	3

Little Men 1:

The most common words are:

said	625
little	611
one	534
mrs	382
like	373
dan	363
bhaer	341
will	336
nat	332
boys	328
mr	286
jo	282
good	280
demi	280
much	275
see	270
now	239
nan	226
well	211
daisy	210

Little Men 2:

The most common unique words are:

dan	363
nat	332
silas	50
dan's	35
brook	25
nat's	23
nan's	21
charlie	21
tommy's	20
toby	20
robby	20
nursey	17
goldilocks	14
owl	13
crane	13
melons	11
beans	10
stuffy's	9
kites	9
rewards	8

Little Women 1:

The most common words are:

jo	1250
said	827
little	727
one	711
meg	635
like	591
amy	572
laurie	549
will	502
good	462
beth	415
now	399
go	394
old	378
never	375
much	373
well	368
see	357
mother	328
away	328

Little women 2:

The most common unique words are:

amy's	80
beth's	52
fred	51
sallie	42
haf	31
moffat	28
kate	28
march's	21
flo	18
davis	17
pickwick	16
carrol	16
tina	15
kirke	13
esther	13
bundles	13
traveling	12
roderigo	12
gif	12
zara	1

Our results from question 3 yielded that the highest negative component was for *The Cask of Amontillado,* followed by *The Man in the Brown Suit,* followed by *The Raven.* The least negative was *Emma.* The most positive on the other hand was *Little Women,* followed by *Little Men,* followed by *Emma.* The least positive was *Poirot Investigates.* The most neutral was also *Poirot Investigates,* by *The Man in the Brown Suit* (the same author), and *The Raven.* This was interesting and it kind of made sense to us. Agatha Christie novels are usually filled with suspense and mysteries while Edgar Allen Poe's books are usually gloomy and about murder. Jane Austen and Louisa May Alcott books are about adventure, coming of age, and love - so really more on the positive side than negative.

The result from the text similarities part for text clustering technique yielded, to our surprise, that *The Man in the Brown Suit* by Agatha Christie was closest to *Little Men* by Louisa May Alcott. This surprised us because Agatha Christie's novel yielded as the second highest neutral toned book as well as only a 0.12 positivity and 0.096 for negativity compared to *Little Men* at 0.178 for positivity and 0.086 for negativity. With Christie's book more negative than Alcott's and less positive, we realized that this text similarity technique looks further than tone.

Reflection

This project was a lot of fun and it gave us a good opportunity to try to implement a lot that we had learned so far as well as learn new techniques. The way we went about this project was to first come up with the texts we wanted to work with and the questions we wanted to answer. We then divided the work for questions 1 and 2 between the two of us while paired programming for question 3 and 4 as they used techniques we both were not familiar with. It was extremely helpful to have previous examples to refer to as well as to be able to ask Professor questions on problems we ran into. Pseudo coding became extremely useful for working through our thought process and implementation of coding. If we could do anything better or had more time, we'd work on how to show question 1 and 2 side by side for each text in order to compare the results.We would definitely work further on our fourth function, to figure out how to make it run and display the chart. We did write most of the code and have left it as Pseudo-code. If we were to do this project all over again, it would have been helpful to have more reading materials on the new techniques in order to figure out how to fully utilize the new skills and toolkits.