Yifan Wang

Professor Zhi Li

MIS3640 - 02

18 October 2020

Project Write-up and Reflection

Project Overview

I used the data souce from Project Gutenberg. To analyze the text, I used five techniques: characterizing by word frequencies, computing summary statistics, doing natural language processing, text similarity and text clustering. I hoped to learn if there will be much similarity among the books of one author and if the words chosen by different authors will be significantly different.

Implementation

I started with *Little Women* and *A Study in Scarlet,* and stored the words appeared in each text and their frequency in two dictionaries. Then I started to figure out the words that appeared most common in each text. To make it more accurate, the function can skips the words in "stopwords.txt". The function can also take in a list of characters' names, and skip those names. , Another function is designed to find out if there is any word that appear in the 100 most common words in one text, but not in the other. However, it was hard to figure out some patterns from a list of words, and thus I started to perform natural language processing and analyzed the sentiment, whether the author has a positive, negative or neutral tone, and whether the two authors will have a different tone.

In addition, I studied the similarity between two texts. Firstly, the function combines the dictionary of the two texts and create a new dictionary with all unique words. Then the function creates two vectors with length equal to the count of the words in the position of the new dictionary, and calculates a cosine similarity of the two vectors. After that, since I wanted to analyze the similarity between more than two texts, I created a new function that calculates the cosine similarity between each pair of the inputs and returns a similarity matrix. Finally, I used Metric Multi-dimensional Scaling to visualize the matrix.

One design decision I made was whether to calculate the matrix manually or to design a new function to calculate the matrix. I decided to design a new function because it is much more convenient when I wanted to add a new text. I did not have to calculate the similarity of the new text with each of the previous texts, and all I have to do is to add a new parameter into the function.

Results

Even though I have skipped the stopwords and names, the 10 most common words in *Little Women* and *A Study in Scarlet* are still hard to interpret. The only thing interesting here is that there are really a lot of conversation, and thus a lot of "said"s in novels.

```
The most common words in Little Women are:
said     827
little   727
one      711
like     591
will     502
good     462
now      399
go       394
old      378
never    375
```

```
The most common words in A Study in Scarlet are:
said     207
upon     201
one      163
man      131
will     95
little   83
now      80
two      79
time     76
come     71
```

The words that are in the top 100 most common words in *A Study in Scarlet*, but not in the top 100 in *Little Women* are as below:

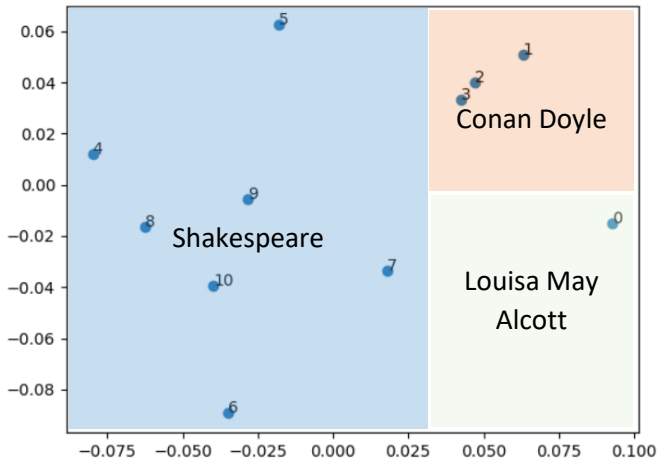
```
['answered', 'room', 'however', 'hope', 'ferrier', 'drebber', 'case', 'must', 'might', 'door', 'every', 'don't'
, 'appeared', 'companion', 'night', 'enough', 'whole', 'small', 'morning', 'cab', 'still', 'round', 'ring', 'na
me', 'men', 'right', 'remarked', 'daughter', 'yet', 'street', 'spoke', 'matter', 'left', 'knew', 'road', 'follo
wed', 'death', 'blood', 'without', 'three', 'seen', 'heard']
```

This word list makes much more sense. We can see words such as "death" and "blood", which certainly are common in detective stories, but probably uncommon in other novels. Also, "cab" should not be common in *Little Women* as cabs did not appear until the late 19th century.

Then, I studied the tone of the two authors. I expected the tone of Conan Doyle to be more negative, but the result is different from what I thought. While both overall tones are positive, Conan Doyle uses more neutral language than Louisa May Alcott. The use of neutral language in *A Study in Scarlet* reflects the objective perspective of Sherlock Holmes, while the tone in *Little Women* seems to be slightly more positive.

```
Little Women {'neg': 0.093, 'neu': 0.719, 'pos': 0.188, 'compound': 1.0}
A Study in Scarlet: {'neg': 0.092, 'neu': 0.804, 'pos': 0.104, 'compound': 1.0}
```

Finally, I studied the similarity between the novels of the same authors and of the different authors. To better see the relationship, I added two more books from Conan Doyle and seven books from Shakespeare. As shown below, we can see that each author seems to occupy part of the graph, and the division from author to author is clear. The differences among the works of the same author really depends on the author. For example, we can see that three novels of Conan Doyle are really similar to each other, with cosine similarity more than 0.9, while the novels of Shakespeare can be very different. Not only is there a huge difference between comedies and tragedies, but there is also different between tragedy and tragedy. Simply from this graph, we can have a brief knowledge of how Shakespeare's writing styles vary and how large his vocabulary size is.



0	Little Women
1	A Study in Scarlet
2	The Sign of The Four
3	The Adventures of Sherlock Holmes
4	Romeo and Juliet
5	Hamlet
6	Othello
7	Macbeth
8	King Lear
9	The Merchant of Venice
10	A midsummer Night's Dream

Reflection

In conclusion, my analysis of texts generally went well. I met my target to find the differences among different authors and the differences among different works of the same author. There are also several things that I can improve. The first thing I can improve is to have a clear plan of each function and then start to code. For this assignment, I wrote so many things in the first function at first that I cannot reuse some of the dictionaries generated during the function. I spent some time to divide the big function into smaller functions, and I believe I would be able to save some time if I plan ahead. In addition, I can also improve my analysis by adding more data source. I would expect the words used in a novel written in last centuries to be different from words used in a movie review today.