Problem Solving & Software Design Assignment 2

Naibo (Ray) Hu

Xuran (Angela) Wang

Feb 26th 2020

We pledge our honor that we have neither received nor given unauthorized assistance during the completion of this work.

1. Project Overview

The data source that our group uses is from Gutenberg, and the three books that we select are *Walden and on the Duty of Civil Disobedience* by Henry David Thoreau, *Beyond Good and Evil* by Friedrich Nietzsche, and *Alice's Adventure in Wonderland* by Lewis Carroll. The book *Walden* is the main book that we study, and the rest two books are supplementary. The techniques that we use **include characterizing word frequencies, computing summary statistics, text similarity, text clustering, and Markov text synthesis.** Our group hopes to gain insight into the themes and study the similarity and differences among the books by analyzing the text.

2. Implementation

In order to characterize word frequencies and complete the top 10 highest-frequent words in the book *Walden*, we first remove stop words and then use the dictionary function, for loop, and if statement to define functions.

In order to study text similarly, we used all three books. We first change the punctuation to spaces, map uppercase to lowercase, and split text-lines into two words. Next, we return a dictionary that matches the words to their frequency. Finally, we calculate the dot product and the vector angle of the two selected books.

In order to compute text clustering, we also used all three books. Based on the result from text similarity, we calculate the dissimilarity, compute embedding, and draw the scatterplot that shows the relationships among three books.

In order to conduct Markov text synthesis, we used the book, *Walden*. We create a corpus that splits the text file into single words. We decide to keep all the punctuation. Next, we create word pairs using dictionary. Finally, we randomize all the words and set the total word limit to be 100.

Our group decides not to do sentiment analysis because as a philosophical book, *Walden* contains complicated emotions. It is hard to track certain emotional words such as "happy" and "sad" to determine the mood of the entire book.

3. Results

Result of summary statistic analysis

The top 10 frequent words in the book *Walden* is shown in the following figure. As we can see, the book is about human nature and life necessities, so it makes sense that words like "house," "life," "man" are high-frequency words.

one 502
man 321
like 310
will 274
men 238
may 222
house 203
pond 200
life 200
day 196

Result of Text Similarity Analysis

From text similarity analysis, we find out that the distance between *Walden* and *Beyond Good and Evil* is 0.270507; the distance between *Walden* and *Alice's Adventure* is 0.525258; the distance between *Beyond Good and Evil* and *Alice's Adventure* is 0.608798.

The distance differences definitely make sense since the book Walden and the book Beyond Good and Evil are both philosophical books about human nature. So, they are similar to each other, resulting in a smaller distance. *Alice's Adventure*, on the other hand, is a fiction book and thus has a larger distance compared to the other two books.

Python output:

File Assignment 2/Walden.txt :

654237 lines,

120565 words,

11793 distinct words

File Assignment 2/Beyond_good_evil.txt :

402269 lines,

67737 words,

8331 distinct words

The distance between the documents is: 0.270507 (radians)

File Assignment 2/Walden.txt :

654237 lines,

120565 words,

11793 distinct words

File Assignment 2/Alice.txt :

164201 lines,

31007 words,

3545 distinct words

The distance between the documents is: 0.525258 (radians)

File Assignment 2/Beyond_good_evil.txt :

402269 lines,

67737 words,

8331 distinct words

File Assignment 2/Alice.txt :

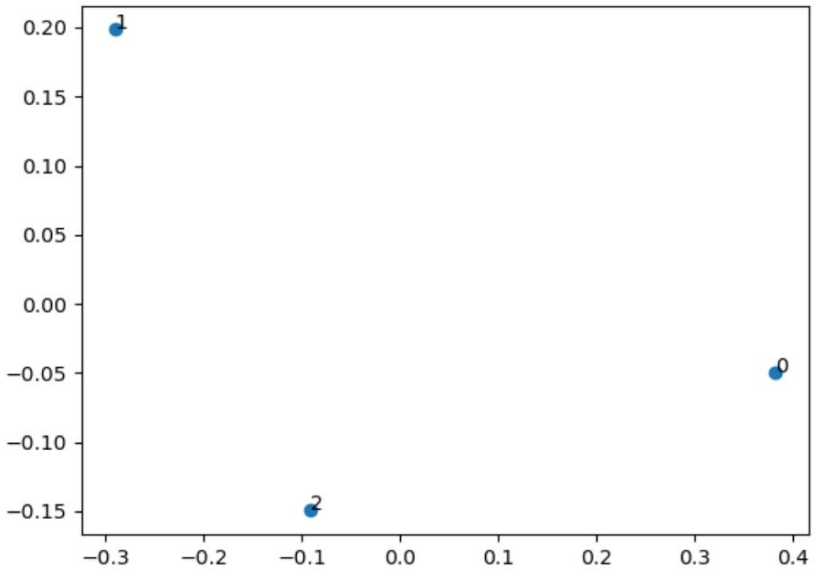
164201 lines,

31007 words,

3545 distinct words

The distance between the documents is: 0.608798 (radians)

Result of Text Clustering Analysis:



0 represents the book Walden

1 represents the book Beyond Good and Evil

2 represents the book Alice's adventure.

The graph above shows the dissimilarity among the three books.

Result of Markov Text Synthesis

The following output is randomly generated from the Markov text synthesis. The logic of this method is that we first count the words that appear after each word and the number of times. When the current word is determined, calculate the percentage of the next word. Use this percentage as the probability of the next word in the text generator for text generation.

Python output:

I would reach his insectivorous fate. Self-emancipation even the midst of their whistle till at sunrise. To act must keep the land by the surface, nor sit in procession with the chimney after staying in his, and a year, for fuel in a small imps that the winter, the staff

of weeds." He would take, and I could. I _survey_, My gay butterfly is eight pounds because they can converse about Project Gutenberg-tm trademark. Project Gutenberg Literary Archive Foundation are the outskirts, having been sawed off, like reason for all, become a reader, a time, and wider still it is, for

4. Reflection

The best part of the project is the working dynamic within our group. We encourage each other to move forward along the way. Our group divides the work evenly. One person is in charge of doing the text similarity, text clustering, and Markov text synthesis. Another person is in charge of computing summary statistics, characterizing by word frequencies, and writing part three deliverable. The problem that we face is that some coding parts are not easy to figure out, so we use Google to search, watch YouTube videos, and ask the professor for help. Next time, we will try out different sources such as Wikipedia, Twitter and so on and use sentiment analysis in our work. Overall, our project is appropriately scoped since we try almost every analysis method. We will definitely improve on not bothering the professor that much! Thanks for helping us, professor!