Pranjal Joshi and Christian Thompson

Problem Solving and Software Design

Professor Zhi Li

11/12/18

Assignment #2: Text Mining and Analysis:

1. Project Overview

Our mini-project is centered around two novels by Agatha Christie titled *The Mysterious Affairs at Styles* and *The Secret Adversary*. The books were downloaded from Project Gutenberg as text files and served as data sources for our project. We aim on using word frequency analysis and computing summary statistics to compare the two different books written by the same author. Through this analysis we want to see if Agatha Christie has a specific writing style that is portrayed in both these books through the use of similar words. We're hoping to learn if this analysis can be done easily to be used on other works by same authors.

2. Implementation

In order to analyze the similarity of the text, we took the text of both the books from Project Gutenberg and adapted it in order for it to be processed by the code. To get more accurate results, we had to eliminate parts of text that were not written by Agatha Christie but put in the text by Project Gutenberg. The one alternative we did was to process the text files locally instead of trying to access them through the web because we could not figure out how to trim away the excess preamble and space present in the text.

Two major components we focused on were using lists and dictionaries so that it would be easy for us to compare the most frequent words in each text along with the words that were common in both the texts. It was important for us to use functions that allowed us to strip the text of stopwords (such as 'and', 'a', 'the', etc.) so that our results were not skewed to provide unreliable data. Furthermore, displaying the results in a list format allowed us to see which non-stopwords overlapped in both the texts. There were more words and phrases that appeared in both the texts but we focused on the ones that appeared more frequently (more than 100 occurrences).

3. Results

The Mysterious Affair At Styles	Frequency	The Secret Adversary	Frequency
Poirot	352	tuppence	548
said	252	tommy	511
mrs	227	said	479
inglethorp	227	"i	321
will	211	julius	286
"i	200	one	240
one	180	mr	238
mr	160	sir	212
john	153	know	208
know	132	don't	165

Additional Output:

The following words appear in both books:

produced, by, charles, keller, the, mysterious, affair, at, agatha, christie, contents, chapter, i, go, to, ii, and, of, july, iii, night, tragedy, iv, v …

There are a lot more words printed out when the file is executed, but to keep the overview short, we limited the amount of output displayed here. We chose not to include this extensive list of words in the above table because this list contained skewing results such as roman numerals that are used to list chapters in the beginning of the book, the author's name, repetition of the book's title, etc.

From the results we can conclude that the most common words used in both the books were the names of the characters along with their suffix (Mr., and Mrs,). Though this may not tell us the writing style of Agatha Christie, it does tell us what the characters' names are along with their gender. This type of data can be used by book review sites or online ebook websites to write a brief description of books. For example, we can see that the main characters in both the books are male. We can assume that the main character in The Mysterious Affair At Styles is called Mr. Poirot and the main character in The Secret Adversary is called Mr. Tommy Tuppence, as these are the most occuring words.

4. Reflection

From a process point of view, we came to a common idea rather quickly. We knew we wanted to explore word frequencies and summary statistics. After reviewing the Project Gutenberg database, we thought about potentially comparing books by the same author, looking

to see if the same author consistently used the same words in their writing, as well as how often they used these words. While the idea came easily, the formulation of the code necessary to complete this analysis was much harder. We kept running into obstacles that prevented us from completing the assignment. In particular, stripping out the preamble and finding ways to print out the top 10 words used in each text took multiple hours of trial and error. In order to improve our project, we feel as though running through our code with Professor Li outside of class in office hours might have helped us improve our analysis. Prior to the project, we planned to evenly divide the work up, so that we each spent some time devising the code and write up for the assignment. While working on this assignment, we stuck to this formula and it paid off, as we both felt that we came away with newfound knowledge regarding text mining. Our hours of hard work were imperative, as we were able to create a program the computed the necessary data and analysis we were seeking. We didn't encounter any problems while completing this project, but if we were to do things differently, we would have investigated even further and done more statistical analysis on our data set.