Data Mining: Analyzing Profanity in Classic Novels

Robbie Siegel

February 2015

1 Project Overview

I decided to analyze books from Project Gutenberg's Top 100 most downloaded list. I used pattern to download these files, analyze them, and after looking through pattern's documentation, decided to check the different texts for how much profanity they included. I thought it would be interesting to see how classic and popular novels use profanity differently.

2 Implementation

My program analyzed the different texts using three simple functions. First, I used the create dictionary function to create a dictionary mapping each word in the text to its number of uses. This was accomplished by passing each line of the text to a function I called inside create dictionary, named update dictionary. This function looked at the string being passed in, stripped out the white space and punctuation, and then added each word to the dictionary of words. If the word is already in the dictionary, it will instead increase the value by 1, effectively keeping track of the uses of each word. Once all the lines of the text have been analyzed, create dictionary will return the dictionary of all the words.

The next function I used, analyze profanity, checked each dictionary for words considered profanity. Whenever a profane word was found in the dictionary it would be added to a list of all the curse words. In addition, I included a variable to keep track of the total number of profane words used throughout the novel by adding the values of each word, or key, that was found to be profane.

The final function I included plotted the data I found from my first three functions. I decided to plot the number of different curse words used on the x-axis and the ratio of profane words to total words in the novel on the y-axis. With these axes, more profane novels were plotted towards the upper right, while those considered less profane were plotted towards the bottom left.

3 Results

I ran my program on six different novels: The Wizard of Oz, Moby Dick, The Divine Comedy, The Adventures of Huckleberry Finn, Beowulf, and The Scarlet Letter. I believed these six novels spanned different time periods and literary styles, so I was interested to see how each novel employed profanity.

In addition, I created a plot to represent the level of profanity in each of these books. I plotted the number of different curse words used in each novel on the x-axis and the ratio of profane words to total words in the novel on the y-axis. I believed this would be an interesting way to represent the data, shown below:

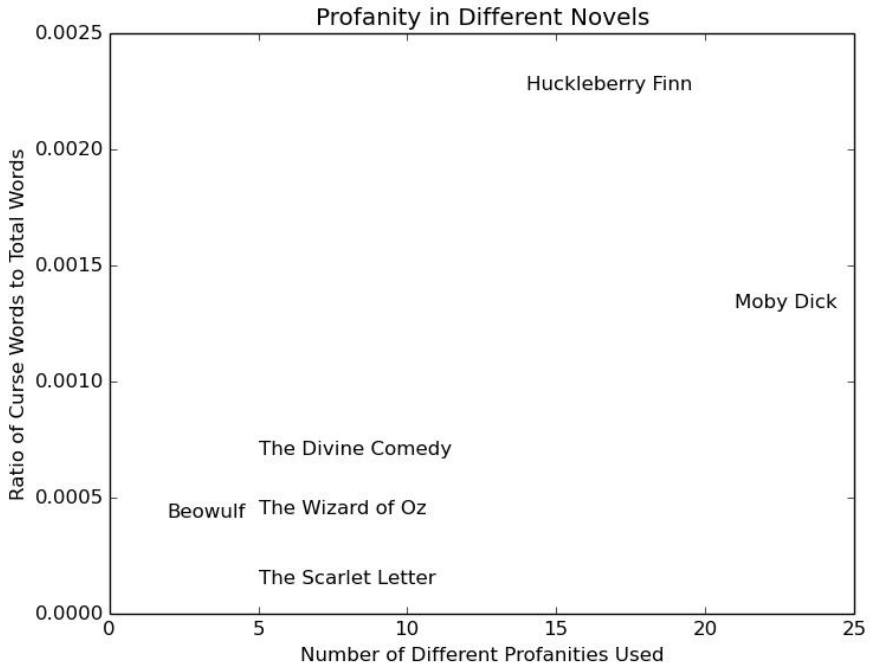

Figure 1: Comparison of Profanity in Five Classic Books

This plot clearly displays a cluster that four novels fall into in the bottom left of the plot. These novels have fewer different curse words and a relatively small ratio of total curse words to total words. The two novels which do not fall in this cluster are The Adventures of Huckleberry Finn and Moby Dick.

4 Reflection

Although I like the topic I chose, in retrospect, I wish I had found some more profane novels to make my results more interesting. The Catcher in the Rye is one novel that I think would be very interesting to analyze, so I was disappointed that I could not legally find an online text file of this book. I also think it may have been more interesting to look at profanity in other data sources, such as data from facebook or twitter.