CS 153 PROGRAM ASSIGNMENT #8

Computational Stylistics (big data)

Ben Durao

November 11, 2016

Who Wrote it?

Given a text by Jane Austin, Charles Dickens, Mark Twain, and Unknown, an investigation was started to figure out who the Unknown author could be. In building relative.c, we are able to calculate the relative frequency of each author’s letter usage. The data found in relative.c was collected in the form of a chart and outputted to a separate file. Once the charts of each author were created RMSdifference.c was constructed to find the root mean square of the difference (RMSD) between two author’s relative frequencies. By creating just a single value of comparison, we have found a way to determine the style in which an author writes by the letters they commonly use. With the idea that an author has a consistent writing style, we can determine who wrote the Unknown sample by comparing each author’s RMSD to it. The results of the investigation can be found below in the confusion matrix. From this data, the investigation can best conclude that Jane Austin is the author of the unknown sample because her RMSD value is the closest to zero out of the suspect writers. This investigation can be improved upon by comparing Unknown to more authors, or by gathering more texts from the current authors to gain a better sense of their overall relative frequency of letters used. By providing more data we can more accurately determine who the Unknown author truly is.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Austin | Dickens | Twain | Unknown |
| Austin | 0.0 | 0.002307 | 0.006807 | 0.001497 |
| Dickens | 0.002307 | 0.0 | 0.005657 | 0.002131 |
| Twain | 0.006807 | 0.005657 | 0.0 | 0.006106 |
| Unknown | 0.001497 | 0.002131 | 0.006106 | 0.0 |