Statistical Case Studies - Assignment 1

The goal of the project is to use statistical techniques to analyze the authorship of a disputed novel. This is an open ended task – you are free to use whichever methods from this course (or elsewhere!) that you please. Your goal should be to analyze the supplied texts, draw a conclusion, and then write a report which presents a coherent argument to support your conclusion.

This is a group project and worth 10% of the total course mark. You should form into groups of size 3. You are free to choose your partner(s), however you also have the option of emailing me (gordon.ross@ed.ac.uk), in which case I will assign you a group. If you have any issues or difficulty with group formation then please send me an email. The project deadline is 4pm on the 25th October.

Background

The novel ''Frankenstein" was published anonymously on January 1st, 1818, and tells the story of the scientist Victor Frankenstein, who creates and animates his monster. Since then, Frankenstein's monster has become famous and regularly features in horror movies. "Frankenstein" is generally attributed to the English author Mary Shelley (1797– 1851) who wrote it when she was 21 years old.

A French edition of "Frankenstein" published in 1821 was the first to credit Mary Shelley as the author, and subsequent revised second and third English editions were published in 1823 and 1831, The 1831 edition is the one that is widely read today. It is generally believed that Mary Shelley is indeed the true author of Frankenstein, but historically this has been disputed, with some people believing that it was written by her husband Percy Shelley.

During the time period when Frankenstein was written, it was common for female authors to publish anonymously, and this was also done by several other well-known authors such as Jane Austen and Frances Burney. Therefore, while the publication process for Frankenstein is not unusual and is perfectly consistent with Mary Shelley's claimed authorship, it has allowed for speculation surrounding Frankenstein's 'true' author. Both Mary and Percy Shelley always claimed that the novel was the sole work of Mary.

Following the publication of Mary Shelley's second novel, ''Valperga", an anonymous 1824 review of the novel noted substantial differences between the quality of Frankenstein and its successor, claiming ''there is not the slightest trace of the same hand", and suggesting that Percy Shelley wrote Frankenstein and that Mary was only responsible for ''Valperga". Several other literary

figures made the same claim, such as the Scottish writer Walter Scott in 1928.

These suppositions were all denied by Percy Shelley - from the start, he claimed to only have been the editor of the novel. It later transpired that he had written the preface of "Frankenstein" as if he were Mary Shelley, as well as the poem Mutability, which was included in the novel uncredited. This helped fuel speculation that he may have written the whole novel. Despite this, most scholars of Mary Shelley today believe that she was the sole author of "Frankenstein"

Recently, John Lauritsen's 2007 book ''The Man Who Wrote Frankenstein" promotes the theory of Percy Shelley's authorship. Several other recent figures have made similar claims. Although these claims were not – and still are not – taken seriously by mainstream scholars of Shelley's work they did receive substantial media attention, with Lauritsen's book receiving the largest share of this attention. The dispute over the authorship has typically involved handwriting analysis and subjective discussions of literary style. Stylometric methods have rarely been used.

Assignment

The goal of this project is to use stylometry to investigate whether Mary Shelley is indeed be the true author of Frankenstein, or whether the theories attributing authorship to Percy Shelley have merit. The Learn page contains a zip file 'frankenstein.zip' which contains a number of folders. Each folder contains the full text of various novels written by an author who lived around the same time as Mary Shelley. Note that the task of attributing the book to Percy Shelley is made more difficult by the fact that he wrote very few prose works. As such, I have also given you function word counts from some of his poetry (which may have a slightly different style).

Your goals

1. Investigate whether stylometric methods from this course are able to accurately identify the authorship of the texts that have a known author (i.e. texts other than Frankenstein). This will demonstrate that these methods are accurate and reliable for studying 19th century English fiction.

2. Perform analysis to determine which of the supplied authors was most likely to have written Frankenstein (which is contained in the 'Unknown' folder).

3. Create a report summarizing your findings. This should be have a maximum page length of 5 pages, not including any pictures.

You are free to structure your report however you like, but I would expect to see the following:

- A brief introduction containing an executive summary of your findings. Try to write this so that it would be understandable to someone who did not know the details of stylometry.

- A description/investigation of the datasets used in your report. You should assume the reader is unfamilar with this.
- A short discussion of the methods which you will use for the analysis.
- Some accuracy metrics that show how well your methods perform when attributing the texts with known authorship (for example, using leave-one-out cross validation which we will discuss next week).
- Your analysis of Frankenstein, and your conclusions.

Marking Scheme

There is no single correct analysis for this type of project, so you will not be marked on the basis of how close you get to some particular model answer. The marks are not subdivided, but will be allocated on a combination of statistical approach and justification, interpretation of results in context and presentation.

- 80 – 100% A report that could be presented to the client or collaborator with little or no revision. Analysis is sound so that conclusions are well-supported statistically. Interpretation is reason- ably mature. The project should demonstrate a clear overview of the work, without getting lost in details, and be free of all but minor statistical errors. The work is to a publishable standard.
- 70-79% A report that could be presented to the client or collaborator with little or no revision. Analysis is sound so that conclusions are well-supported statistically. Interpretation is reason- ably mature. The project should demonstrate a clear overview of the work, without getting lost in details, and be free of all but minor statistical errors.
- 60 – 69% A project that could be presented after a round of revision, but without having to re- do much of the actual analysis. Some flaws in the analysis or presentation (or minor flaws in both), but basically sound. A good grasp of the statistics and context, so that interpretation is reasonable.
- 50 - 59% Major re-working required before the project could be presented, but containing some sound statistics demonstrating understanding of statistical modelling and its application. Rea- sonable presentation and organisation.
- 40 – 49% Major flaws in analysis and presentation, but demonstrating some understanding of statistics, and a reasonable attempt to present the results.
- Fail (below 40%) Flawed analysis demonstrating little or no understanding of statistics, and/or incomprehensible or very badly organised presentation.