Exercise 3

by Oleg Sobchuk and Joe Stubbersfield

Background

"Romantic" emotions are frequently present in narratives. This was true for the old narrative forms, such as folktales, and this remains true for the complex contemporary fictional narratives: novels, Hollywood films, or TV series. Often, such emotions seem to be more strongly connected to female characters. (For example, the genre of "romance" novels is often associated with the female audience.) Is "sensibility" or "love" more strongly connected to female characters than male characters in fiction?

Recently, digital corpora and text mining tools allowed us to answer such questions empirically. In this exercise, we will use the programming language R for some basic text mining and visualization, using an online collection of books: Project Gutenberg.

The Exercise

Step 1 - R stats

If you are familiar with R already, feel free to skip to the next step, if not please continue. This exercise uses the program 'R', this is open source, so is entirely free for you to download and install on your computer.

R can be downloaded here: https://www.r-project.org/

It's also a good idea to download RStudio, as this provides a more user friendly interface for R and has several useful features.

RStudio can be downloaded here: https://rstudio.com/

R is a 'command line' program, which means that you tell the program what to do by typing commands rather than pointing and clicking as you do in Excel, for example. It has a lot of flexibility with what you can do with it, from basic arithmetic to complex mathematical models - and allows you to do things like the text mining we'll do here. Many of these functions require the installation of 'packages' inside R (don't concern yourself about that for now, it will be explained).

R can be quite intimidating if you are not used to command line programs, however, in this exercise you will initially just be running code which has already been written.

Once you have installed both R and RStudio, move onto Step 2.

Step 2 - Text mining and visualization

Run the script "Gender in books.R" by clicking on it. It uses the books of Jane Austen, one of the key figures in the 19th century romance genre, and her male contemporary: Walter Scott. Running the script, section by section, you will examine which exact words are associated with "he" and "she" pronouns in the works of these two authors. The script contains two different kinds of visualization of these pronoun-linked words. Detailed instructions on how to run the script are provided in the script itself, all information after a hash (#) is for you and won't impact on what R does.

Tasks

- Examine the script, get the sense of it: which part does what?
- Try changing the number of words displayed on the plots.
- Use this script for learning about the gender-connected words in other authors present on Project Gutenberg.

After running through the script and producing the visualizations you can compare them to the model answer. After completing the exercise, reflect on the key questions below.

Key Questions

Is there any predominance of "romantic", or simply "emotional" words associated with the female characters?

Do the words associated with female and male characters differ in other ways? How do they reflect gender stereotypes?

What are the advantages of text mining? What are the disadvantages?

Further exercise

If you feel confident with the R script provided, try writing a new script to conduct text mining with different authors. You can do this by copying the existing script and editing it so it applies to the texts of a new author, compare the differences between the script for Austen and Scott to work out which parts to change. The start of this is provided for you using an earlier author (William Shakespeare) and a later one (Charlotte Brontë), relative to the first pair of authors.

After creating and running your script you can compare it to 'further exercise model script.R'. After completing the exercise, reflect on the key questions below.

Key questions

Is there any historical change in the use of "romantic" or "emotional" words for male and female characters across all these authors? (Try finding both predictable and unexpected examples in other authors!)

How might this reflect their narratives 'fitting' to their social environment?

Of course, by looking at individual authors alone, we obtain no more than anecdotal evidence about the historical change. And yet, this may point out to some potentially interesting large-scale historical trends.