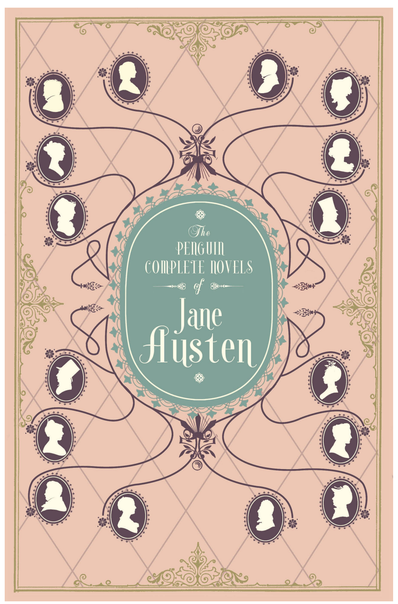
Exploring the Role of Gender in 19th Century Fiction Through the Lens of Word Embeddings

Siobhán Grayson, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene

Insight
Centre for Data Analytics

Motivation

Aim: To analyse how female and male authors of 19th century literature use terms relating to gender via word embeddings.



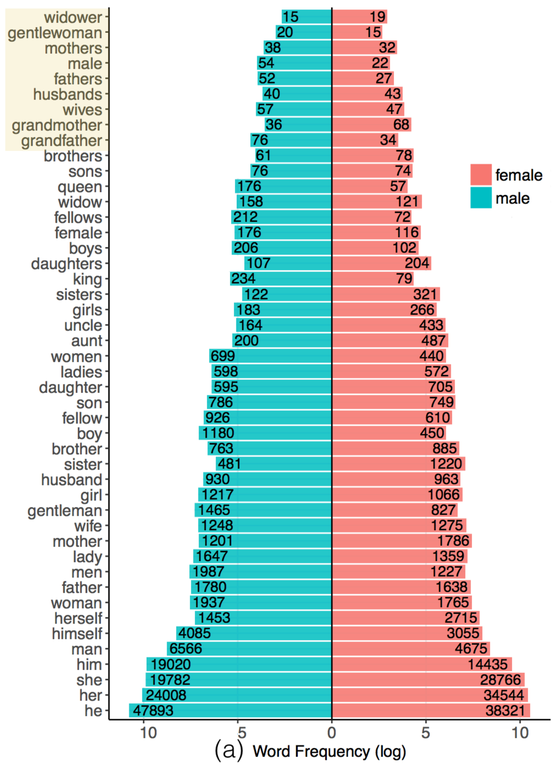
Word embedding algorithms derive a set of real-valued vectors representing the vocabulary of a text corpus in a new embedded space. This provides a useful means of measuring the underlying similarity between words.

We focus on uncovering the contexts in which female and male authors of the 19th century engage with gender specific words, by compiling a list of gender-encoded unigrams, such as 'she' and 'he', and then annotating their occurrences within our corpus to reflect the author's gender of the text they appear in ('she_female', 'he_female').

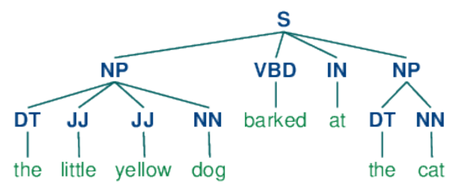
Methodology

Data Preprocessing

1. The full text of each chapter of the novel was annotated by a literary scholar to identify all characters and their aliases.



2. Each occurrence of gender encoded unigrams (Fig. a) was labelled to reflect the author's gender of the text they appear in.



3. Part-of-speech tagging (POS tagging) was applied to each text using the Natural Language Toolkit (NLTK) PerceptronTagger Implementation.

Gender	#Authors	#Novels	#Characters	#Chapters	#Sentences	#Words	%Words
Female	11	22	4005	816	111,102	2,707,884	46%
Male	18	26	6436	983	136,023	3,130,090	54%
Total	29	48	10,441	1,799	247,125	5,837,974	

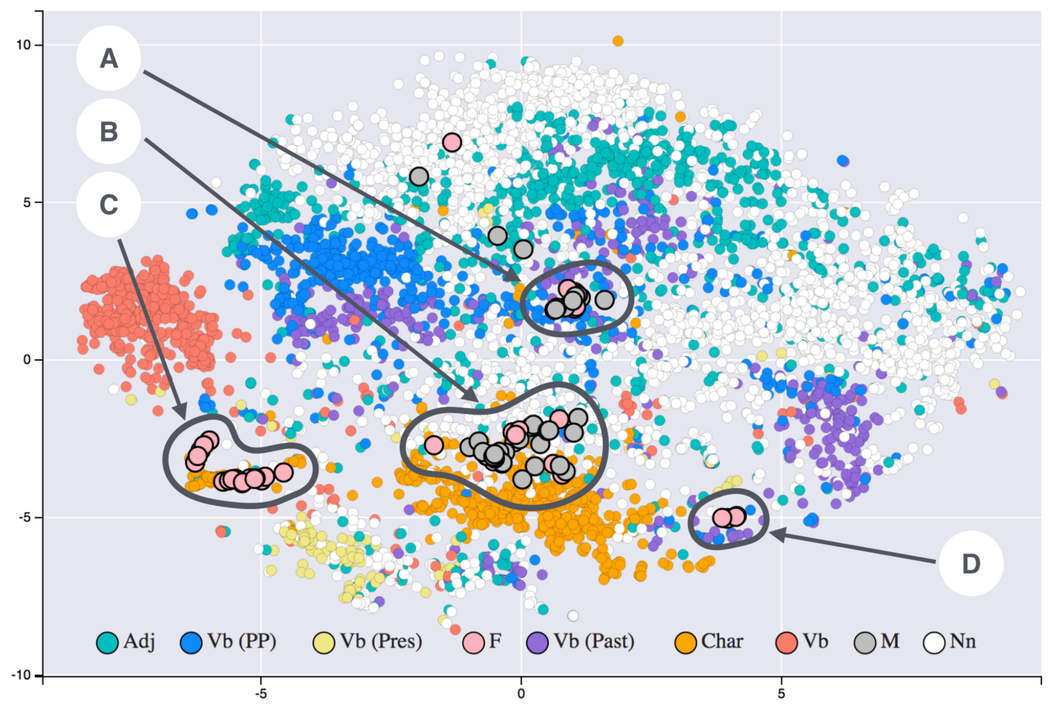
Word Embedding Generation

Word2vec is a two-layer neural network that processes text into a set of feature vectors distributed within a dense dimensional space.

1. For our purposes, we used the Gensim word2vec implementation.
2. The results presented on this poster were generated using the following parameters: Skip-Gram Negative Sampling (SGNS), 300 Neural Network Layers, Context Window = 5, and Min Word Frequency = 50.
3. Reduced word embeddings to 2D space using t-SNE and visualised.

Results and Analysis

Gender-encoded unigrams by female authors are depicted as large, pink circles while the corresponding male authored unigrams are depicted as large, grey circles. Gender-encoded embeddings occupy four different spaces within our embeddings projection annotated A-D in the figure below.



A Female- and male-authored plural nouns {fellows, women, men,..} surrounded by past-participles verbs. No family related nouns such as {daughters, sisters, brothers} by female authors despite presence of male-authored counterparts.

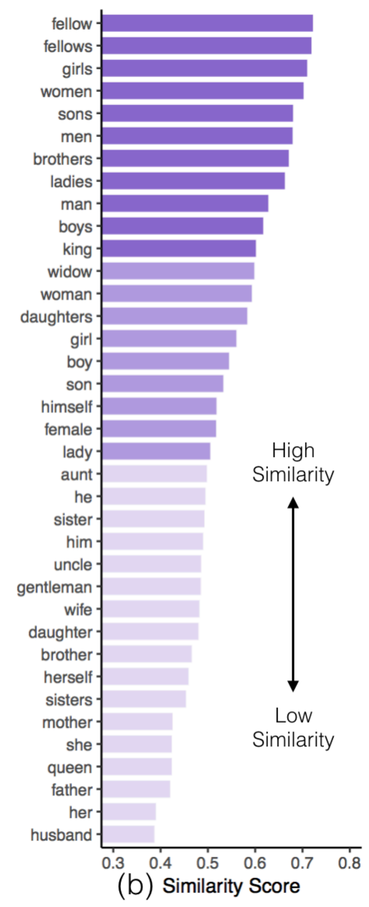
B Singular gender-encoded nouns by both female and male authors nested within nouns referring to (typically male) occupations {priest, clerk, magistrate, farmer,..}. All male-authored pronouns but only one female authored pronoun, "himself".

C Family related nouns (singular and plural) by only female authors, nested within a cluster of characters predominately from Jane Austen's novels.

D Female authored pronouns next to past-participles and past verbs. Provides interesting counterpoint to Argamon et al. [1] who found differences in how women and men use words particularly personal pronouns.

Cosine similarity of female and male annotated embeddings are displayed in Fig (b): higher scores equate to greater semantic similarity and visa versa.

Word	Gender	8 Nearest Neighbours
He	F	she^{f}, him^{f}, her^{f}, he^{m}, himself^{f}, vaguely, nervously, trembling
	M	she^{m}, him^{m}, himself^{m}, his, he^{f}, her^{m}, it, that
Lady	F	gentleman^{f}, woman^{f}, girl^{f}, ladies^{f}, heiress, lady^{m}, widow^{f}, maid
	M	woman^{m}, gentleman^{m}, girl^{m}, aunt^{m}, widow^{m}, major, maid, friend
Gentleman	F	lady^{f}, man^{f}, farmer, clergyman, bachelor, barrister, nobleman, lawyer
	M	soldier, man^{m}, lady^{m}, officer, magistrate, farmer, nobleman, colonel



Conclusion and Future Work

By generating word embeddings we have presented a new way of representing and visualising how gender is represented in well-known literary texts that complement traditional "close reading" techniques. In future work, we hope to extend our analysis to diachronic word embeddings to discover how word usage within our corpus changes over time.

[1] Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. TEXT 23, 321–346 (2003)

A World Leading SFI Research Centre

The Insight Centre for Data Analytics is supported by
Science Foundation Ireland under Grant Number SFI/12/RC/2289