Introduction

For this project, I am using a corpus of early 20th century British science fiction novels.

I collaborated with Nick Kalinowski and Humaira Halim on this project but deliverables are my own.

This is composed of four novels:

1. Frankenstein by Mary Shelley
2. The Invisible Man by HG Wells
3. Flatland: A Romance of Many Dimensions by Edwin Abbott
4. The Last Man by Mary Shelley.

Some of the questions I originally asked:

1. Are there similarities between the sentiment of male and female authors in this genre?
2. How socially relevant are the terms in these novels? Do they mimic the hardships of the times?
3. Is there a way to summarize some of the themes throughout these novels and compare them?
4. How do the male authors speak on interpersonal relationships as opposed to the female authors?

This report aims to discover relationships inherent in the textual structure of these novels to answer these questions.

Source Data

Provenance: The CSVs used for these came from Project Gutenberg

Location: The four text files can be found on my GitHub:
https://github.com/namannaUVA/DS5001FinalProject/tree/main/Data

Description: This corpus contains four novels from early 20th century British authors within the science fiction genre.

Format: All files are in CSV format.

Data Model.

Describe the analytical tables you generated in the process of tokenization, annotation, and analysis of your corpus. You provide a list of tables with field names and their definition, along with URLs to each associated CSV file.

Word2Vec

- Coordinates

	term_str	vector	n	max_pos
0	the	[0.07280652, -0.09591325, 0.073596425, 0.23378...	20756	DT
1	of	[-0.06330172, -0.27578464, -0.013809047, 0.044...	12461	IN
2	and	[0.036803376, -0.09782517, 0.07025755, 0.03682...	12152	CC
3	to	[-0.050558247, -0.07689632, 0.09011419, -0.065...	9153	TO
4	i	[-0.17635722, -0.024475018, 0.13872357, 0.0789...	7314	PRP
...	...	...	...	...
642	carried	[-0.0143698035, -0.104446515, 0.020087391, 0.0...	50	VBD
643	move	[-0.0343015, -0.110164605, 0.052107602, 0.0797...	50	VB
644	creatures	[-0.014922435, -0.11850482, 0.060565293, 0.083...	50	NN
645	walls	[-0.010944443, -0.113337934, 0.0360156, 0.0711...	50	NNS
646	letter	[-0.03745977, -0.10404693, 0.036653895, 0.0687...	50	NN

A DOCS table was created from the TOKENS table by applying a variety of transformations on the tokens to get a list of words. This was then used in the word2vec function, along with setting the vector size to 256 as a default size and for ensuring good quality vector representations. Additionally, a default word count of 50 was set so that words with fewer than 50 occurences would be dropped. Single word docs were also removed.

Upon creating the word2vec object of embeddings, the coordinates could be created and plotted using tSNE to visualize relationships between words in these specific texts.

Latent Dirichlet Allocation

To do LDA on the TOKENS table, it was necessary to restructure the DOCS table in a similar way to the TOKENS from the word2vec part. After this, According to the Sci-Kit Learn documentation, Count Vectorizer converts text documents to token count matrices. This is then used in the count model engine. The LDA is calculated and can be used for the three important tables needed for analysis: theta, phi, and topics.

Phi: distribution of words over topics

Theta: distribution of topics over documents

Topics: distrubution of topics over paragraphs (as specified in code)

Phi

term_str topic_id	abandoned	abandonment	abdication	abdication king	abhorrence	ability	abode	abodes	abruptly	absence	...	yonder	youd	young	younger
T00	0.100005	0.100000	0.100005	0.100013	0.100001	0.100000	0.100016	1.099920	0.1	1.014437	...	0.100000	0.100001	0.100000	0.100030
T01	1.099991	0.100050	0.100012	0.100009	7.099939	0.100038	22.647189	0.100004	0.1	18.006883	...	0.100004	0.100000	5.861380	0.100036
T02	0.100000	0.100000	0.100000	0.100000	0.100000	0.100032	0.100014	4.099940	0.1	2.222671	...	3.922737	0.100000	0.100034	0.100000
T03	0.100008	0.100000	0.100000	0.100000	0.100008	1.099956	3.029299	0.100073	0.1	0.767979	...	0.100000	0.100000	0.100006	1.354997
T04	2.099996	1.336874	0.100000	0.100000	0.100014	3.099925	5.444600	1.099998	0.1	0.100015	...	0.152938	0.100002	1.338567	0.100000

Theta

book_id	chap_id	para_num	T00	T01	T02	T03	T04	T05	T06	T07	T08	T09
84	29	2	0.009091	0.475304	0.009093	0.451959	0.009093	0.009091	0.009091	0.009093	0.009093	0.009092
		3	0.001250	0.001250	0.001251	0.362882	0.001250	0.001250	0.001250	0.001251	0.001250	0.627115
		4	0.002779	0.385952	0.002779	0.002778	0.400907	0.002779	0.002779	0.193691	0.002779	0.002778
		5	0.005883	0.823040	0.129895	0.005883	0.005884	0.005883	0.005883	0.005883	0.005884	0.005883
		6	0.003847	0.003847	0.003848	0.497071	0.003847	0.472153	0.003847	0.003847	0.003847	0.003847

Topics

term_str topic_id	0	1	2	3	4	5	6	7	8	label	doc_weight_sum_brit	term_freq
T00	door	room	house	head	face	man	hand	window	moment	T00 door, room, house, head, face, man, hand, ...	411.595845	0.087919
T01	heart	life	love	mind	time	father	years	eyes	affection	T01 heart, life, love, mind, time, father, yea...	505.731137	0.182578
T02	nature	words	time	people	hand	heart	man	men	matter	T02 nature, words, time, people, hand, heart, ...	288.065053	0.068614
T03	sea	sun	wind	heart	day	earth	night	trees	waves	T03 sea, sun, wind, heart, day, earth, night, ...	316.159082	0.118245
T04	life	man	day	eyes	time	sight	moment	city	heart	T04 life, man, day, eyes, time, sight, moment,...	317.342492	0.116046

Sentiment Analysis

- Word Sentiment Table

term_str	n	tfidf_x	dfidf	syu_sentiment	weighted_sentiment
abandon	1.000000	0.023150	9.954560	-0.75	-0.017362
abandoned	2.500000	0.026620	20.353905	-0.50	-0.013310
abandonment	2.000000	0.022307	9.954560	-0.25	-0.005577
abhor	1.666667	0.054018	13.176952	-0.50	-0.027009
abhorred	1.500000	0.027870	23.818239	-1.00	-0.027870

Mean Sentiment

	title	origin	weighted_sentiment
0	Flatland: A Romance of Many Dimensions	British	0.000978
1	Frankenstein	British	0.000248
2	The Invisible Man	British	-0.003968
3	The Last Man	British	-0.000075

The Syuzhet CSV provided a long list of words and their associated sentiments that was combined with the VOCAB table, where tfidf was calculated. From here, two different sentiment tables were created that allowed for a very straightforward analysis. The word sentiment table provided sentiments per term, and the mean sentiment table provided averaged values per novel.

Principle Component Analysis

Loadings table

pc_id	PC0	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
term_str										
exposed	-0.011620	0.001462	-0.002703	0.005957	0.004643	0.016688	-0.004008	0.016406	-0.011850	0.021119
infinite	-0.011820	-0.007547	-0.013205	0.002865	-0.013922	0.015097	0.005726	0.021334	0.033177	0.013924
touched	-0.012181	0.000369	-0.009283	0.001932	-0.007678	0.006784	0.002135	0.007797	0.010295	0.005982
wont	-0.012240	-0.018905	0.016001	-0.010159	-0.001800	-0.010270	0.006991	-0.009915	-0.009040	-0.015620
shattered	-0.012365	0.001309	-0.000009	-0.019665	-0.007222	0.006969	0.009276	0.004640	-0.008718	-0.014219
...	...	...	...	...	...	...	...	...	...	...
hall	-0.126333	-0.227806	0.401170	0.346889	0.087894	-0.112281	0.113419	-0.104703	-0.084986	0.001761
our	-0.135608	0.116748	-0.015746	-0.014202	-0.171251	-0.198249	-0.050196	0.046549	0.004856	-0.017122
mr	-0.144359	-0.253745	0.339734	-0.285428	0.052235	-0.034354	0.079589	-0.070495	0.043730	-0.267471
she	-0.182353	0.054148	0.121564	0.163076	0.284902	-0.068885	-0.265929	0.228618	-0.137266	0.215218
kemp	-0.195471	-0.452738	-0.634795	0.092000	0.139752	-0.162108	0.021475	-0.095647	-0.173229	-0.037079

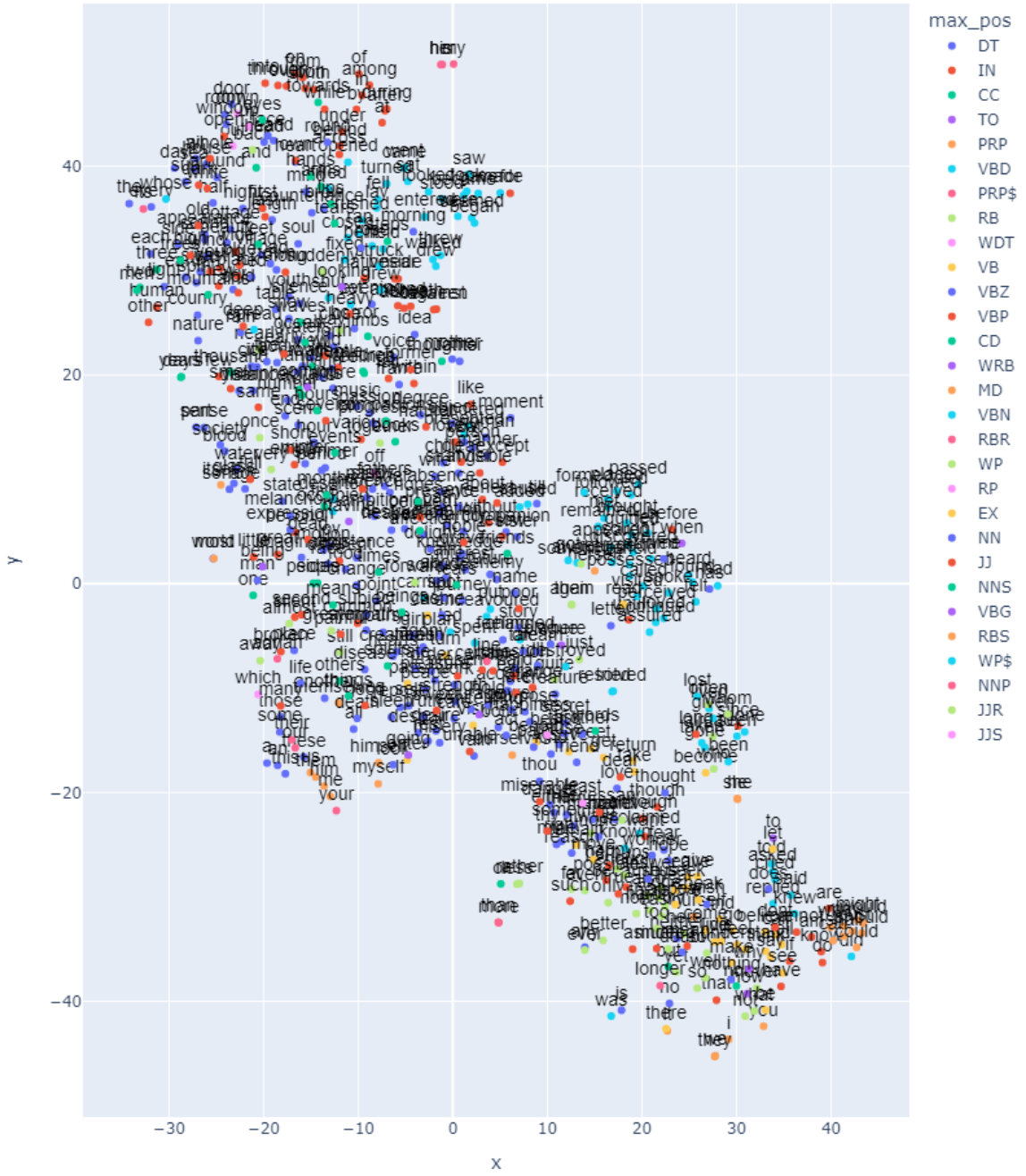
Components Table

pc_id	eig_val	doubt	soul	above	everything	forth	means	minutes	desire
PC0	0.200347	-0.023621	-0.045339	-0.025642	-0.021043	-0.026697	-0.032620	-0.018875	-0.028309
PC1	0.106165	-0.009401	0.034634	0.003735	-0.016714	0.006707	0.016509	-0.012757	0.022872
PC2	0.068671	-0.011295	-0.005652	-0.005926	0.013049	0.000841	-0.007838	0.003473	-0.002881
PC3	0.044279	-0.003717	-0.002010	-0.003765	0.005464	-0.000920	0.008204	-0.000383	0.000080
PC4	0.030105	-0.010208	-0.007023	-0.028981	-0.006214	0.005574	-0.011116	-0.002130	-0.015424
PC5	0.027574	0.000438	-0.044107	0.001277	0.012335	0.041600	0.020526	0.021184	-0.026391
PC6	0.024501	0.005815	0.007543	0.000162	0.000344	-0.013370	-0.053265	0.002020	0.000219
PC7	0.023399	0.020596	-0.031850	0.039526	0.005066	-0.005526	0.010276	0.013499	-0.041338
PC8	0.021461	0.007095	0.060603	-0.002203	0.014529	-0.003036	-0.049085	0.013130	0.011297
PC9	0.018854	-0.045560	-0.086640	0.006615	0.015642	0.023246	0.001696	0.011690	-0.001262

For PCA, I created standard functions to get the bag of words and to calculate tfidf. Using chapters for the bag of words and then calculating max for the tfidf allowed for creation of the loadings and components tables. These tables were then passed into a function that calculated pca, which was then plotted for further analysis on individual principle components.

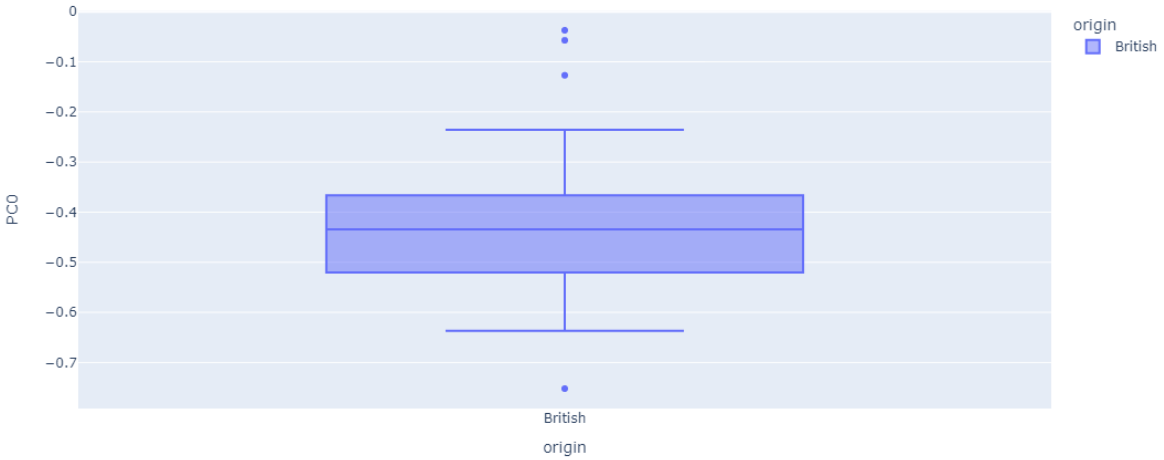
Exploration

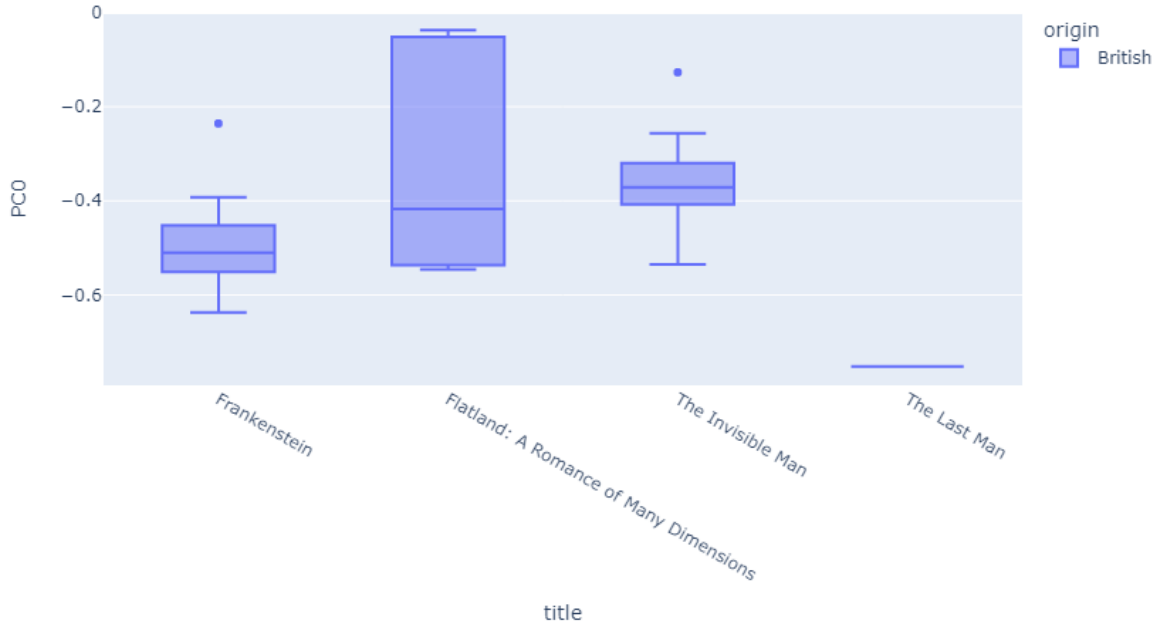
Scatterplot of Word2Vec



The word embeddings from the word2vec coordinate plot told me a lot about the relationships between different words as categorized by their part of speech. For example, determiners (modifies/describes nouns) were scattered throughout the plot, but modals (expressing attitudes) were focused in one area.

PCA Dispersion Plot

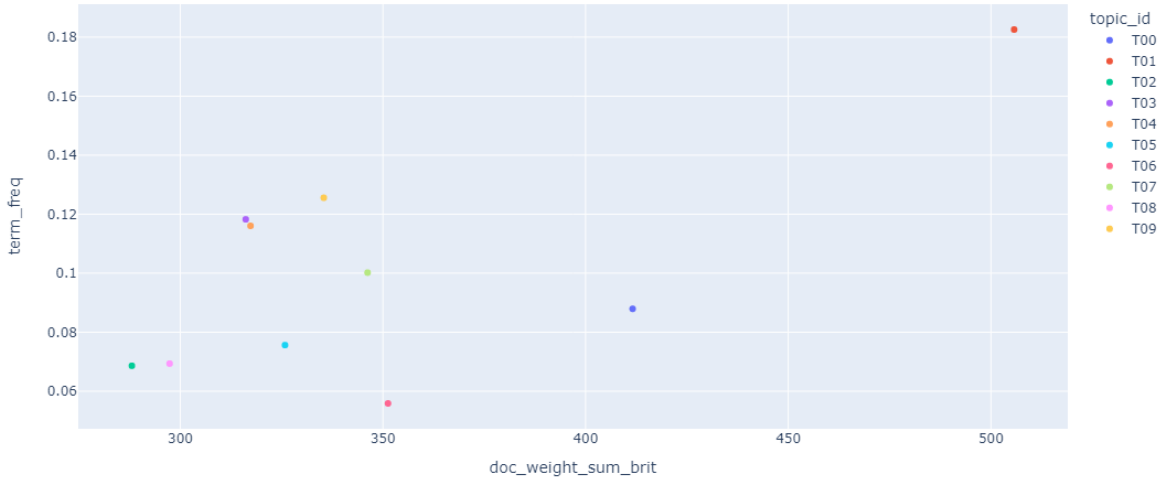

origin
British
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
PC0
British
origin


origin
British
0
−0.2
−0.4
−0.6
PC0
Frankenstein
Flatland: A Romance of Many Dimensions
The Invisible Man
The Last Man
title

The second plot of dispersion by novel was created to look at individual impacts of the principle component by book. Where Flatland had great dispersion, The Last Man lacked any. In the bar chart, it seems that only PC2 and PC8 had any positive correlations.

LDA Heatmap

book_id	chap_id	para_num	T00	T01	T02	T03	T04	T05	T06	T07	T08	T09
5230	19	36	0.050000	0.050000	0.050000	0.050001	0.050000	0.050006	0.549993	0.050000	0.050000	0.050000
	23	10	0.516701	0.007143	0.007144	0.007145	0.007144	0.007146	0.007145	0.007144	0.007144	0.426145
	27	10	0.100000	0.100000	0.100000	0.100000	0.100000	0.100000	0.100000	0.100000	0.100000	0.100000
201	4	24	0.020000	0.020001	0.020004	0.020000	0.020000	0.020002	0.020001	0.819986	0.020005	0.020001
18247	1	593	0.007694	0.007695	0.930756	0.007693	0.007694	0.007694	0.007694	0.007693	0.007694	0.007694
		470	0.050008	0.050000	0.050000	0.050000	0.050000	0.549992	0.050000	0.050000	0.050000	0.050000
		442	0.003030	0.003031	0.003031	0.003031	0.003031	0.003031	0.003030	0.972722	0.003031	0.003031
5230	44	81	0.939992	0.006667	0.006667	0.006668	0.006668	0.006668	0.006669	0.006667	0.006667	0.006667
		8	0.424675	0.010000	0.010001	0.010002	0.010001	0.010003	0.010000	0.010000	0.010001	0.495316
84	42	9	0.009095	0.517948	0.009095	0.009095	0.409306	0.009091	0.009092	0.009093	0.009092	0.009093
201	2	180	0.006250	0.006253	0.006251	0.369189	0.006251	0.006253	0.006252	0.580800	0.006251	0.006251
18247	1	991	0.004546	0.004547	0.004547	0.212774	0.214932	0.004546	0.004546	0.340852	0.004547	0.204164
84	48	21	0.005886	0.005885	0.005883	0.947047	0.005883	0.005883	0.005883	0.005883	0.005883	0.005883
5230	18	11	0.033345	0.033349	0.033342	0.033343	0.033364	0.699871	0.033358	0.033351	0.033343	0.033333
	44	31	0.550000	0.050000	0.050000	0.050000	0.050000	0.050000	0.050000	0.050000	0.050000	0.050000
18247	1	159	0.005266	0.952608	0.005264	0.005265	0.005268	0.005266	0.005267	0.005264	0.005267	0.005266
84	53	25	0.020000	0.020004	0.020002	0.020002	0.020002	0.020003	0.020004	0.819975	0.020003	0.020004
	56	81	0.007144	0.007145	0.007144	0.480613	0.462234	0.007145	0.007143	0.007143	0.007144	0.007145
5230	7	12	0.943747	0.006250	0.006250	0.006250	0.006251	0.006250	0.006250	0.006250	0.006250	0.006250
	1	24	0.006251	0.006251	0.442234	0.006251	0.120193	0.006251	0.263963	0.006252	0.006253	0.136100

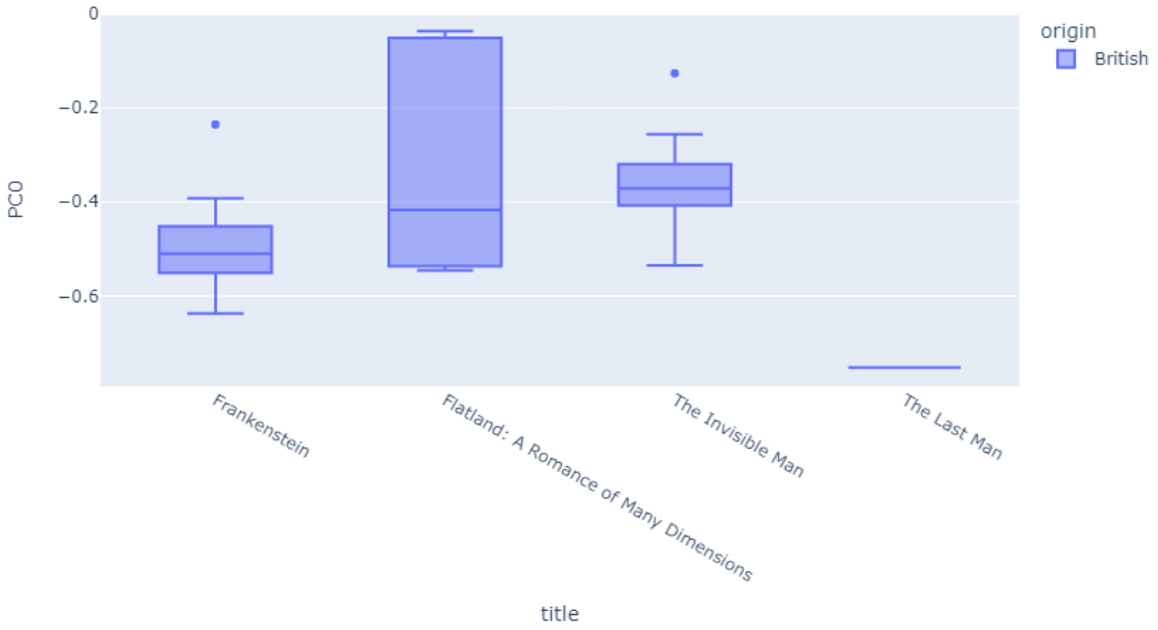


The heatmap associated with the novels indicate the significance of certain topics in each of the books- a metric that can be helpful in future analysis of social relecance of specific books. The scatterplot provides a look at term frequency against document weight throughout the topics, where most of them were around a certain area in the chart and a few topics were completely different.

Interpretation.

Provide your interpretation of the results of exploration, and any conclusion if you are comfortable making them.

	title	origin	weighted_sentiment
0	Flatland: A Romance of Many Dimensions	British	0.000978
1	Frankenstein	British	0.000248
2	The Invisible Man	British	-0.003968
3	The Last Man	British	-0.000075



From a purely visual perspective, the mean sentiment table provided explicit information on the overall tone of each of the novels. Some questions that arose: Why were the Invisible Man and the Last Man both negative sentiment? Upon further research, it seemed that The Invisible Man, by H.G. Wells, was banned from certain school districts for its vulgarity. This "vulgarity" is entirely subjective, the overall image of H.G. Wells being a purveyor of progressiveness, especially sexually. Additionally, The Last Man by Mary Shelley, an already well-known figure within the space of turbulent and tragic literature, is supported by the added dimension of the PCA plots and its lack of dispersion.

Upon further analysis of the tSNE plot, words like "sister" and "beloved" are grouped together, indicating a certain language towards social settings of the times.

Looking at the TOPICS table, the topics seem "softer"- does this indicate a distinction in the rhetoric of female writers within the science fiction space? This brings on additional questions on if Mary Shelley, one of the few female science fiction authors of the time, wasn't included in the corpus and the impact on the groupings of topics.

Interestingly, it doesn't seem like there is much consistency between looking at gender as a factor in affecting a novel's content and context within the time it was written. While there may be certain distinctions in things like sentiment, other metrics are inconsistent with proving that female authors write in a collectively similar way or extremely different from their equally famous, male counterparts.