Digital appendix 7. Parameters of 100-topic model, including justification for the representativeness of sampled text files

The 100-topic model explored in chapter six only includes one version of each story in the curated dataset (based on a shared common title). As investigated in chapter five, many titles appear more than once in the curated dataset, reflecting the importance of fiction reprinting, and fiction syndication specifically, in the nineteenth century: for Australian newspapers and for global print culture in general. However, when modeling is conducted on a corpus that includes multiple copies of the same story, the results tend to feature topics dominated by single titles. This is a reassuring outcome from one perspective, in that it shows the model grouping what *are* very similar documents (chapters from the same literary work, as it appeared in different newspapers). Indeed, this outcome suggests that topic modeling could assist in identifying different versions of the same literary work within mass-digitized collections. But my question was what are the similarities among – and differences between – different works, not what installments belong to the same work. For this reason excluding reprinted titles was a necessary first step in preparing my sample for analysis.

Even excluding reprinted fiction, modeling all the text files attached to unique titles would have overloaded the computing power available to me and resulted in imbalanced results due to the variable number of installments harvested for different titles. Both problems could have been addressed by analyzing a single text file for all unique titles, but that strategy would have produced topics that were perhaps only relevant to a particular installment of a story (for example, a title dominated by a topic relating to sea travel when only one chapter of the work concerned that journey).

Ultimately, I settled on modeling the first three text files attached to a title (and excluding titles without at least three text files attached) so as to encompass as wide a range as possible of unique titles in the curated dataset, while skewing the focus of analysis to the beginnings of works. These I considered more likely than a random selection of text files to contain a range of issues or topics subsequently explored by the work (because the start of stories tend to introduce their main themes). Restricting the sample to titles with at least three text files also excludes most of those completed in two newspaper issues, thus focusing analysis on the longer fiction in the curated dataset.

The resulting dataset encompasses 75% of unique titles, or 81% of titles in the curated dataset. The table below establishes the representative nature of the sampled titles with respect to the gender and nationality of authors and the type (metropolitan, provincial, or suburban). Some slight variability in the representativeness of the sample occurs because certain categories of title are more or less likely to appear only once in the curated dataset, thus affecting the likelihood of there being a version with at least three text files attached. For instance, titles by unknown authors (either those who were published anonymously or those whose identities are now lost to literary history) are more likely to appear only once in the curated dataset; as a consequence, fiction by authors of known genders is slightly overrepresented in the fiction sampled.

Although the sampled text files are representative with respect to metropolitan and provincial publication, the proportions of unique titles in the two sites are quite different to those analyzed in chapters four and five. There, the proportion of fiction published in provincial newspapers was significantly higher than in metropolitan ones. When only unique titles are considered these proportions are reversed because the fiction in provincial newspapers was much more likely to be reprinted.

Relationship between curated dataset and sampled text files

Category	**Subcategory**	**Curated dataset**				**Sampled text files**			
		All titles		**Unique titles**		**Represented titles**		**Unique titles**	
		#	%	#	%	#	%	#	%
Total	n/a	9263	**100**	6015	**100**	7541	**81**	4491	**75**
Period	Pre-1865	179	**2**	168	**3**	125	**2**	116	**3**
	1865–1869	242	**3**	214	**4**	157	**2**	136	**3**
	1870–1874	425	**5**	352	**6**	300	**4**	238	**5**
	1875–1879	668	**7**	552	**9**	476	**6**	383	**9**
	1880–1884	1435	**15**	908	**15**	1158	**15**	667	**15**
	1885–1889	1806	**19**	1169	**19**	1469	**19**	871	**19**
	1890–1894	2297	**25**	1314	**22**	1976	**26**	1031	**23**
	1895–1899	2211	**24**	1338	**22**	1880	**25**	1049	**23**
Author gender	Female	2020	**22**	1238	**21**	1792	**24**	1046	**23**
	Male	3774	**41**	2052	**34**	3300	**44**	1658	**37**
	Unknown	3465	**37**	2725	**45**	2445	**32**	1787	**40**
Author nationality	American	1318	**14**	660	**11**	1180	**16**	548	**12**
	Australian	1402	**15**	736	**12**	1227	**16**	580	**13**
	British	2946	**32**	1791	**30**	2608	**35**	1495	**33**
	Other	241	**3**	157	**3**	208	**3**	131	**3**
	Unknown	3356	**36**	2671	**44**	2318	**31**	1737	**39**
Newspaper type	Metropolitan	3795	**41**	2759	**46**	2939	**39**	1970	**44**
	Provincial	5262	**57**	2857	**47**	4420	**59**	2158	**48**
	Suburban	206	**2**	23	**>1**	182	**2**	11	**>1**
	Multiple	n/a	n/a	376	**6**	n/a	n/a	352	**8**