Curatr: A Platform for Semantic Analysis and Curation of Historical Literary Texts

Susan Leavy, Gerardine Meaney, Karen Wade, and Derek Greene

University College Dublin, Ireland
{susan.leavy,gerardine.meaney,karen.wade,derek.greene}@ucd.ie

Abstract. The increasing availability of digital collections of historical and contemporary literature presents a wealth of possibilities for new research in the humanities. The scale and diversity of such collections however, presents particular challenges in identifying and extracting relevant content. This paper presents *Curatr*, an online platform for the exploration and curation of literature with machine learning-supported semantic search, designed within the context of digital humanities scholarship. The platform provides a text mining workflow that combines neural word embeddings with expert domain knowledge to enable the generation of thematic lexicons, allowing researches to curate relevant sub-corpora from a large corpus of 18th and 19th century digitised texts.

Keywords: Text mining · Digital humanities · Corpus curation

1 Introduction

The interpretability of the algorithmic process and the incorporation of domain knowledge are essential to the use of machine learning and text mining in the semantic analysis of literature. The absence of these factors can inhibit adoption of machine learning approaches to text mining in the humanities, due to issues of accuracy and trust in what is often regarded as a 'black-box' process [6, 13, 15]. This paper presents *Curatr*, an online platform that incorporates domain expertise and imparts transparency in the use of machine learning for literary analysis. The system supports a corpus curation workflow that addresses the requirements of scholars in the humanities who are increasingly working with large collections of unstructured text and facilitates the development of sub-corpora from large digital collections.

Selection, curation and interpretation is central to knowledge generation in the humanities [18, 37]. This process is supported in *Curatr* with conceptual search functionality that uses neural word embeddings to building conceptual lexicons specific to a given theme or topic. These thematic lexicons can then be used to mine relevant texts to form curated literary collections that may be saved, further modified, or exported as sub-corpora. The platform was developed based on a collection of 35,918 English language digital texts from the British Library[1].

[1] British Library Labs: https://www.bl.uk/projects/british-library-labs

An evaluation of *Curatr* was conducted in conjunction with an associated project examining the relationship between societal views of migration, ethnicity, and concepts concerning contagion and disease in 19th century Britain and Ireland. In order to explore the cultural representation of migrants, the study focused on their representation within historical fiction which comprises 16,426 texts of the British Library digital collection. Given the largest communities of migrants to London during the late 19th century were Irish and Jewish, this study focused on their portrayal in relation to prevailing concepts of contagion, disease and migration. Lexicons related to these themes were generated through recommendations derived from word embedding models and text were retrieved based on relevance of the texts. The findings were evaluated in terms of the overall requirements of a humanities scholar along with the relevance of the texts uncovered. We describe this case study in more detail in Section 4.

2 Related Work

2.1 Text Mining in the Humanities

The work builds on a range of literature that demonstrates requirements for digital humanities platforms. Close reading functionality to provide context is an essential aspect of humanities research, as evidenced in the provision of close reading functionality along with quantitative analysis in systems developed by Hinricks et al. [16] and Vane [35]. Domain knowledge was combined with automated text classification to provide more accurate retrieval results in a system developed by Sweetnam and Fennel [1] to explore early-modern English texts. A system based on semantic search was demonstrated by Kopaczyk et al [20] for the analysis of Scottish legal documents from the 16th century. Other systems, such as that proposed by Jockers [18], have focused primarily on the use of machine learning methods for text analysis.

In a study of the update of machine learning in industry, Chiticariu et al. [6] noted a gap in the volume of academic research on machine learning, compared with lower levels of uptake within industry and found the causes of this pertained to training data, interpretability and incorporation of domain knowledge. Similarly, the relatively low uptake of machine learning methods in the digital humanities has been attributed to issues pertaining to interpretation and trust [34, 13, 15]. Imparting domain knowledge into the process of text analysis through interpretation and annotation is also central to humanities research [17, 37]. The *Curatr* platform addresses these specific requirements of digital humanities research by incorporating domain knowledge and transparency within a text mining workflow.

2.2 Concept Modelling with Word Embeddings

Word embedding refers to a family of methods from natural language processing that involve mapping words or phrases appearing in large text corpora to dense,

low-dimensional numeric representations. Typically, each unique word in the corpus vocabulary will be represented by its own vector. By transforming textual data in this way, we can use the new representation to capture the semantic similarity between pairs or groups of words. Word embedding methods have been used in digital humanities research to generate semantic lexicons for a range of purposes including detecting language change over time [14], extracting social networks from literary texts [36], sentiment analysis [32], and semantic annotation [21]. An interactive strategy whereby a user incrementally creates a lexicon based on recommendations for similar words as recommended by a word embedding model has been demonstrated in a number of works [10, 27].

A variety of different approaches have been proposed in the literature to construct embeddings. The word embedding algorithm used in this research is *word2Vec* [22], which generates distributed representations of words that can be used to interpret their meaning. This approach captures the concept from distributed semantics that the meaning of a word "can be determined by the company it keeps" [11]. Word co-occurrence is identified over an entire corpus and each word along with the words found beside it in the text are represented by a vector. The similarity of terms can then be derived based on whether they are used alongside similar words, or in a similar context. This approach to generating lexicons has been shown to be useful where the language of a particular corpus is highly specific [5] and where existing general-purpose lexicons are not appropriate. This approach is therefore highly relevant to digital humanities where language is often specific to a particular research project or domain.

It has also been pointed out that pre-processing and representation approaches in automated processes can have particular significance within a digital humanities context [4, 12]. These decisions are a crucial aspect of the evaluation of results within humanities scholarship. Given that using word embedding in text analysis has been critiqued for its lack of transparency [31], it is crucial within a digital humanities context to impart transparency into the text mining workflow.

3 Curatr Design

3.1 Platform Overview

Curatr implements a text mining framework involving conceptual search using word embeddings to dynamically build a semantic lexicon specific to a given literary corpus. The humanities researcher begins with seed terms and these are expanded using neural word embedding through an interactive online interface to produce semantic lexicons. The *Curatr* system also provides for keyword search, filtering based on metadata, ngram frequencies, and categorisation based on the original British Library topical classifications (Fig. 3). The information retrieval component of the system involves the indexing of texts and using the open source Apache Solr engine[2].

[2] `http://lucene.apache.org/solr/`

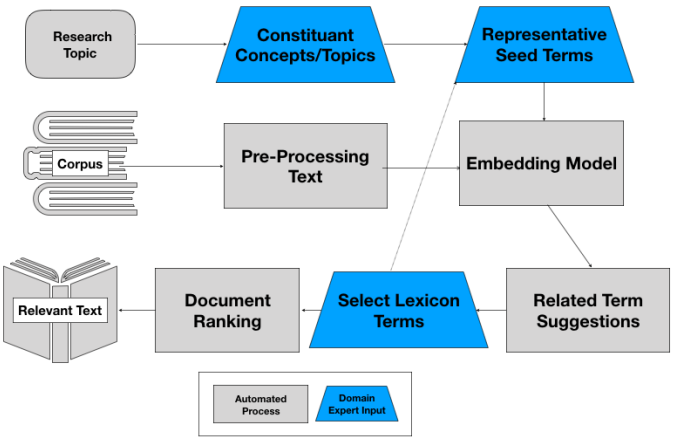

Fig. 1: *Curatr* workflow for digital humanities text mining.

3.2 Concept Lexicon Generation

The conceptual search workflow outlined in Fig. 1 enables the compilation of seed terms associated with a given concept or topic by the humanities researcher. *Curatr* allows for the expansion of these terms to form a lexicon of semantically similar words based on associations suggested from the querying of a neural word embedding model developed from the corpus. Word embedding models were generated from the complete English language corpus using the *word2vec* approach [22] yielding real-valued, low-dimensional representations of words based on lexical co-occurrences. The use of word embeddings rather than more complex language models were deemed appropriate due to the lack of structure in the text and OCR errors introduced through the process of digitisation. The specific embedding variant used in this work is a 100-dimensional Continuous Bag-Of-Words (CBOW) *word2vec* model, trained on the full-text volumes of the corpus. To address the levels of transparency required in digital humanities scholarship regarding approaches to text representation and parameter settings and their potential effects on results, decisions regarding text processing options and parameter strategies in generating the neural word embedding models are available to the user.

Based on the embedding model, the top 20 words found to be similar to the seed words are recommended to expand the current lexicon. The user selects the subset of recommendations to be included in the conceptual lexicon. This inclusion of a "human in the loop" ensures that the process of generating a lexicon is informed by domain knowledge of the user. Multiple iterations of this semantic search allow the researcher to refine the lexicon and augment the conceptual category on each iteration, simulating the process of knowledge generation from close reading and annotation.

Title:	Bess
Volume:	Volume 1 of 1
Author(s):	Helen M. Boulton
Published:	1896 – London (Osgood, McIlvaine & Co.)
Edition:	N/A
Classification:	Fiction — English 3 Volume Novels
Physical Description:	334 pages
Shelfmark:	HMNTS 012626.g.17.
Mudie's Library Author:	No matching author
Related Links:	Find Similar Volumes — British Library Catalogue

BESS

BESS BY HELEN M. BOULTON AUTHOR OF "JOSEPHINE CREWE " " Wrong is not only different from Right, but it is in strict scientific terms infinitely different ; even as the gaining of the whole world set against the losing of one's own soul." Carlyle. LONDON OSGOOD, McILVAINE & COMPANY 45 ALBEMARLE STREET, W. 1896 [All rights reserved]

BESS. CHAPTER I. " The childhood shows the man As morning shows the day." ? Milton. THE village of Torcliff lay sleeping in the hot June sunshine, its sinuous length stretched by the river side like a creature basking in the heat. The sun glared down on the dusty road and on the buff cottages that lined it irregularly on either hand ; the heat danced against the dark pine-woods covering the slopes of the steep hills that closed the village in, and the light was hard and dazzling on

Fig. 2: The *Curatr* close reading interface.

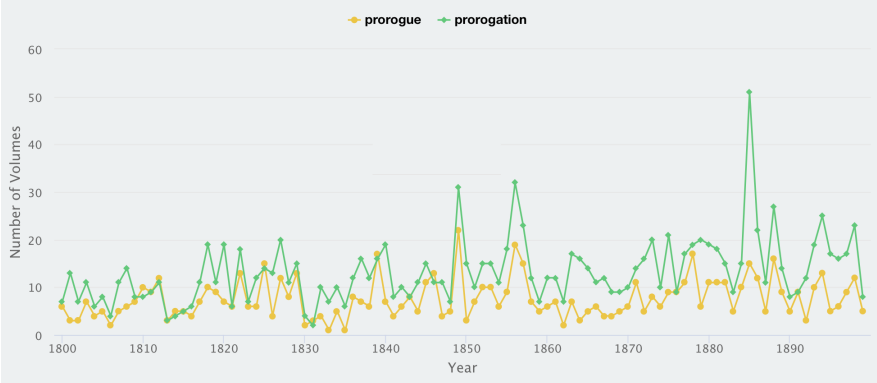

Fig. 3: Sample *Curatr* n-gram frequency analysis.

3.3 Corpus Curation

The finalised semantic lexicons are used as a basis for ranked volume retrieval from the indexed library corpus. Texts are ranked according to the frequency of occurrence of terms from each generated lexicon relative to the document length. This process uncovers documents that pertain to the given conceptual category. These document rankings may then be saved by the user for later use.

Curatr also facilitates the amendment of the sub-corpus to remove non-relevant texts. Based on the results of this search, the user may revise the contents of the conceptual lexicon and rerun the text mining process. Along with the provision of a close reading interface for examining individual texts (Fig. 2), the platform also enables the export of curated sub-corpora to download for further analysis.

4 Case Study Evaluation

The case study involved an exploratory investigation of historical cultural attitudes towards migration in Britain and their associations with concepts of contagion and disease [26, 19]. Poverty-induced migration from Ireland to Britain during the Great Famine (1845-1851) is cited as generating a fear of transmission of contagious disease [23]. The association of fear of contagion with historical sentiments towards migrants, according to Samuel K. Cohn however, requires more systematic study and more nuanced interpretation of both historical and cultural archives [8]. The aim of this exploratory study is to uncover new texts and literary representations that are less well known to researchers and which might challenge current understandings of the relationship between migration and concepts of contagion and disease.

Curatr was used to uncover texts that addressed concepts pertaining to the dynamics of bio-politics and migration in Britain from the set of 16,426 works of fiction in the corpus. Each step of the text mining workflow is evaluated in the context of the value gained from the process in terms of the exploratory phase of corpus analysis for humanities research. The use of machine learning to support the text mining process is evaluated on the basis of the value the approach brings to the iterative theory-building research process of the humanities.

Measuring the usefulness of documents retrieved in the context of humanities research can be complex and does not always fit with standard information retrieval precision metrics. Particularly in the exploratory phase of research, the topics under study themselves constantly transform and thus, negate the potential to establish a ground truth. This is particularly pertinent in the preliminary exploratory phase of humanities research, where new and diverse texts are sought by the researcher and their relevance to the research topic is often not immediately apparent [7]. The relevance of a text as a stable, independent, and consistent attribute does not always apply within the context of digital humanities where relevance is individual to each researcher and their evolving research themes. The concept of relevance may also be different in the exploratory stage of research than at other stages [30]. The value of text mining in the humanities is linked with its ability to uncover novel and diverse texts that have the potential to challenge prevailing theories on a given topic [7]. The framework utilised for evaluation in this case study therefore draws on Bates et al. [3], who examined what constitutes relevance of a document retrieved through a text mining system in the context of humanities research.

Bates et al. [3] defined two aspects of relevance in humanities research, *content relevance* and *utility relevance*. Content relevance refers to whether the document that was retrieved reflects the terms in the search query. Utility relevance refers to whether the texts retrieved are of value in terms of the research topic of the scholar. Particularly in the exploratory phase of humanities research, the researcher requires texts that address the research topic. Texts which are deemed to be most relevant in terms of utility are often not those with the highest content relevance, but are those that alter the judgement and challenge existing

theories of the researcher. Document familiarity is therefore an attribute that 'may be presumed to swamp all other considerations' [3, p. 702].

Categories of unfamiliarity of a text to a humanities research were outlined by Barry et al. [2]. This framework outlines how a text retrieved may be novel in terms of content, source or stimulus. *Content novelty* describes how the content of a text itself provides new knowledge to a researcher. *Source novelty* refers to whether the text is written by a previously unknown author or publisher. *Stimulus novelty* indicates whether the text itself is new to the researcher. This framework captures the complexity of the judgements of value and relevance of a text to a humanities researcher and is drawn upon in the evaluation of the *Curatr* system in this paper.

4.1 Conceptual Lexicons

Seed terms were identified that represented the key thematic strands of the case study. The themes pertained to 'migration', 'illness', 'contagion', and Irish and Jewish ethnic identities. The word embedding model was queried to uncover words used in a similar context to the seed terms in the fiction corpus. The results were selectively added based on contextual knowledge and interpretation on the part of the humanities researcher through an interactive lexicon building interface. The resulting expanded semantic lexicons were then used as a basis for ranking the documents to uncover texts that capture the key themes of the study. See Table 1 for examples of lexicons generated in this way.

Analysis of the recommended terms suggested associations in the fiction corpus that were not immediately apparent from an initial close reading. For example, terms pertaining to the theme of civil unrest *(rebels, dynamiters, conspirators, incendiarism, anarchistic, nihilists, revolutionary, sedition, informers, radicals*) make up a substantial portion of the lexicon pertaining to 'ethnic identity'. As a result of these suggestions, aligning with the iterative theory-building approach of research in the humanities, the suggested words enriched the original conceptualisation of the research theme of ethnic identity to incorporate relationships with political ideology. Additionally, the expanded lexicons indicated points where conceptual ontologies overlap.

Lexicon	**Seed Terms**	**Recommended Words**
Ethnic identity	irish, fenian, papist, jewish, jew	jews, fenianism, hibernian, usurer, celtic, rebels, invincibles, whiteboys, incendiary, brogue, chartists, irishman, catholics, ringleaders, rabbis
Migration	immigrant, alien, interloper, migrant	civilisers, doavn trodden, self governing, circumcised, peoples, separatists, cousinhood, usurpers, interloper, aliens, intruder, middlemen, ryots, kinless, alien
Contagion	infect, epidemic, inoculate, contagion, contaminate, vaccinate	infection, contagion, infectious, infected, fever, contagious, epidemic, plague, epidemic, disease, epidemics, fever, malarial, endemic, malady
Disease	disease, smallpox, cholera, fever, pestilence	scarlet fever, epidemics, morbus, cancers, tumour, incurable, malaria, sickness, cancer, brain, diseases, distempers, typhus, anaemia, heart disease

Table 1: Examples of top recommended words for different conceptual lexicons.

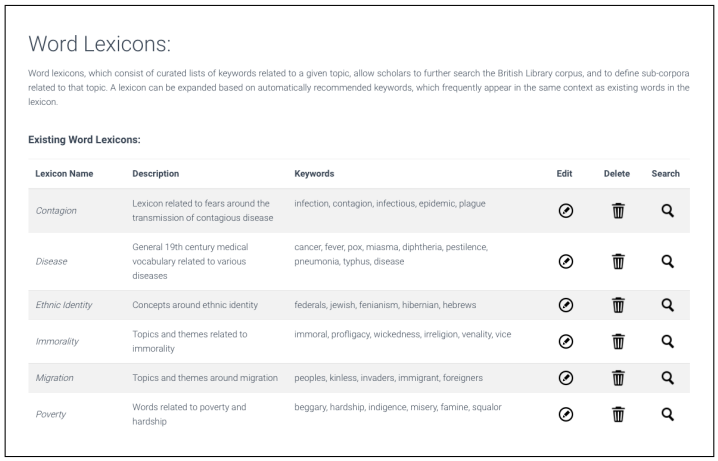

Fig. 4: The *Curatr* lexicon management interface.

The suggested related terms in *Curatr* also provided indications of relevant but unfamiliar or archaic terminology. For instance, in this study the term 'distemper', which is more commonly applied to animal illnesses nowadays, appears as a term that is relevant to the concept of contagion within these works. The method was also effective in dealing with OCR (optical character recognition) errors as well as spelling errors and variants; variations on relevant terms (such as the misspelling of fever as 'fea er'), which would not be identified by a simple search, appeared in their correct contexts. Addressing issues related to OCR quality is a recurring issue in digital humanities projects, and this research demonstrates how neural word embedding can render OCR errors less problematic.

4.2 Discussion of Results

Drawing upon the framework of retrieval value outlined by Bates et al. [3] and Barry et al. [2], the top 10 retrieved texts in each category were analysed in terms of novelty to the humanities researcher involved in the project. The qualitative difference in text retrieved before and after the query expansion phase was also examined. Texts in the corpus were ranked according to the frequency by which the curated conceptual lexicons were mentioned, relative to the length of the documents (Fig. 4). The iterative process whereby the researcher returns to edit the conceptualisation of the key thematic trends of the research topic through editing the semantic lexicons aligns with the process of grounded theory research and necessitated a close reading of the results by humanities researchers to evaluate the value of the retrieved documents.

Source novelty. A striking pattern among the texts retrieved was the relative obscurity of some of the authors of the texts retrieved using expanded query

Prior to Query Expansion	Post Query Expansion
1894 The Captain's Youngest, Frances H. Burnett	1891 The Year of Miracle, Fergus Hume
1888 The Devil's Die, Allen Grant	1897 The Sign of the Red Cross, Evelyn E. Green
1894 The Azrael of Anarchy, Gustave Linbach	1855 Old Saint Paul's, William H. Ainsworth
1891 The Year of Miracle, Fergus Hume	1894 The Azrael of Anarchy, Gustave Linbach
1870 Unawares, Frances Mary Peard	1847 A Tale of the Irish Famine, Unknown
1893 Doctor, or Lover?, Faber Vance	1885 The Legend of Samandal, James Fer
1888 A Crown of Shame, Florence Marryat	1855 The Wood-Spirit, Ernest C. Jones
1875 Ashes to ashes, Hugh R. Haweis	1888 The Devil's Die, Grant Allen
1865 Not Proven, Christina B Cameron	1898 Vanya, Orlova Olga
1892 The Medicine Lady, Elizabeth T Meade	1897 A Literary Gent, J.C. Kernahan

Table 2: Texts retrieved before and after query expansion, for concepts related to 'illness' and 'contagion'.

Prior to Query Expansion	Post Query Expansion
1898 For Lilias, Rosa Nouchette Carey	1881 Gifts and Favours Doctor Olloed, Unknown
1875 The Golden Shaft, George C. Davies	1871 Ierne, William R. Trench
1876 The Youth of the Period, James F.S. Kennedy	1863 Sackville Chase, Charles J. Collins
1876 Her Dearest Foe, Alexander	1894 Ivanda or the Pilgrim's quest, Claude A. Bray
1886 A Modern Telemachus, Charlotte M. Yonge	1894 Doctor Izard, Anna K. Green
1887 Major Lawrence, Emily Lawless	1865 The Crusader or the Witch of Finchley, Unknown
1857 Guy Fawkes, William H. Ainsworth	1864 The Bee-Hunters, Gustave Aimard
1846 The Moor, the Mine and the Forest, William Heatherbred	1850 Helen Porter, Thomas P. Prest
1865 The Notting Hill Mystery, Charles Felix	1852 Idone, James H. L. Archer
1897 Owen Tanat, Alfred N. Palmer	1886 A Mysterious Trust, Edmund Mitchell

Table 3: Texts retrieved before and after query expansion, for a concept related to 'migration'.

terms. In analysing the top ten texts retrieved using the seed terms alone, five of the migration category and four of illness category were immediately familiar to the researchers (Tables 2 and 3). However, after the list was expanded based on suggestions from neural word embedding, this figure reduced to three for the concept of illness, and none were previously well known in relation to the concept of migration.

A similar trend is evident in the texts retrieved using the semantic lexicons of ethnic identity. Only one of the authors in the top ten texts related to Irish identity were widely known, while three of those relating to Jewish identity were (see Fig. 4). The texts retrieved based on the semantic lexicon of Irish ethnicity contained none by widely read Irish authors, although Ulick Ralph Burke and Albert Stratford George Canning do merit an entry in the database of Irish Literature[3]. A work by the prolific English author Fergus Hume was retrieved, as were works by the British authors Mabel Collins and Edith Cuthell, the latter novel being its author's only work based in an Irish setting. This identification of a number of writers less prominent in the canon, but who wrote on topics relating to the key themes of the case study, demonstrate how the expansion of

[3] Ricorso: Database of Irish writers `http://www.ricorso.net`

Irish Lexicon
1897 Sweet Irish Eyes, Edith E. Cuthell
1893 The Harlequin Opal, Fergus Hume
1875 The Autobiography of a Man-o-War's Bell, Charles R. Low
1883 The Wild Rose of Lough Gill, Patrick G. Smyth
1880 Loyal and Lawless, Ulick R. Burke
1867 Baldearg O'Donnell, Albert A. G. Canning
1893 The Great War of 189-. A forecast, F. Villiers
1896 The Idyll of the White Lotus, Mabel Collins
1886 Our Radicals. A tale of love and politics, Fred Burnaby
1891 The Last Great Naval War, A. N. Seaforth
1890 Heir and no Heir Canning, Albert S. G. Canning

Jewish Lexicon
1890 The Prophet. A parable, Thomas H. H. Caine
1897 A Rogue's Conscience, David C. Murray
1864 The Hekim Bashi, Humphry Sandwith
1874 Jessie Trim, Benjamin L. Farjeon
1853 The Turk and the Hebrew, Unknown
1865 The Crusader or the Witch of Finchley, Unknown
1895 Maid Marian and Crotchet Castle, Thomas L. Peacock
1898 Dreamers of the Ghetto, Israel Zangwill
1838 Oliver Twist, Charles Dickens
1869 Count Teleki: A story of modern Jewish life, UN

Table 4: Texts retrieved for the 'Irish' and 'Jewish' conceptual lexicons.

query terms in this context has the potential to uncover relevant texts which might otherwise be overlooked.

Stimulus novelty. In this case study the texts retrieved presented a number of works that represented not only less well-known authors of the 18th and 19th centuries, but also less well-known texts written by more prominent authors. In relation to the texts retrieved using the conceptual lexicon of illness for instance, the majority of retrieved titles were not commonly known and none pertaining to Irish identity attracted significant bibliographic or critical commentary. The text marked as most relevant to Irish identity, *Sweet Irish Eyes*, was written by an English author and was their only Irish-themed work.

The most relevant of the retrieved set of texts pertaining to Jewish identity was that by British novelist Hall Caine. This work is referred to in studies of the Gothic (e.g. Mulvany-Roberts [24]), but there is very little reference to this novel and its portrayal of Jewish characters. It is notable that Anglo-Jewish writers (Zangwill and Farjeon[4]) feature prominently alongside one of Dickens' best known novels. Charles Dickens' *Oliver Twist* and Israel Zangwil's *Dreamers of the Ghetto*, which appear on the list of novels pertaining to Jewishness, were the only two novels on the list that have previously attracted sustained scholarly attention. For example, Udelson [33], Rochelson [28, 29] and Murray [25] have all written on Zangwill, though he is far less well known to the public than Dickens.

Using the expanded list of query terms to retrieve texts also uncovered slightly older texts. For instance the top 10 pertaining to topics of illness and dis-

[4] Farjeon features more prominently in histories of Australian literature as he emigrated there.

ease were on average five years older. This points to the value of word embedding in uncovering historical linguistic associations and thereby compiling a semantic lexicon that is less biased towards current linguistic norms. The text retrieval process also highlights the influence of historical change in publication practices. Comparing text ranking methods using raw frequency counts or conceptual lexicons with frequencies in relation to document length demonstrates how these two approaches illuminate different aspects of the corpus. Relative frequency locates texts that would otherwise be almost impossible to find, as evidenced by their having attracted little or no critical commentary. It works better for the latter part of the century when publication, distribution and creative practices favoured shorter texts and prevents the findings being swamped by multi-volume and very long form serial fiction. Raw frequency however highlights references to concepts across a broad range of extremely long mid-century texts and makes it possible to work with more long novels and long running serials.

Content novelty. While uncovering authors and titles that were previously less familiar to a researcher is beneficial, its value is ultimately dependent on the relevance of the content of the books in relation to the research topic. Evaluation of relevance in this study was assessed through the extent to which the content of the books retrieved offered new insights and challenged existing theories in relation to the research topics. Close analysis showed that the expanded set of query terms retrieved documents that were less specific to the topics than those retrieved using the manually generated set of seed terms. This process increased the novelty of the text retrieved thorough the suggestion of new associations and challenged the researcher's original definition of conceptual categories. For instance, in relation to the concept of illness, a text pertaining to the Irish Famine and texts addressing political issues were returned. This challenged the researcher's original concept of illness and broadened it to consider historical associations with political ideology and historical events. The text identified as most relevant to the concept of illness using an expanded lexicon was Evelyn Everett Green's novel, *The Sign of the Red Cross*. The content of this novel describes details of a plague in London and was critiqued at the time for its graphic nature [9]. The novel *Old Saint Paul's* was also concerned with the plague in London and provides a wealth of information relevant to the topic. While these novels are not among the most well known currently, their graphic account of plague in London provides a wealth of insight regarding the historical conceptualisation of illness in Britain.

Overall the content of texts retrieved pertaining to Irish and Jewish ethnic identities were highly relevant to the topic and uncovered texts that are less well known, thus presenting new opportunities to examine cultural representations of ethnicity. It also suggests new and unexpected associations for further investigation. For instance, the theme of seafaring in the titles pertaining to Irish identity is an association that was new to the researchers. The content that was returned in relation to the concept of migration returned more titles that addressed mi-

gration in the broader sense in connection with international political events, in contrast with the narrower focus of the results retrieved using seed terms alone.

The findings uncovered by the *Curatr* system demonstrate that domain-specific conceptual modelling with neural word embedding is effective in uncovering texts that capture given concepts. The incorporation of domain expertise within the workflow uncovers texts relevant to the research topic and also meets the requirements of interpretability and inclusion of domain knowledge that are vital to humanities scholarship. Evaluation was conducted via a case study examining understandings of migration, ethnic identity and concepts pertaining to disease and contagion in 19th-century Britain. The findings showed that based on a minimal list of seed terms, new and unexpected documents were uncovered that addressed the concepts under investigation.

5 Conclusion

This research presents an approach whereby machine learning and text analysis is used within a text mining workflow designed for digital humanities research. The corpus curation workflow in *Curatr* is supported by neural word embedding and through an interactive online platform, ensures transparency and incorporates domain knowledge. The approach demonstrates how machine learning techniques may be used to enhance the curation process for humanities scholars. Evaluation of the system found that texts retrieved using the *Curatr* text mining workflow were particularly useful in the context of exploratory humanities research. The system uncovered works that had not previously attracted sustained scholarly attention, and which presented opportunities to uncover new insights in relation to the given research topic. In future work the addition of new sources to the platform would further enrich this process. Through the exploratory text mining workflow supported by *Curatr*, humanities scholars are enabled to challenge prevailing methods of canonisation of historical fiction.

Acknowledgements. This research was partially supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

1. Bailey, E., Lawless, S., O'Connor, A., Sweetnam, S., Conlan, O., Hampson, C., Wade, V.: Cultura: supporting enhanced exploration of cultural archives through personalisation. In: the Proceedings of the 2nd International Conference on Humanities, Society and Culture, ICHSC. ICHSC (2012)
2. Barry, C.L.: User-defined relevance criteria: An exploratory study. Journal of the American Society for Information Science **45**(3), 149–159 (1994)
3. Bates, M.J.: The getty end-user online searching project in the humanities: Report no. 6: Overview and conclusions. College & Research Libraries **57**(6), 514–523 (1996)

4. Camacho-Collados, J., Pilehvar, M.T.: On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. arXiv preprint arXiv:1707.01780 (2017)
5. Chanen, A.: Deep learning for extracting word-level meaning from safety report narratives. In: Integrated Communications Navigation and Surveillance (ICNS), 2016. pp. 5D2–1. IEEE (2016)
6. Chiticariu, L., Li, Y., Reiss, F.R.: Rule-based information extraction is dead! long live rule-based information extraction systems! In: Proceedings of the 2013 conference on empirical methods in natural language processing. pp. 827–832 (2013)
7. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 659–666. ACM (2008)
8. Cohn, S.K.: Pandemics: waves of disease, waves of hate from the plague of athens to aids. Historical Research **85**(230), 535–555 (2012)
9. Dempster, J.A.: Thomas nelson and sons in the late nineteenth century: A study in motivation. part 1. Publishing History **13**, 41 (1983)
10. Fast, E., Chen, B., Bernstein, M.S.: Empath: Understanding topic signals in large-scale text. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. pp. 4647–4657. ACM (2016)
11. Firth, J.R.: A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis (1957)
12. Flanders, J., Jannidis, F.: The Shape of Data in Digital Humanities: Modeling Texts and Text-based Resources. Routledge (2018)
13. Frank, A., Bögel, T., Hellwig, O., Reiter, N.: Semantic annotation for the digital humanities. Linguistic Issues in Language Technology **7**(1), 1–21 (2012)
14. Hamilton, W.L., Clark, K., Leskovec, J., Jurafsky, D.: Inducing domain-specific sentiment lexicons from unlabeled corpora. In: Proc. EMNLP 2016. vol. 2016, p. 595. NIH Public Access (2016)
15. Hampson, C., Munnelly, G., Bailey, E., Lawless, S., Conlan, O.: Improving user control and transparency in the digital humanities. In: 2013 International Conference on Culture and Computing (Culture Computing). pp. 196–197. IEEE (2013)
16. Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E., Coates, C.M.: Trading consequences: A case study of combining text mining and visualization to facilitate document exploration. Digital Scholarship in the Humanities **30**(suppl_1), i50–i75 (2015)
17. Jackson, H.J.: Marginalia: Readers writing in books. Yale University Press (2002)
18. Jockers, M.: Detecting and characterizing national style in the 19th century novel. In: Digital Humanities 2011, Stanford, CA (2011)
19. Kinealy, C.: This Great Calamity: The Great Irish Famine: The Irish Famine 1845-52. Gill & Macmillan Ltd (2006)
20. Kopaczyk, J.: The Legal Language of Scottish Burghs: Standardization and Lexical Bundles (1380-1560). Oxford University Press (2013)
21. Leavy, S., Pine, E., Keane, M.T.: Industrial memories: Exploring the findings of government inquiries with neural word embedding and machine learning. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018 (2018)
22. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
23. Morash, C.: The Hungry Voice: The Poetry of the Irish Famine. Irish Academic Press (2009)

24. Mulvey-Roberts, M.: The Handbook of the Gothic. Springer (2016)
25. Murray, J.: The social enterprise law market. Md. L. Rev. **75**, 541 (2015)
26. Nelkin, D., Gilman, S.L.: Placing blame for devastating disease. Social Research pp. 361–378 (1988)
27. Park, D., Kim, S., Lee, J., Choo, J., Diakopoulos, N., Elmqvist, N.: Conceptvector: text visual analytics via interactive lexicon building using word embedding. IEEE transactions on visualization and computer graphics **24**(1), 361–370 (2018)
28. Rochelson, M.J.: they that walk in darkness: Ghetto tragedies: The uses of christianity in israel zangwills fiction. Victorian Literature and Culture **27**(1), 219–233 (1999)
29. Rochelson, M.J.: A Jew in the Public Arena: The Career of Israel Zangwill. Wayne State University Press (2010)
30. Spink, A., Greisdorf, H., Bateman, J.: From highly relevant to not relevant: Examining different regions of relevance1. Information Processing & Management **34**(5), 599–621 (1998)
31. Subramanian, A., Pruthi, D., Jhamtani, H., Berg-Kirkpatrick, T., Hovy, E.: Spine: Sparse interpretable neural embeddings. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
32. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Learning sentiment-specific word embedding for twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1555–1565 (2014)
33. Udelson, J.H.: Dreamer of the Ghetto: The Life and Works of Israel Zangwill. University of Alabama Press (1990)
34. Van Cranenburgh, A., van Dalen-Oskam, K., van Zundert, J.: Vector space explorations of literary language. Language Resources and Evaluation (Feb 2019)
35. Vane, O.: Text visualisation tool for exploring digitised historical documents. In: Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility. pp. 153–158. ACM (2018)
36. Wohlgenannt, G., Chernyak, E., Ilvovsky, D.: Extracting social networks from literary text with word embedding tools. In: Proc. Workshop on Language Technology Resources and Tools for Digital Humanities. pp. 18–25 (2016)
37. Wolfe, J.: Annotations and the collaborative digital library: Effects of an aligned annotation interface on student argumentation and reading strategies. International Journal of Computer-Supported Collaborative Learning **3**(2), 141 (2008)