Tales and Truths: Exploring the Linguistic Journey of 19th Century Literature and Non-Fiction

Suchana Datta[1][0000-0001-9220-6652], Dwaipayan Roy[2][0000-0002-5962-5983], Derek Greene[1][0000-0001-8065-5418], and Gerardine Meaney[1][0000-0002-5412-5007]

[1] University College Dublin, Ireland
[2] Indian Institute of Science Education and Research, Kolkata, India
suchana.datta@ucd.ie, dwaipayan.roy@iiserkol.ac.in, derek.greene@ucd.ie, gerardine.meaney@ucd.ie

Abstract. In this work, we explore the potential of using the lens of information retrieval to reveal societal themes within historical texts. We specifically investigate how term usage evolves over time in the 19th century texts categorised as either fiction or non-fiction. By applying Pseudo-relevance Feedback to a collection of texts from the British Library, segmented by decade, we analyse changes in related terms over time within each category. Our analysis employs standard metrics, such as Kendall's τ, Jaccard similarity, and Jensen-Shannon divergence, to assess overlaps and shifts in these expanded term sets. The results reveal significant divergences in related terms across decades, highlighting key linguistic and conceptual changes during the 19th century.

Keywords: Digital Collection · 19th century fiction · 19th century non-fiction · Information Retrieval · Relevance Feedback.

1 Introduction and Background

The 19th century marked a profound literary transition, with novels evolving from the *Romanticism* of the early 1800s to the *Realism* and *Naturalism* of the later decades, reflecting complex socio-political and cultural shifts. Similarly, the language and vocabulary used in the literature also developed over the century. In information retrieval (IR), handling the evolution of terms over time is essential for creating search systems capable of effectively retrieving relevant information from historical collections. *Vocabulary mismatch* problem has been identified as one of the most challenging issues that is faced by the researchers working with traditional search systems. This issue arises when the terms in a query differ from those in relevant documents, often due to variations in wording for the same concept, leading to low term overlap and reduced retrieval performance. To address vocabulary mismatch problem, *query expansion* is commonly applied in information retrieval to improve search effectiveness by incorporating semantically related terms into the original query [1,3,13,16,29]. In addition, numerous studies have examined the impact of temporal effects on search performance [4,5,7,8,17]. It has been established that incorporating the recency of events generally improves the modelling of temporal aspects in queries and documents, thereby enhancing retrieval effectiveness. However, the vocabulary mismatch problem continues to pose difficulties when dealing with historical data sources, particularly when they span across decades or even an entire century, where the usage of words naturally changes over time. This means that a term that is significant in one decade may lose its relevance in another or even fall out of common use entirely. Conversely, new terms may emerge at any time, as language continually evolves to mirror changing societal, cultural, technological, or intellectual trends. For example, the term 'scientist' – which was first coined by William Whewell in the 19th century – emerged at a time when the study of science was becoming more specialised and institutionalised [27,28].

In the past, various approaches have been employed to analyse word usage trends in extensive text corpora. Simple techniques, such as word frequency analysis and N-gram analysis, exemplified by Google Books Ngram Viewer [20], can provide basic insights. However, the interpretation of such results are likely to be nuanced, depending on factors such as corpus composition and semantic shifts over time [21]. To examine cultural phenomena in more depth, researchers have turned to distributional similarity models [10], which capture word meanings based on co-occurrence patterns. By training word embeddings on corpora from different time periods, it becomes possible to compare semantic representations and identify evolving contexts for specific words. However, this approach necessitates substantial data from each period to ensure robust embeddings [11,14].

Building upon existing research that explored the sociological impacts of language evolution in 19th century fiction [6], we consider the culturally significant British Library Digital Collection (BL19)[1] – a digitised archive comprising a rich and diverse selection of books collected by the British Library during the 19th century. Extending over previous work, we compare texts categorised as either fiction and non-fiction, with the latter serving as an important baseline for comparison. We examine these categories separately by decade, exploring how expansion terms identified through the Relevance-based Language Model (RLM) [12,15] change over time. Our analysis compares these term variations both within and between the two categories. We hypothesise that the highest-weighted expansion terms will show notable differences across decades, reflecting broader linguistic evolution. By comparing expansion terms generated from individual decades across genre against those from the complete collection, we reveal subtle shifts in semantic patterns and language use throughout the period.

In summary, this study addresses the following research questions:

RQ1 - Given the evident evolution of word usage over time, can retrieval techniques be employed to effectively highlight this changing phenomenon?

RQ2 - How does the variation in term usage across different decades impact the searching experience of a user for a specific concept? Specifically, does the temporal shift in terms influence the search results for a particular topic?

[1] See the British Library's Digital Collections: `https://labs.biblios.tech`

RQ3 - Do fiction and non-fiction texts exhibit distinct patterns of language change, or are there overlapping trends in their linguistic evolution? In other words, can we observe a variation in behaviour between the fiction and non-fiction categories when addressing **RQ1** and **RQ2**?

Following from the research questions above, the primary contributions of this work are as follows:

- We perform a detailed analysis of the evolution of specific concepts within the BL19 collection, based on semantically similar terms identified by the pseudo-relevance feedback. Specifically, we assess the correlation between the expansion terms derived from the relevance feedback model for a particular decade in the 19th century and those obtained from the full collection.
- We demonstrate that applying relevance feedback on a given decade from the 19th century exhibits different relevance behaviour in comparison to feedback from the entire time span of the collection.

2 Quantifying Evolution of Concepts

Understanding how ideas and language evolve over time has long fascinated scholars, but quantifying these changes presents unique challenges. By tracking shifts in term associations and meanings across decades, we can quantify how concepts evolved and transformed throughout historical periods. As a step in this direction, our aim in this paper is to study how the relevance and meaning of key concepts have changed over different decades by analysing term overlaps, rank correlations, and semantic shifts for understanding the temporal dynamics of language and themes in literary texts. Our work advances this line of inquiry by examining how the relevance and meaning of key concepts shifted across decades, using term overlaps, rank correlations, and semantic shift analysis to illuminate the temporal dynamics of language and themes in literary texts.

Traditional studies of literary landscape shifts typically require extensive linguistic expertise. To overcome this limitation, in this study, we use a statistical approach that captures conceptual evolution through automated analysis of term relationships. Our methodology examines how the relevance and meaning of key concepts changed across decades by analysing term overlaps, rank correlations, and semantic shifts in literary texts. Our approach consists of three main analytical components. *First*, we examine the overlap of significant expansion terms selected by the RLM between fiction and non-fiction sub-collections, comparing top terms generated for each decade to identify temporal consistency patterns. *Second*, we quantify cross-decade term overlap using Jaccard similarity between feedback queries, revealing which terms maintain relevance over time and which are period-specific. Finally, we assess the alignment between decade-specific and full-collection queries using Kendall's τ rank correlation and Jensen-Shannon divergence measures, providing insight into how well individual decades represent overall linguistic patterns.

Table 1: Number of fiction and non-fiction texts for each decade in the BL19 collection, corresponding to the period from 1831 to 1899.

Decades	1831-40	1841-50	1851-60	1861-70	1871-80	1881-90	1891-99
#Fiction	20	144	746	1139	1750	2034	4377
#Non-fiction	1656	2155	2405	2364	2886	3920	3947

Relevance feedback in temporal migration. In the Pseudo-relevance feedback model (RLM), we aim to find a set of terms that denote a group of concepts related to the main topic of the query. The purpose of this model is to select additional potentially relevant terms to expand the initial query, as a query consisting of only a single term or a few terms is often insufficient to fully represent the user's information need. These expanded terms are selected from the top-ranked documents, under the assumption that these documents provide a good representation of the informational intent, thereby selecting terms that are likely to improve retrieval effectiveness. Formally, we estimate a standard relevance model, θ as $P(w|\theta) = \sum_{j=1}^{M} P(w|d_j) \prod_{q \in Q} P(q|d_j)$, where the weights $P(w|\theta)$ capture the co-occurrences between terms in the M top-retrieved documents, and a query term $q \in Q$. These M documents are selected from the collection $\mathcal{C}$ following an initial retrieval process using the query Q. When there are different subsets $\mathcal{C}^i$ (or sub-collections) characterised by distinct attributes (e.g., thematic or temporal features), the selection of these M documents can either be drawn from the entire collection $\mathcal{C}$ or from a specific subset $\mathcal{C}^i$.

RLM has been widely used in the information retrieval community for enhancing retrieval effectiveness [9,25,24,18]. In this study, we employ RLM to explore how term usage evolves both across decades and within collections of diverse nature. Specifically, we conduct experiments on both $\mathcal{C}_f$ and $\mathcal{C}_n$ collections comprising works published over a century, providing deeper insights into the retrieval process when temporal dimensions are considered.

3 Experiments

Collection. In our experiments, we study two distinct subsets of the complete British Library Digital Collection (BL19), based on book classification information provided by the British Library. *First*, we consider the set of $10,210$ English-language fictional works, with publication dates that span from 1830 to the end of 19th century. The set includes well-known novels by authors such as Charles Dickens and Jane Austen, alongside numerous lesser-known works. *Second*, we consider a collection of 15,780 English-language non-fiction books from the same period, covering a broad spectrum of subjects. These include history, geography, philosophy, and travel, representing the diversity of books collected by the British Library during this time. Exploring these two distinct sets offers a unique lens through which to observe the evolution of language and its connection to key societal shifts during the 19th century. Furthermore, by exam-

Table 2: Lists of keyword queries for three different categories that are used to study the linguistic changes in the BL19 collection.

Thematic	immigrant, emigrant, foreign, newcomer, alien, enslaved, colony, vampire
Plot	engagement, proposal, wedding, suitor, lover, betrothal, eligible, consent, love, mesalliance, heiress, eviction
Genre	crime, murder, mystery, villain, adventure

ining how words and their meanings changed in the context of both fictional and non-fiction texts, we gain insights into cultural and societal transformations.

For the purpose of our analysis, both sets are separately grouped into seven decades, spanning from the 1830s to the 1890s. The distribution of texts across these decades is uneven, particularly among fictional texts, which show a marked increase in the late 19th century (see Table 1). This pattern reflects both the composition of the BL19 collection and the broader rise in British fiction publication during that period. Henceforth, we will refer to the entire collection by $\mathcal{C}$. The sub-collections consisting of fiction novels and non-fiction texts will be denoted as $\mathcal{C}_f$ and $\mathcal{C}_n$, respectively. Furthermore, texts published in the 1830s, 1840s, and subsequent decades will be denoted by $\mathcal{C}_i^{30}$, $\mathcal{C}_i^{40}$, and so on, where $i \in \{f, n\}$. More generally, let $\mathcal{C}^m$ denote any sub-collection corresponding to a specific decade m. For our experiments, we split each sub-collection $\mathcal{C}^m$ in the unit of paragraphs prior to the retrieval.

Topics. As a starting point, a set of 25 queries was identified by scholars with expertise in 19th century British and Irish literature, each of which can be broadly categorised as either *thematic*, *plot*-driven, or *genre*-specific. As we can see from Table 2, these keywords reflect areas of societal change during the 19th century, including gender and migration. Thematic words associated with migration could be expected to occur more frequently as both inward and outward migration increased exponentially from Victorian Britain. However, the fact that the lexical search indicates a strong association between the queries ‘immigrant’ and ‘vampire’ would indicate that this topic is not treated with realism in this body of fiction and that fear and suspicion dominate. Narratives around ‘engagements’ and ‘proposals’ are key elements in both the domestic, realist fiction which came to dominate fiction from the early to the mid-19th century. The relative frequency of terms associated with this kind of ‘marriage’ plot as the century progressed offers an insight into the extent to which this form of fiction dominates the corpus. It also indicates the extent to which plots and issues primarily concerned with the life of young women are a central focus of that fiction over the period of analysis.

The emergence of crime fiction as a distinct genre can be traced through associated terms in the BL19 collection. While literary historians point to Edgar Allan Poe’s 1840s short stories and Arthur Conan Doyle’s 1880s novels as pivotal moments in the genre’s development, our initial lexical analysis reveals a more complex picture. The co-occurrence of terms such as ‘crime’ and ‘mystery’ with 'adventure' suggests that genre boundaries remained fluid throughout the 19th century, challenging conventional narratives about the genre's evolution. For non-fiction, the prominence of crime-related terms within historical records, news reports, and contemporary essays of the 19th century illustrates societal concerns and public discourse around crime, justice, and social order. Reports on criminal investigations, legal proceedings, and social critiques often intersected with fictional representations, revealing how the portrayal of crime shaped and was shaped by societal narratives.

Experimental settings. We use the Lucene implementation[2] of the probabilistic retrieval model BM25 [22,23]. For our experiments, we use the default values of the two parameters, k_1 and b[3]. For RLM, we experimented with the number of top-ranked documents, M, varying in the range $\{50, 100, \ldots, 1000\}$. Another parameter is the number of terms, T, having the highest weight values, $P(w|\theta)$ (in Section 2) which are used to compute the Kullback-Leibler (KL) or, Jensen-Shannon (JS) divergences for re-ranking in a standard RLM setup [15]. The parameter T is varied in the range from 40 to 120 varying in steps of 10. However, the overall observations remain similar with varying the two parameters. Therefore, we report the results with both M and T set to 100.

4 Results and Analysis

4.1 Observations for feedback term selection

We explore whether the demarcation of the collection by decade leads to different outcomes by comparing the two expanded queries generated from the individual sub-collections based on the decades. Specifically, we form expanded queries $EQ_{\mathcal{C}}$ and $EQ^i_{\mathcal{C}}$ using RLM, respectively from $\mathcal{C}$ and $\mathcal{C}^i$. We then perform a retrieval on the entire collection $\mathcal{C}$ both by $EQ_{\mathcal{C}}$ and $EQ^i_{\mathcal{C}}$. Table 3 presents the top 15 terms with the highest weights, as determined by RLM, for three queries – one from each of the three categories ('immigrant', 'mesalliance', and 'murder') as in Table 2. These results are generated from both the individual sub-collections by decade (i.e. $\mathcal{C}^i$) and the entire collection ($\mathcal{C}$). We repeat these experiments on both $\mathcal{C}_f$ as well as $\mathcal{C}_n$, separately. For each sub-collection ($\mathcal{C}_f$ and $\mathcal{C}_n$), the terms that appear consistently across different decades are highlighted in bold.

Our analysis reveals striking differences in how terms evolved across the century. For example, words associated with 'immigrant' showed substantial variation between decades, while those related to 'murder' remained relatively stable (in Table 3). This observation can be supported based on the evolving discourse around immigration in the media, political rhetoric, and public policy during the period in question. Over time, discussions around immigration often shifted from focusing on assimilation and economic contribution to concerns over national security, cultural identity, and the impact of undocumented immigration. These

[2] `https://lucene.apache.org/`

[3] Since explicit relevance judgments are not available for the queries, we are unable to tune the parameters.

Table 3: Feedback term (stemmed) overlaps estimated from both $\mathcal{C}_f$ and $\mathcal{C}_n$ for each decade of the time period from 1831 to 1899. The overlaps are shown in the set of top-scored 15 terms estimated by RLM for three queries, each from a different category (Table 2). For each decade group, the top row contains the feedback terms selected from the $\mathcal{C}_f$ of that corresponding decade and that of $\mathcal{C}_n$ in the bottom row (in blue). Terms that appear in each decade across both the $\mathcal{C}_f$ and $\mathcal{C}_n$ are shown in bold, while the decade in which the corresponding query term does not occur is greyed out.

Query	Decade	Feedback terms
Immigrant (Thematic)	1831-40	
		popul, countri, emigr, state, canada, number, american, year, british, land, western, proport, unit, settler, gener
	1841-50	pudent, avez, censu, raison, soulless, asop, axl, aptli, diadem, altitud, radic, fry, shred, wight, cape
		coloni, emigr, labour, popul, year, land, increas, number, wale, countri, state, person, free, femal, arriv
	1851-60	trade, **countri**, **emigr**, soil, saxon, great, flourish, melbourn, cite, irish, gener, furrugn, nativ, anccstii, intrus
		number, coloni, popul, labour, arriv, state, **emigr**, **countri**, year, increas, port, foreign, total, cooli, unit
	1861-70	britain, **land**, foreign, **govern**, race, distdleri, saxon, countri, law, influx, great, thed, tax, compani, natur
		labour, immi, emigr, state, wage, year, employ, arriv, agent, coloni, **govern**, popul, number, **land**, obtain
	1871-80	coloni, govern, **land**, **countri**, peopl, home, durban, australia, **popul**, **state**, settler, dimsdal, good, **number**, nativ
		state, **land**, **popul**, emigr, year, **countri**, agent, person, settl, agricultur, labour, climat, make, induc, **number**
	1881-90	**land**, **countri**, queensland, room, nativ, work, men, australia, home, great, **year**, place, chines, trial, time
		coloni, labour, emigr, popul, **countri**, agent, **year**, employ, number, arriv, state, **land**, govern, increas, class
	1891-99	alien, favish, land, time, young, turn, yenta, chananya, **year**, man, mendel, dai, america, good, found
		popul, emigr, countri, **year**, state, increas, number, arriv, govern, peopl, unit, decad, labor, europ, coloni
	1831-99	alien, **coloni**, favish, **countri**, land, turn, **emigr**, chananya, **year**, found, time, good, make, australia, thing
		arriv, **emigr**, state, **year**, agent, immi, popul, labour, foreign, unit, number, **countri**, person, europ, **coloni**
Mesalliance (Plot)	1831-40	
		famili, ladi, class, societi, man, marri, decri, digniti, dinnersof, atior, lineag, beampt, bracelet, bonaventuri, honoratior
	1841-50	ohvia, aunt, plebeian, kennyfeck, **famili**, opinion, chateaufort, montoheu, **wife**, sainvil, young, courtois, inept, intermarriag, titut
		sister, ladi, husband, **famili**, butterfli, brush, caught, beauti, father, charm, brother, **wife**, marri, handsom, respect
	1851-60	**marriag**, **marri**, son, ladi, **famili**, daughter, made, young, thing, father, lord, mother, pride, man, posit
		famili, **marri**, **marriag**, madam, parent, liaison, worthi, tast, rank, inexor, sponsor, mistress, girl, hermenbrud, manbrud
	1861-70	**marriag**, famili, marri, made, ladi, **man**, thought, make, heart, daughter, mother, woman, world, love, feel
		marriag, **man**, attaint, soudane, breatralban, address, baron, creat, haynault, blood, regnoit, enchanteress, godrick, countersunk, decid
	1871-80	**ladi**, **marri**, son, famili, **marriag**, **wife**, birth, man, make, poor, girl, **made**, **mother**, **daughter**, love
		marriag, **marri**, husband, **daughter**, lauzun, **wife**, **mother**, put, **ladi**, princ, **made**, fortun, time, antitrinitarian, young
	1881-90	**marri**, **marriag**, **famili**, **ladi**, made, wife, **daughter**, make, **girl**, son, dear, **man**, poor, **young**, love
		daughter, **marriag**, **marri**, **ladi**, **famili**, **girl**, **young**, women, hous, race, wealth, life, **man**, good, gener
	1891-99	**marri**, **famili**, father, **marriag**, **daughter**, mother, make, **man**, **son**, girl, good, thing, **love**, hesseltin, young
		marriag, **marri**, **daughter**, grevi, societi, ladi, life, read, **son**, **famili**, heart, pari, **man**, women, **love**
	1831-99	**marri**, **marriag**, **famili**, **ladi**, son, **daughter**, make, **love**, **girl**, good, **man**, feel, lord, thing, **life**
		marriag, **marri**, **daughter**, **famili**, **ladi**, princess, **life**, **man**, young, **love**, princ, **girl**, wife, made, father
Murder (Genre)	1831-40	gipsei, girl, walsingham, oliv, **blood**, clifford, **man**, deed, exclaim, twist, **time**, fear, heard, moment, repli
		commit, punish, death, crime, **man**, kill, execut, person, guilti, life, **time**, escap, demand, **blood**, prison
	1841-50	**man**, **commit**, **bodi**, cri, **crime**, guilti, found, kill, evid, case, accus, person, heard, head, deed
		commit, kill, **man**, **crime**, escap, counti, wife, hang, drown, protest, death, men, **bodi**, time, evid
	1851-60	**commit**, man, deed, **crime**, cri, accus, **death**, word, blood, answer, exclaim, **case**, bodi, head, evid
		forgeri, **crime**, guilti, **commit**, trial, robberi, wife, hang, child, execut, acquit, **case**, kill, **death**, treason
	1861-70	**man**, **commit**, **crime**, bodi, dead, heard, cri, word, mr, blood, men, horror, prove, found, hear
		guilti, **commit**, trial, execut, acquit, **crime**, **man**, forgeri, kill, hang, death, john, william, convict, wife
	1871-80	**commit**, **man**, **blood**, **death**, **bodi**, heard, finn, **crime**, **found**, night, mr, bonteen, **guilti**, evid, phinea
		commit, trial, **crime**, **man**, sentenc, **death**, **guilti**, juri, verdict, **found**, dead, **blood**, convict, **bodi**, execut
	1881-90	**commit**, **man**, **crime**, victim, polic, **kill**, found, **case**, mysteri, **arrest**, hand, mr, dai, **blood**, time
		crime, **man**, **commit**, death, convict, trial, **case**, guilti, **kill**, **blood**, prison, **arrest**, april, aug, execut
	1891-99	**man**, **commit**, **crime**, dead, polic, **kill**, horror, face, bodi, peopl, word, **blood**, found, hand, cri
		crime, **man**, trial, **kill**, guilti, execut, convict, **commit**, law, death, case, hang, **blood**, crimin, sentenc
	1831-99	**commit**, **man**, horror, **kill**, **dead**, **crime**, cri, victim, **bodi**, **blood**, hand, **death**, call, life, hear
		blood, **commit**, **kill**, guilti, **death**, **bodi**, **crime**, case, arrest, **man**, **dead**, execut, convict, law, polic

shifts reflect changing political landscapes, global crises, and public perceptions. Meanwhile, discussions about crime, while subject to some shifts in terms of specific criminal activities, largely maintained consistent associations with law and order, punishment, and justice. As a result, the concept of crime has remained more stable compared to the often fluid and contentious conversations surrounding immigration. Particularly notable is the prominence of 'Australia' among the expansion terms in the fictional collection. This reflects the impact of British colonization of Australia, which began with the First Fleet's arrival in 1787-1788[4]. The frequent appearance of Australia-related terms in British literature suggests how deeply colonial expansion influenced the literary imagination of the period, as authors grappled with themes of empire, settlement, and cultural transformation.

The evolution of the term 'alien' presents another fascinating pattern. Though derived from the Latin 'alienus' (meaning "belonging to another") and codified in British law through the 1793 *Aliens Act*, its association with immigration in novels emerged primarily in the late 1800s. This linguistic shift hints at changing attitudes toward foreignness and belonging in British society, demonstrating how literature both reflected and shaped evolving social perspectives. Interestingly, this trend was less pronounced in the non-fiction collection.

In general, the terms chosen in the $\mathcal{C}_n$ collection tended to be more universal, such as 'country', 'year', 'government', 'population' etc. These terms are more neutral and factual, reflecting the emphasis on documenting or discussing real-world events and concepts. The language in non-fiction is often constrained by the need for precision and clarity, leading to the use of general terms that are broadly applicable across different contexts. This contrasts with fictional works, where language is often more flexible and creative, allowing for a broader range of specialised or abstract terms that serve specific narrative purposes.

By comparing the across-collection expansion term overlap in Table 3, we can observe that there is generally a lower degree of term overlap across both decades and collection types for thematic queries (e.g. 'immigrant'), reflecting the conceptual diversity and nuanced nature of these queries. Conversely, there is a greater degree of term overlap within specific genre ('murder' in our example) categories across decades, as well as within the same collection type. This increased overlap is due to the conceptually focused nature of genre-specific terms, which often adhere to established conventions and recurring motifs, making them more consistent over time and across related collections[5].

4.2 Correlation in the lists of re-ranked documents

We compute the rank correlation between the two ranked lists produced by $EQ_{\mathcal{C}}$ and $EQ^i_{\mathcal{C}}$ in different settings. In Table 4, we report the rank correlations (Kendall's τ) between the ranked lists with the two expanded queries for $\mathcal{C}_f$ as well as $\mathcal{C}_n$. The results demonstrate that the correlation between the ranked lists is generally quite low, with the highest observed correlation being a 0.256 for the query *wedding* within $\mathcal{C}_n$ and only 0.189 for the query *murder* within $\mathcal{C}_f$. This indicates that the expanded queries derived from different sub-collections produce markedly distinct retrieval outcomes. This observations are graphically shown in Figure 1, depicting the intensity of the rank correlation values for $\mathcal{C}_f$ and $\mathcal{C}_n$.

[4] `https://en.wikipedia.org/wiki/Immigration_history_of_Australia` and `https://www.homeaffairs.gov.au/about-us-subsite/files/immigration-history.pdf`.

[5] Codes and artifacts are available at https://github.com/suchanadatta/RLM-BL19

Table 4: Comparisons of rank correlation values (measured with Kendall's τ) between two re-ranked lists retrieved by RLM augmented queries, where one query is obtained from the entire century collection and the other is estimated from an individual decade, both for fiction ($\mathcal{C}_f$) and non-fiction ($\mathcal{C}_n$) collection. For each query, the correlations are to be compared (i) between two types of collections ($\mathcal{C}_f$ and $\mathcal{C}_n$) of the same decade, and (ii) across decades, i.e. along the columns. For each decade (column groups), the highest and lowest correlations are highlighted, respectively in green and red; and for each query (across rows) the highest across all the decades are bold-faced. Grey cells indicate the absence of the query term in the corresponding decade's collection.

		1831-40		1841-50		1851-60		1861-70		1871-80		1881-90		1891-99	
Type	Query	$\mathcal{C}_f$	$\mathcal{C}_n$	$\mathcal{C}_f$	$\mathcal{C}_n$	$\mathcal{C}_f$	$\mathcal{C}_n$	$\mathcal{C}_f$	$\mathcal{C}_n$	$\mathcal{C}_f$	$\mathcal{C}_n$	$\mathcal{C}_f$	$\mathcal{C}_n$	$\mathcal{C}_f$	$\mathcal{C}_n$
Thematic	Immigrant		0.032	-0.020	0.055	-0.024	0.052	0.063	-0.048	-0.001	**0.095**	-0.016	-0.001	0.025	0.093
	Emigrant	0.069	0.070	-0.060	0.048	-0.074	-0.071	-0.001	0.067	0.091	0.086	**0.125**	0.055	-0.012	0.082
	Foreign	0.010	0.050	-0.073	-0.072	0.095	-0.041	0.041	0.036	-0.077	**0.121**	0.042	0.067	0.036	0.004
	Newcomer	-0.115	0.028	-0.062	**0.161**	-0.015	-0.034	-0.010	0.076	-0.011	0.039	0.061	-0.031	0.030	0.057
	Alien	-0.041	0.015	-0.114	-0.028	-0.013	0.052	**0.141**	0.390	-0.080	0.021	-0.070	0.098	0.051	0.053
	Enslaved	-0.049	0.059	-0.042	0.061	0.034	0.017	**0.072**	0.052	-0.064	-0.061	0.068	0.065	0.038	0.069
	Colony	-0.021	0.070	0.014	0.047	0.035	-0.008	0.068	**0.075**	0.016	0.017	-0.033	0.071	-0.018	0.019
	Vampire		-0.004	0.020	0.028	0.098	**0.123**	0.075	0.055	0.018	0.061	0.027	0.066	0.038	0.106
Plot	Engagement	0.013	0.029	0.006	0.071	-0.003	0.061	0.043	-0.014	0.109	0.158	0.105	0.115	0.031	**0.126**
	Proposal	-0.004	0.010	-0.068	0.081	-0.006	0.065	0.101	**0.182**	-0.064	-0.059	0.037	0.040	0.073	0.103
	Wedding	-0.031	-0.100	-0.001	0.015	-0.128	0.022	0.060	0.014	-0.056	**0.256**	-0.076	-0.116	-0.030	0.008
	Suitor	0.011	-0.059	-0.110	**0.057**	-0.062	-0.021	0.055	0.009	-0.015	0.038	-0.117	0.054	-0.105	0.045
	Lover	0.007	0.030	0.003	-0.037	0.027	-0.006	0.085	-0.016	0.057	-0.084	0.006	0.046	0.078	**0.108**
	Betrothal	-0.038	-0.050	0.105	-0.026	-0.008	0.002	0.016	-0.025	0.060	**0.118**	-0.103	0.080	-0.059	-0.055
	Eligible	-0.004	0.027	-0.056	**0.199**	0.080	-0.006	0.032	-0.058	-0.167	-0.169	0.113	-0.058	-0.044	-0.046
	Consent	0.042	-0.025	-0.074	0.066	-0.045	0.010	0.080	**0.121**	0.068	-0.080	0.007	-0.072	0.074	-0.109
	Love	0.019	-0.051	0.089	0.034	-0.027	0.020	0.013	0.073	-0.019	-0.173	**0.174**	-0.060	-0.037	0.137
	Mesalliance		**0.128**	0.107	0.018	0.041	0.018	0.065	0.069	-0.027	-0.133	0.034	0.033	0.012	0.048
	Heiress	-0.009	**0.142**	-0.027	-0.065	-0.024	-0.063	0.140	-0.059	0.050	0.084	0.027	0.066	0.098	0.031
	Eviction		0.015	-0.021	-0.015	0.019	0.031	-0.043	-0.026	-0.026	0.038	-0.108	0.039	0.069	**0.085**
Genre	Crime	0.122	0.004	0.024	0.006	-0.013	-0.054	-0.023	0.136	0.040	0.007	0.011	**0.187**	0.038	0.051
	Murder	0.103	0.003	0.005	-0.010	0.037	0.038	0.141	0.086	**0.189**	-0.028	0.073	-0.053	0.050	0.129
	Mystery	-0.005	-0.016	0.037	-0.093	0.060	0.105	0.081	-0.010	-0.108	0.040	-0.104	0.083	-0.153	-0.151
	Villain	-0.022	-0.010	-0.003	0.037	0.015	0.052	0.067	-0.063	0.064	**0.084**	0.048	-0.017	-0.145	-0.035
	Adventure	-0.031	0.115	-0.006	-0.025	-0.078	-0.068	-0.036	0.160	-0.087	-0.001	0.049	-0.065	0.018	0.131

From the Figure, we can observe that overall, there are not much correlations both for $\mathcal{C}_f$ and $\mathcal{C}_n$. However, on a closer look, it can be realised that the correlation is relatively better for $\mathcal{C}_n$. This is because the rank correlations between the re-ranked lists obtained from feedback queries generated from a specific decade and those generated from the entire collection tend to be higher for $\mathcal{C}_n$.

4.3 Similarities and diversions in feedback query term distributions

Next, Jaccard similarity is evaluated between the entire collection $\mathcal{C}$ and any pair of sub-collections $\mathcal{C}^i$. We measure similarities between pairs of sub-collections as well. We calculate the average Jaccard similarity for each pair of decades, $\mathcal{C}^i$ and $\mathcal{C}^j$, using the expanded queries obtained by RLM for each of the 25 keyword queries. This process is repeated separately for both collections ($\mathcal{C}_f$ and $\mathcal{C}_n$), and the results are reported in two groups in Table 5 (left and right, respectively). The standard deviation of these Jaccard similarities is also reported in order to quantify the variations in term overlaps across different topic pairs. In Table 5, the upper triangle of each group displays the average Jaccard similarity between

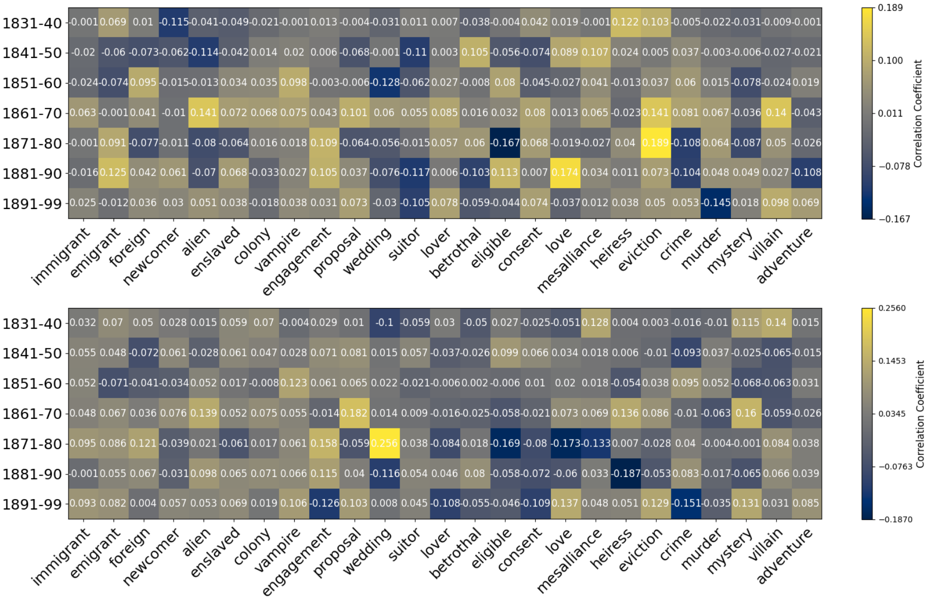

Fig. 1: Heatmap visualising rank correlations (Kendall's τ) between re-ranked document lists, retrieved by two feedback queries, (i) generated from the decade subcollections $\mathcal{C}^i$, and (ii) from the full collection $\mathcal{C}$ – both for $\mathcal{C}_f$ (top) and $\mathcal{C}_n$ (bottom).

the expanded query term sets for each of the 25 topics. The mean Jaccard similarity between collections $\mathcal{C}^i$ and $\mathcal{C}^j$ is shown in each cell (i, j) of this triangle. The standard deviation of these Jaccard similarities is provided in parenthesis inside each cell to provide a measure of dispersion. According to Table 5, the 1890s decade and the complete collection have the highest similarity, with an average Jaccard similarity of more than 0.44.

To supplement the Jaccard similarity analysis, the Jensen-Shanon (JS) divergence is also calculated and shown in the lower part of Table 5. JS divergence considers term weights and determines the divergence between two expanded queries from two distinct sub-collections, whereas the Jaccard measure only takes into account the overlap of words in two expanded query sets, disregarding term weights. Interestingly, the average JS divergence for all decade pairs stays reasonably consistent at 0.5, with the lowest divergence value of 0.48 for $\mathcal{C}_f$, even though there are variations in the composition of specific terms (indicated by lower Jaccard values).

The non-fiction collection $\mathcal{C}_n$, on the other hand, has a generally lower divergence, with the $\mathcal{C}_n^{40}$ and $\mathcal{C}_n$ showing a minimum value of 0.36. A comparatively lower standard deviation for all queries suggests that there is minimal variation in the values, indicating that both the JS divergence and Jaccard similarity for all queries within a decade exhibit consistent trends. This illustrates how language and conceptual frameworks change over time and points to a growing

Table 5: Average variations of feedback term distributions for 25 queries obtained by RLM, both for fiction, $\mathcal{C}_f$ (left group) and non-fiction, $\mathcal{C}_n$ (right group) collection, comparing decades of 19th century. Upper triangles (grey cells) depict Jaccard similarities of feedback queries, whereas lower halves (white cells) show JS divergence between two feedback term distributions. Average variation and standard deviation of 25 queries are also reported in each cell (in brackets). Highest Jaccard similarity and lowest JS divergence are bold-faced in each group.

	Fiction ($\mathcal{C}_f$)								**Non-fiction ($\mathcal{C}_n$)**							
	1830s	**1840s**	**1850s**	**1860s**	**1870s**	**1880s**	**1890s**	**1831-99**	**1830s**	**1840s**	**1850s**	**1860s**	**1870s**	**1880s**	**1890s**	**1831-99**
1830s	1 (0)	0.1756 (0.1392)	0.1718 (0.1430)	0.1517 (0.1245)	0.1541 (0.1275)	0.1539 (0.1193)	0.1322 (0.1160)	0.1405 (0.1220)	1 (0)	0.3438 (0.1309)	0.3215 (0.1266)	0.2848 (0.1160)	0.2607 (0.1294)	0.3068 (0.1110)	0.3213 (0.1109)	0.3068 (0.1190)
1840s	0.5249 (0.2084)	1 (0)	0.2618 (0.1422)	0.2541 (0.1314)	0.2465 (0.1295)	0.2383 (0.1241)	0.2245 (0.1361)	0.2257 (0.1377)	0.4403 (0.1828)	1 (0)	0.3474 (0.1171)	0.3125 (0.1100)	0.2911 (0.1303)	0.3184 (0.1227)	0.3338 (0.1009)	0.3158 (0.1230)
1850s	0.5190 (0.2140)	0.5188 (0.1954)	1 (0)	0.3214 (0.1108)	0.2967 (0.1181)	0.2801 (0.1047)	0.2717 (0.1256)	0.3166 (0.1307)	0.4421 (0.1873)	0.4011 (0.1911)	1 (0)	0.3352 (0.1164)	0.2817 (0.1208)	0.3275 (0.1052)	0.3369 (0.1067)	0.3561 (0.1277)
1860s	0.5341 (0.1575)	0.5616 (0.1434)	0.5469 (0.1770)	1 (0)	0.3304 (0.1117)	0.3166 (0.1173)	0.2946 (0.1214)	0.3286 (0.1248)	0.4429 (0.1497)	0.4413 (0.1354)	0.4633 (0.1430)	1 (0)	0.2810 (0.1268)	0.3183 (0.1132)	0.3353 (0.1302)	0.3513 (0.1245)
1870s	0.5633 (0.1880)	0.5403 (0.1776)	0.5274 (0.1963)	0.5493 (0.1766)	1 (0)	0.3450 (0.0928)	0.3064 (0.1198)	0.3543 (0.1217)	0.4212 (0.1855)	0.4622 (0.1420)	0.4411 (0.1705)	0.4236 (0.1652)	1 (0)	0.3481 (0.0921)	0.3106 (0.1150)	0.3479 (0.1216)
1880s	0.5912 (0.2082)	0.5443 (0.1726)	0.5899 (0.1064)	0.5267 (0.1971)	0.5283 (0.1975)	1 (0)	0.3354 (0.1126)	0.3646 (0.1190)	0.4608 (0.1542)	0.4621 (0.1406)	0.4648 (0.1094)	0.4420 (0.1693)	0.4825 (0.1020)	1 (0)	0.3757 (0.1143)	0.3961 (0.1128)
1890s	0.5234 (0.1594)	**0.4817** (**0.2233**)	0.5891 (0.1031)	0.5892 (0.1072)	0.5918 (0.1004)	0.5732 (0.1393)	1 (0)	**0.4491** (**0.1484**)	0.4644 (0.1491)	0.4031 (0.2074)	0.4669 (0.0956)	0.5048 (0.0098)	0.4823 (0.0934)	0.3640 (0.0093)	1 (0)	**0.4748** (**0.1206**)
1831-99	0.5483 (0.1593)	0.5835 (0.1078)	0.5055 (0.2121)	0.5912 (0.1005)	0.4861 (0.2309)	0.6133 (0.0085)	0.5505 (0.1704)	1 (0)	0.4441 (0.1467)	**0.3632** (**0.1009**)	0.4452 (0.1692)	0.4649 (0.0916)	0.4613 (0.1445)	0.4821 (0.0719)	0.4443 (0.1638)	1 (0)

divergence in the semantic space of query expansions. Additionally, this exhibits a constant degree of semantic separation which varies rarely between the various historical periods. In summary, we can conclude that the overall similarity in the $\mathcal{C}_n$ is higher compared to $\mathcal{C}_f$, exhibiting a lower level of divergence. This higher similarity and lower divergence in $\mathcal{C}_n$ likely stem from its more consistent terminology and subject matter, as opposed to the greater conceptual diversity and evolving narratives of fiction.

To compare the feedback terms across the collections ($\mathcal{C}_f$ and $\mathcal{C}_n$), we report the average Jaccard similarity (measuring term overlap) and the JS divergence (considering term weights) in Table 6. The table reveals a relatively stable level of Jaccard similarities and JS divergence between the $\mathcal{C}_f$ and $\mathcal{C}_n$(indicated by the lower standard deviation). While the highest Jaccard similarity of 0.21 is observed between the entire $\mathcal{C}_f$ and $\mathcal{C}_n^{90}$, the corresponding JS divergence is 0.59, indicating substantial semantic differences. Conversely, the lowest divergence is found (between $\mathcal{C}_f^{30}$ and $\mathcal{C}_n^{70}$), with a Jaccard similarity of only 0.08, suggesting minimal term overlap. These findings suggest that while there may be some overlap in the types of terms selected for feedback in fiction and non-fiction collections, the overall semantic space of these terms differs significantly.

5 Discussion

From Table 3, we observe significant variations in term selection when using the same technique (in our case, RLM, which identifies contextually similar terms). This highlights the model's ability to capture subtle changes in word usage and associations. Thus, in response to **RQ1**, we conclude that RLM [15] is indeed

Table 6: Average Jaccard similarities (left group) and JS divergences (right group) of 25 queries, measured between feedback queries generated from the fiction collection ($\mathcal{C}_f$) and that of its non-fiction ($\mathcal{C}_n$) counterpart. Similar to Table 5, each cell shows the average variation and the standard deviation (in brackets) of 25 keyword queries for the corresponding decade. Highest Jaccard similarity and lowest JS divergence are highlighted respectively in blue and green.

	Jaccard								**JS Divergence**							
	1830s	**1840s**	**1850s**	**1860s**	**1870s**	**1880s**	**1890s**	**1831-99**	**1830s**	**1840s**	**1850s**	**1860s**	**1870s**	**1880s**	**1890s**	**1831-99**
1830s	0.1115 (0.0986)	0.1082 (0.0963)	0.1006 (0.0943)	0.0946 (0.0797)	0.0894 (0.0812)	0.0980 (0.0841)	0.1031 (0.0815)	0.0723 (0.0647)	0.4624 (0.1483)	0.4602 (0.1556)	0.4615 (0.1577)	0.4426 (0.1780)	0.4190 (0.2039)	0.5021 (0.0069)	0.4197 (0.2039)	0.5016 (0.0073)
1840s	0.1577 (0.0898)	0.1650 (0.0932)	0.1542 (0.0911)	0.1428 (0.0867)	0.1426 (0.1006)	0.1528 (0.0887)	0.1651 (0.0892)	0.1207 (0.0752)	0.5998 (0.0079)	0.5795 (0.1132)	0.5374 (0.1865)	0.5582 (0.1579)	0.5998 (0.0067)	0.5806 (0.1081)	0.6027 (0.0082)	0.5811 (0.0097)
1850s	0.1665 (0.0866)	0.1751 (0.0913)	0.1697 (0.0827)	0.1610 (0.0749)	0.1636 (0.0884)	0.1736 (0.0764)	0.1828 (0.0775)	0.1411 (0.0715)	0.5857 (0.1107)	0.5436 (0.1888)	0.5860 (0.1164)	0.5476 (0.1509)	0.5661 (0.1513)	0.5849 (0.0074)	0.5683 (0.1551)	0.5899 (0.0979)
1860s	0.1592 (0.0779)	0.1750 (0.0888)	0.1618 (0.0738)	0.1633 (0.0791)	0.1628 (0.0799)	0.1762 (0.0786)	0.1812 (0.0742)	0.1380 (0.0665)	0.5657 (0.1560)	0.5458 (0.1549)	0.5850 (0.1153)	0.5655 (0.1543)	0.5856 (0.1124)	0.5869 (0.1129)	0.5672 (0.1584)	0.5887 (0.0966)
1870s	0.1680 (0.0820)	0.1801 (0.0867)	0.1665 (0.0694)	0.1643 (0.0739)	0.1777 (0.0867)	0.1905 (0.0719)	0.1967 (0.0665)	0.1483 (0.0665)	0.5883 (0.0077)	0.5280 (0.2046)	0.5683 (0.1510)	0.5244 (0.2087)	0.5692 (0.1445)	0.5879 (0.0084)	0.5889 (0.1127)	0.5892 (0.1074)
1880s	0.1735 (0.0744)	0.1853 (0.0799)	0.1798 (0.0750)	0.1737 (0.0817)	0.1763 (0.0864)	0.2075 (0.0918)	0.2089 (0.0717)	0.1599 (0.0768)	0.5858 (0.1160)	0.5254 (0.2046)	0.5272 (0.2043)	0.5239 (0.1897)	0.5467 (0.1836)	0.6093 (0.0075)	0.5894 (0.1053)	0.5885 (0.1156)
1890s	0.1590 (0.0822)	0.1737 (0.0897)	0.1617 (0.0816)	0.1638 (0.0802)	0.1659 (0.0887)	0.1844 (0.0830)	0.1978 (0.0767)	0.1516 (0.0799)	0.5677 (0.1513)	0.6086 (0.0075)	0.5667 (0.1537)	0.5873 (0.1094)	0.5675 (0.1499)	0.5660 (0.1535)	0.5488 (0.1849)	0.6087 (0.0095)
1831-99	0.1700 (0.0771)	0.1875 (0.0944)	0.1772 (0.0777)	0.1794 (0.0878)	0.1745 (0.0816)	0.1997 (0.0808)	0.2113 (0.0742)	0.1643 (0.0781)	0.5681 (0.1477)	0.5271 (0.1781)	0.6093 (0.0073)	0.5467 (0.1809)	0.5478 (0.1800)	0.5880 (0.1086)	0.5902 (0.1056)	0.5694 (0.0942)

capable of capturing the evolution of word associations over time. Applying RLM to the entire collection could result in penalties for search intent due to changes in how query terms and their associated terms are used. The results in Tables 4 and 5 confirm that applying RLM on a decade-wise basis is more effective. Feedback terms selected for a specific decade differ significantly from those chosen from the entire collection, showing minimal overlap as illustrated in Table 3. Hence, in response to **RQ2**, we conclude that the performance of a feedback-based retrieval model can be influenced if the underlying feedback pool is altered (e.g. from $\mathcal{C}^p$ to $\mathcal{C}^q$). This is valid for both the fiction ($\mathcal{C}_f$) as well as non-fiction ($\mathcal{C}_n$) collections. Additionally, we observe that feedback term overlaps between $\mathcal{C}_f$ and $\mathcal{C}_n$ are similarly limited (see Table 3).

We see that for $\mathcal{C}_f$, the average Jaccard similarity between decades is relatively low, while the JS divergence is high, indicating greater variability in term usage across decades. Conversely, in $\mathcal{C}_n$, we noted higher Jaccard similarities and slightly lower JS divergences as compared to $\mathcal{C}_f$, reflecting more consistency in language usage over time (refer to Table 5). Furthermore, when performing re-ranking using expanded queries generated by RLM, terms selected from a specific decade led to significantly different ranked lists compared to those generated using the entire collection. This was evident through the Kendall's τ values, which often showed negative correlations as in Table 4. In contrast, re-ranking within the $\mathcal{C}_n$ demonstrated a relatively higher correlation between the decade-specific ($\mathcal{C}^i$) and that of the full collection ($\mathcal{C}$) ranked lists (higher Kendall's τ values are observed in Table 4 and Figure 1). This indicates that applying RLM on the entire $\mathcal{C}_n$ for re-ranking is less detrimental compared to its application on the fictional counterpart, $\mathcal{C}_f$. This difference can be attributed to the greater conceptual diversity and subjectivity inherent in $\mathcal{C}_f$, as opposed to the relative consistency and factual focus found in $\mathcal{C}_n$. Hence, in addressing **RQ3**, we conclude that although both fictional and non-fictional collections display temporal variations in language usage, the specific patterns and the extent of these variations differ between the two genres. Thus, we hypothesise that, due to the significant evolution of terms within the 19th century fiction collection, applying relevance feedback techniques such as RLM [12,15] should be performed on a decade-specific sub-collection, rather than on the entire collection.

An important distinction in our approach, compared to traditional lexicographical resources such as the Oxford English Dictionary (OED), is the focus on the prevalence of word usage in context rather than merely tracking its first recorded occurrence. The OED is known for its historical lexicography, offering etymologies and first-known instances of words [26]. This chronological emphasis is useful for tracing linguistic origins. However, it may not indicate the same when a particular meaning or usage becomes widespread and embedded in language. In contrast, our corpus-based methods that leverage statistical modelling and retrieval techniques offer a broader understanding of the evolution of word usage by analysing word frequency and semantic shifts over time (as presented in [19]). Our approach highlights when new meanings or trends achieve prominence within a corpus, providing insights into broader societal or cultural changes reflected through language [2]. By analysing term usage in context, we can capture subtle but meaningful linguistic shifts and how these shifts mirror or influence historical and socio-cultural dynamics.

6 Conclusion

In this paper, we presented our findings on the study of evolving language usage in 19th century texts, considering both fiction and non-fiction, using an information retrieval approach. Specifically, we investigated how the usage of words has changed over time by applying a Relevance-based Language Model, an effective relevance feedback technique that leverages the statistical occurrences of words. Our results indicate a significant variation in related terms over decades, showing important language and conceptual developments during the 19th century.

As part of future work, we plan to explore the effectiveness of applying IR models to capture and quantify sociological trends in literature and non-fiction, with the goal of enhancing search capabilities for historical collections such as BL19. This investigation will require annotated relevance judgements to evaluate retrieval effectiveness, which we plan to collect in a formal setting with input from domain experts. Additionally, we aim to apply these approaches to other collections (e.g. historical and modern newspaper collections) to study their patterns regarding the evolution of term usage.

Acknowledgement. This publication is part of a project that has received funding from (i) the European Research Council (ERC) under the Horizon 2020 research and innovation program (Grant agreement No. 884951); (ii) Science Foundation Ireland (SFI) to the Insight Centre for Data Analytics under grant No 12/RC/2289_P2.

References

1. Azad, H.K., Deepak, A.: Query expansion techniques for information retrieval: A survey. Information Processing & Management **56**(5), 1698–1735 (2019)
2. Baker, C.: Foundations of Bilingual Education and Bilingualism. Bilingual education and bilingualism, Multilingual Matters (2006)
3. Bassani, E., Tonellotto, N., Pasi, G.: Personalized query expansion with contextual word embeddings. ACM Trans. Inf. Syst. **42**(2) (Dec 2023)
4. Craveiro, O., Macedo, J., Madeira, H.: Query expansion with temporal segmented texts. In: Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval - Volume 8416. p. 612–617. ECIR 2014, Springer-Verlag, Berlin, Heidelberg (2014)
5. Dakka, W., Gravano, L., Ipeirotis, P.G.: Answering general time sensitive queries. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. p. 1437–1438. CIKM '08, Association for Computing Machinery, New York, NY, USA (2008)
6. Datta, S., Roy, D., Greene, D., Meaney, G.: Unveiling temporal trends in 19th century literature: An information retrieval approach. arXiv preprint arXiv:2501.06833 (2025), accepted at JCDL 2024.
7. Efron, M., Golovchinsky, G.: Estimation methods for ranking recent information. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 495–504. SIGIR '11, Association for Computing Machinery, New York, NY, USA (2011)
8. Efron, M., Lin, J., He, J., de Vries, A.: Temporal feedback for tweet search with non-parametric density estimation. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. p. 33–42. SIGIR '14, Association for Computing Machinery, New York, NY, USA (2014)
9. Ganguly, D., Leveling, J., Jones, G.: Cross-lingual topical relevance models. In: Proceedings of COLING 2012. pp. 927–942. The COLING 2012 Organizing Committee, Mumbai, India (Dec 2012)
10. Gulordava, K., Baroni, M.: A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In: Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics. pp. 67–71 (2011)
11. Hamilton, W.L., Leskovec, J., Jurafsky, D.: Diachronic word embeddings reveal statistical laws of semantic change. In: Proc. 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1489–1501 (2016)
12. Jaleel, N.A., Allan, J., Croft, W.B., Diaz, F., Larkey, L.S., Li, X., Smucker, M.D., Wade, C.: Umass at TREC 2004: Novelty and HARD. In: Proceedings of the Thirteenth Text REtrieval Conference. NIST Special Publication, vol. 500-261. National Institute of Standards and Technology (NIST) (2004)
13. Kung, P.P.H., Fan, Z., Zhao, T., Liu, Y., Lai, Z., Shi, J., Wu, Y., Yu, J., Shah, N., Venkataraman, G.: Improving embedding-based retrieval in friend recommendation with ann query expansion. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2930–2934. SIGIR '24, Association for Computing Machinery, New York, NY, USA (2024)
14. Kutuzov, A., Øvrelid, L., Szymanski, T., Velldal, E.: Diachronic word embeddings and semantic shifts: A survey. arXiv preprint arXiv:1806.03537 (2018)

15. Lavrenko, V., Croft, W.B.: Relevance based language models. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 120–127. SIGIR '01, Association for Computing Machinery, New York, NY, USA (2001)
16. Li, M., Zhuang, H., Hui, K., Qin, Z., Lin, J., Jagerman, R., Wang, X., Bendersky, M.: Can query expansion improve generalization of strong cross-encoder rankers? In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2321–2326. SIGIR '24, Association for Computing Machinery, New York, NY, USA (2024)
17. Li, X., Croft, W.B.: Time-based language models. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management. p. 469–475. CIKM '03, Association for Computing Machinery, New York, NY, USA (2003)
18. Mackie, I., Chatterjee, S., Dalton, J.: Generative relevance feedback with large language models. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2026–2031. SIGIR'23, Association for Computing Machinery (2023)
19. McEnery, T., Hardie, A.: Corpus Linguistics: Method, Theory and Practice. Cambridge Textbooks in Linguistics, Cambridge University Press (2011)
20. Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Team, T.G.B., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., Aiden, E.L.: Quantitative analysis of culture using millions of digitized books. Science **331**(6014), 176–182 (2011)
21. Pettit, M.: Historical time in the age of big data: Cultural psychology, historical change, and the google books ngram viewer. History of Psychology **19**(2), 141 (2016)
22. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: SIGIR '94. pp. 232–241. Springer London, London (1994)
23. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval **3**(4), 333–389 (apr 2009)
24. Roy, D., Ganguly, D., Mitra, M., Jones, G.J.: Word vector compositionality based relevance feedback using kernel density estimation. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. p. 1281–1290. CIKM '16, Association for Computing Machinery, New York, NY, USA (2016)
25. Salakhutdinov, R., Mnih, A.: Bayesian probabilistic matrix factorization using markov chain monte carlo. In: Proceedings of the 25th International Conference on Machine Learning. p. 880–887. ICML '08, Association for Computing Machinery, New York, NY, USA (2008)
26. Simpson, J., Weiner, E.: The Oxford English dictionary Vol. 2. Oxford : Clarendon Press ; Oxford ; New York : Oxford University Press, 2nd ed. vol. 2 edn. (1989)
27. Whewell, W.: History of the Inductive Sciences: From the Earliest to the Present Times. No. v. 1 in History of the Inductive Sciences: From the Earliest to the Present Times, J.W. Parker (1837)
28. Whewell, W.: History of the Inductive Sciences: From the Earliest to the Present Times. Cambridge University Press (Sep 2010)
29. Zhao, L., Callan, J.: Automatic term mismatch diagnosis for selective query expansion. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 515–524. SIGIR '12, Association for Computing Machinery, New York, NY, USA (2012)