CHAPTER 6

Greening a Post-Industrial City: Applying Keyword Extractor Methods to Monitor a Fast-Changing Environmental Narrative

Sarah Luria
Department of English, College of the Holy Cross, Worcester, Massachusetts, USA

Ricardo Campos
Ci2 – Smart Cities Research Center – ICT Departmental UNIT, Polytechnic Institute of Tomar, Portugal & INESC TEC., Portugal

Worcester, Massachusetts (population just under 200,000) is the second-largest city after Boston in Massachusetts. It grew exponentially during the American industrial revolution; factories dominated its landscape until the mid-20th century when the city slid physically, and psychologically, into post-industrial decline. As one local magazine reports, residents took an abject view of the city: 'Nothing happens in this dump. There's never any development' (Whitmore, Bernard, 2018).

Since the 1960s, Worcester's city government and residents have tried to "revitalize" the city. Some grandiose plans failed, but other small, creative changes took root. Today momentum is building, and the city is now

How to cite this book chapter:
Luria, Sarah and Ricardo Campos (2022). "Greening a Post-Industrial City: Applying Keyword Extractor Methods to Monitor a Fast-Changing Environmental Narrative." In: *Unlocking Environmental Narratives: Towards Understanding Human Environment Interactions through Computational Text Analysis.* Ed. by Ross S. Purves, Olga Koblet, and Benjamin Adams. London: Ubiquity Press, pp. 109–132. DOI: https://doi.org/10.5334/bcs.f.

experiencing a visible 'renaissance' (Whitmore, Bernard, 2018). The city is being touted as a model of urban green development: its railroad is being upgraded to provide improved service to Boston, just 40 miles away. Enthusiasm for investment is concentrating mainly in the city's historic Canal District – named for the canal that connected the neighbourhood's factories to the nearby Blackstone River and the Atlantic Seacoast beyond. Those old factories are being converted into trendy loft apartments and hipster bars and cafés. A new baseball stadium, Polar Park, has been built right in the heart of the District, and this paves the way for development from outside investors on a larger scale, including a new complex of hotels, apartments and stores. National Public Radio recently dubbed the city as the 'New "It" Town" – all of a sudden, 'everyone wants a piece of Worcester' (Schacter, Aaron, 2018a).

The Canal District was first developed in the 1800s by the Irish, who dug its canal, the bold entrepreneurs who built the factories and the successive waves of immigrants who came to work in them. Churches were established, Jewish peddlers sold goods and opened shops, and home distilleries, often run by Irish women in their kitchens, brought neighborhood residents together. This stalwart history and old industrial fabric create the Canal District's allure, but current up-scale development threatens to displace the present immigrant population (largely Latino) that continues the neighbourhood's working-class identity. This probable result, which would lead to loss of homes and community, raises the question of how exactly history is being preserved through the Canal District's development, and whether it is possible to 'green' communities for people other than the well-to-do – to avert what can be called 'the greening of green'. As one activist's sign in Chicago put it, 'Now that the neighborhood is nice, why do I have to move?' (Saunders, 2018).

In this project, we focus on the stories being told about the Canal District that help drive its current development. How did it go from being spoken of as 'a dump' to 'the new "It" town', and how does such language and its narratives impact the city and the market? Our aim is to survey a small corpus of texts describing the Canal District's 'revitalization' and discuss the efficacy of computational methods through an expert reading of the same texts. At first glance, Worcester might present an all-too-familiar global story of gentrification, the loss of city identity and community, but does even a small survey of historical and local public discourse confirm this view? Or does a more complicated local story surface?

To answer these questions, literary geographer Sarah Luria teams up with computer scientist Ricardo Campos, developer of the keyword extractor YAKE! (Campos et al., 2018a; Campos et al., 2020) to discover if YAKE! can create a helpful digest of stories told about a neighbourhood over time. We believe this interdisciplinary work can play a crucial role by showing how computational analysis can track this fast-developing story of urban revitalisation. Development has a momentum that can overwhelm local dissent as signs in Worcester already show. 'All this fake love, the "New Worcester", this whole

new wave that's coming in, that's not Worcester', one resident complains. Some see Worcester's development as at a 'tipping point', where Worcester could 'stay Worcester' or be fundamentally redefined, as has happened in places such as Harlem, New York and San Francisco (Schacter, Aaron, 2018b). Such voices can get buried in the attractive 'renaissance' rhetoric of up-scale 'green development'. We hope that by being able to process a range of documents and sources about Worcester, we can produce a more representative picture of public discourse at this critical time.

6.1 Theoretical Approach

Recently economist Robert J. Shiller has argued for the importance of 'economic narratives' and the study of their 'powerful stories' that spread quickly and can influence market behaviours and 'real estate booms' (Shiller, Robert J., 2019b). Worcester offers a powerful example of an appealing economic narrative – something like a phoenix rising from the ashes – of a beleaguered industrial city being 'reclaimed'. Shiller stresses that economic narratives become 'contagious' and are often helped by being promoted by a celebrity (e.g., Ronald Reagan's promotion of Reaganomics) (Shiller, Robert J., 2019a). Such dynamics help demystify why some narratives may get heard more than others in the public debate about revitalisation. Furthermore, economic narratives are marketed by particular buzz words, as Neil Smith showed in his seminal study of the rise of gentrification after the 1960s. Smith tracked how, concerned by the increasingly negative connotations of 'gentrification', savvy real-estate developers appropriated the 'language of revitalization, recycling, upgrading, and renaissance' to build support for upscale development (Smith, 1996). Language, as geographers would put it, 'makes place'. Crucially, Sharon Zukin has shown how the success of an economic narrative can be due not to its broad-based appeal but to the conscious agenda of the local power elite and the media outlets that support them. Zukin studied how New York City's elite achieved the conversion of factories into loft apartments in order to create a more high-end real estate in Manhattan. She points to the strategic role the *New York Times* played through its aggressively positive reporting on that trend (Zukin, 1982). Indeed, the conventional media coverage of Worcester's revitalisation today often reads more like a boosterish advertisement than reporting (*Booming Worcester Real Estate* 2019; *Why Worcester Works* 2019).

Shiller, Smith and Zukin make it clear that if we are to be critical readers of the de-industrialising of cities towards green urban development, we must tune our ears to the narratives that guide it, the discourse that triggers it and the forces that shape it. Shiller counsels us to be on the lookout for narratives that are becoming 'contagious' in today's Canal District. Smith teaches us to be on the lookout for red-flag keywords such as 'revitalization', and 'renaissance' and to keep asking the question just what 'revitalization' means, and for whom? Zukin warns us to track just which sources and personalities dominate the

re-development discussion and suggests we try to level the playing field of public discourse by giving more attention to other less powerful local views. How does Worcester talk about itself? Many investors and future residents are coming from outside of Worcester and may encounter what Boston's newspapers or National Public Radio say about the city, rather than the city's main newspaper *The Worcester Telegram Gazette*, and other important local sources. In hopes of making this complex narrative more accessible, we explore the power of computational analysis to digest the copy generated by such a hot debate.

To accomplish this objective, we aim to apply keyword extractor algorithms. The problem of identifying keywords to track narratives within texts is long-standing (Meehan, 1976), but only recently has attracted more attention from computational linguistics (Vossen, Caselli, and Kontzopoulou, 2015; Campos et al., 2018b). With so much information made available online, getting insightful knowledge from unstructured clinical documents (Conway et al., 2019) or news articles (Martinez-Alvarez et al., 2016), to name but a few is now strictly dependent on algorithms to automate this process and reduce the effort of doing this manually. Recent advances in natural language processing (NLP) have made it possible to extract, summarise and create narratives from texts more easily than ever before (Jorge et al., 2019a; Jorge et al., 2019b). Several diverse ways of representing the overall idea of a text or group of texts exist, ranging from TF-IDF and topic modelling as introduced in Chapter 3 through to visualisation approaches, including keyword clouds (Martinez-Alvarez et al., 2016), visual storytelling (Jorge et al., 2019a), and timeline summarisation (McCreadie et al., 2018; Pasquali et al., 2019). Extracting relevant keywords from texts may be one such potential solution. In this work, we aim to apply YAKE! keyword extractor[1] to a set of texts about Worcester's Canal District to see if this tool can help identify the most important topics and keywords of the input text, without actually having to read the whole document, which, even in the case of a small story like the development of one neighborhood, becomes less and less feasible, due to the vast amount of information that is available today (both digitised historic material and current digital media).

6.2 Sources

A wide survey of documents about the Canal District was made using online search engines, including archive.org (digitised historical documents), Academic Search Premier, Nexis Uni, Proquest, and the digital archives of the *New York Times*, *Boston Globe*, *Worcester Telegram Gazette (WTG)* and *MassLive.com*. Google searches were useful for identifying other Worcester media sources such as the Worcester's alternative digital newspaper *InCity Times*; and its local magazine's *Vitality*, *Worcester Magazine*, and *Worcester Business Journal*. Our search was helped by Sarah Luria having already been

[1] http://yake.inesctec.pt

engaged in researching the Canal District's development history and uncovered sources.

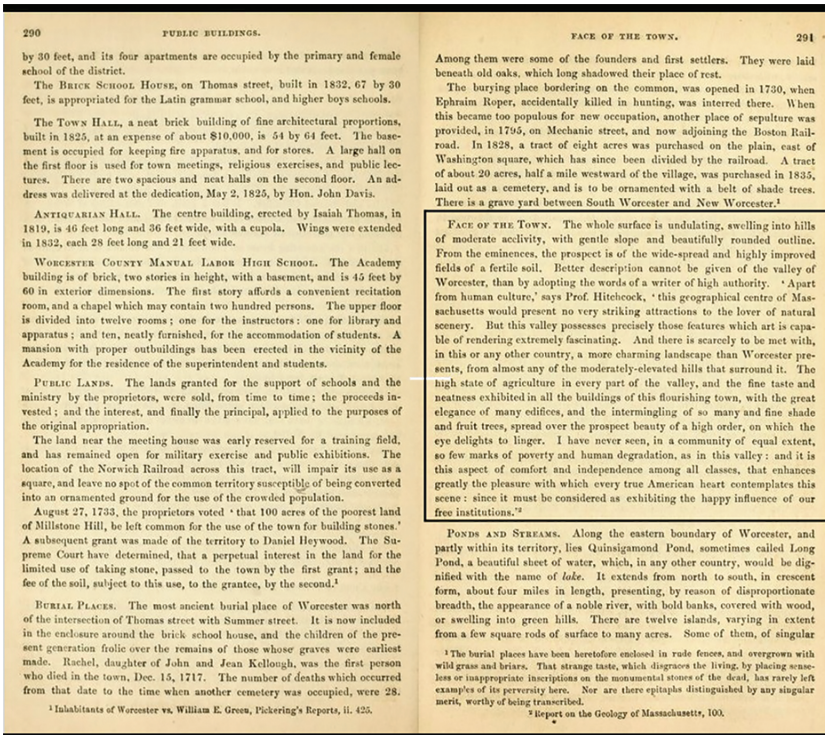
For the purpose of this experiment, we limited our corpus to 26 English-language texts that describe the Canal District over time. We felt 26 texts was a sufficient sampling to analyse and still be able, with some confidence, to make conclusions about YAKE!'s efficacy from one humanist user's point of view and yield some interesting results. If we were to continue the experiment, the corpus could be much larger. The shortest text of our curated corpus contained 72 tokens and the longest one, 4165 tokens. Titles of texts, which often summarise its argument, were included in our extracts. Texts were selected to represent some important features of the history of the District, including the creation of the Blackstone Canal, descriptions of its Irish working-class neighborhood, the city's postindustrial decline, and some of the stages in its efforts toward revitalisation. The majority of the texts are from 2018–2019, an intense period of revitalisation efforts, but examples from 1862, 1917, the 1980s and 1990s provide some historic range. A range of voices was sought and include past and current residents, renters and property owners, city leaders (mayor, city manager, city councilors, community activists), and local and outside developers. While most come from the major local newspaper *The Worcester Telegram Gazette*, the corpus includes articles from *InCity Times*, *Vitality*, *Worcester Magazine*, *Worcester Business Journal*, the *New York Times*, *Boston Globe* and *National Public Radio*. Included in this corpus is an excerpt from an acclaimed historical study, which describes the Irish kitchen barrooms of the Canal District, and a 1984 poem by Worcester-born Mary Fell[2], which describes the neighborhood during its decline.

6.3 Method

Responding to the need to deal with today's abundance of information, researchers have increasingly resorted to computer science as a means to extract, understand and create meaningful stories from large samplings of texts. This is easiest in digitally born documents, where the data is directly available ready to be processed, but offers some additional challenges in analog texts, as is the case of this project, which includes scanned historical documents not captured as plain text. Figure 6.1 shows a sample of one of our texts.

In our corpus, texts were in five different formats, including images, PDFs with images, PDFs with plain text, MS Word documents and text documents. Extracting information from the first two types implies a pre-processing stage that involves the use of Optical Character Recognition (OCR), a machine learning technique used to transform images that contain text (e.g., old text that has been scanned, handwritten, typed, etc.) into text itself. In these cases, we resorted to *tesseract*, an open-source OCR package developed by Google, and to

[2] Stanzas from the poem "The Prophecy" by Mary Fell have been used with permission of the author. All rights reserved for all elements of the poem.



290 PUBLIC BUILDINGS.

by 30 feet, and its four apartments are occupied by the primary and female school of the district.

The Brick School House, on Thomas street, built in 1832, 67 by 30 feet, is appropriated for the Latin grammar school, and higher boys schools.

The Town Hall, a neat brick building of fine architectural proportions, built in 1825, at an expense of about $10,000, is 54 by 64 feet. The basement is occupied for keeping fire apparatus, and for stores. A large hall on the first floor is used for town meetings, religious exercises, and public lectures. There are two spacious and neat halls on the second floor. An address was delivered at the dedication, May 2, 1825, by Hon. John Davis.

Antiquarian Hall. The centre building, erected by Isaiah Thomas, in 1819, is 46 feet long and 36 feet wide, with a cupola. Wings were extended in 1832, each 28 feet long and 21 feet wide.

Worcester County Manual Labor High School. The Academy building is of brick, two stories in height, with a basement, and is 45 feet by 60 in exterior dimensions. The first story affords a convenient recitation room, and a chapel which may contain two hundred persons. The upper floor is divided into twelve rooms; one for the instructors: one for library and apparatus; and ten, neatly furnished, for the accommodation of students. A mansion with proper outbuildings has been erected in the vicinity of the Academy for the residence of the superintendent and students.

Public Lands. The lands granted for the support of schools and the ministry by the proprietors, were sold, from time to time; the proceeds invested; and the interest, and finally the principal, applied to the purposes of the original appropriation.

The land near the meeting house was early reserved for a training field, and has remained open for military exercise and public exhibitions. The location of the Norwich Railroad across this tract, will impair its use as a square, and leave no spot of the common territory susceptible of being converted into an ornamented ground for the use of the crowded population.

August 27, 1733, the proprietors voted 'that 100 acres of the poorest land of Millstone Hill, be left common for the use of the town for building stones.' A subsequent grant was made of the territory to Daniel Heywood. The Supreme Court have determined, that a perpetual interest in the land for the limited use of taking stone, passed to the town by the first grant; and the fee of the soil, subject to this use, to the grantee, by the second.[1]

Burial Places. The most ancient burial place of Worcester was north of the intersection of Thomas street with Summer street. It is now included in the enclosure around the brick school house, and the children of the present generation frolic over the remains of those whose graves were earliest made. Rachel, daughter of John and Jean Kellough, was the first person who died in the town, Dec. 15, 1717. The number of deaths which occurred from that date to the time when another cemetery was occupied, were 28.

[1] Inhabitants of Worcester vs. William E. Green, Pickering's Reports, ii. 425.

FACE OF THE TOWN. 291

Among them were some of the founders and first settlers. They were laid beneath old oaks, which long shadowed their place of rest.

The burying place bordering on the common, was opened in 1730, when Ephraim Roper, accidentally killed in hunting, was interred there. When this became too populous for new occupation, another place of sepulture was provided, in 1795, on Mechanic street, and now adjoining the Boston Railroad. In 1828, a tract of eight acres was purchased on the plain, east of Washington square, which has since been divided by the railroad. A tract of about 20 acres, half a mile westward of the village, was purchased in 1835, laid out as a cemetery, and is to be ornamented with a belt of shade trees. There is a grave yard between South Worcester and New Worcester.[1]

Face of the Town. The whole surface is undulating, swelling into hills of moderate acclivity, with gentle slope and beautifully rounded outline. From the eminences, the prospect is of the wide-spread and highly improved fields of a fertile soil. Better description cannot be given of the valley of Worcester, than by adopting the words of a writer of high authority. 'Apart from human culture,' says Prof. Hitchcock, 'this geographical centre of Massachusetts would present no very striking attractions to the lover of natural scenery. But this valley possesses precisely those features which art is capable of rendering extremely fascinating. And there is scarcely to be met with, in this or any other country, a more charming landscape than Worcester presents, from almost any of the moderately-elevated hills that surround it. The high state of agriculture in every part of the valley, and the fine taste and neatness exhibited in all the buildings of this flourishing town, with the great elegance of many edifices, and the intermingling of so many and fine shade and fruit trees, spread over the prospect beauty of a high order, on which the eye delights to linger. I have never seen, in a community of equal extent, so few marks of poverty and human degradation, as in this valley: and it is this aspect of comfort and independence among all classes, that enhances greatly the pleasure with which every true American heart contemplates this scene: since it must be considered as exhibiting the happy influence of our free institutions.'[2]

Ponds and Streams. Along the eastern boundary of Worcester, and partly within its territory, lies Quinsigamond Pond, sometimes called Long Pond, a beautiful sheet of water, which, in any other country, would be dignified with the name of *lake*. It extends from north to south, in crescent form, about four miles in length, presenting, by reason of disproportionate breadth, the appearance of a noble river, with bold banks, covered with wood, or swelling into green hills. There are twelve islands, varying in extent from a few square rods of surface to many acres. Some of them, of singular

[1] The burial places have been heretofore enclosed in rude fences, and overgrown with wild grass and briars. That strange taste, which disgraces the living, by placing senseless or inappropriate inscriptions on the monumental stones of the dead, has rarely left examples of its perversity here. Nor are there epitaphs distinguished by any singular merit, worthy of being transcribed.

[2] Report on the Geology of Massachusetts, 100.

Figure 6.1: Lincoln, William. History of Worcester, Massachusetts, from its earliest settlement to September, 1836 : with various notices relating to the history of Worcester County. Worcester, 1862, p. 291. Accessed through archive.org 14 Oct. 2019. Framed text shows material excerpted for corpus selection.

some Python libraries that ease the extraction process. Afterward, we conducted a curated process to remove noisy information and clean-up the text, thus guaranteeing the quality of the data extracted. Then, we applied a keyword extractor system to capture the fundamental idea of the documents. Typically, keyword extractors make available a shortlist of relevant keywords (with one or more terms, and not necessarily the most frequent ones), thus instantaneously providing users and machines with a summary of the document. In an era where most of the information available is unstructured, having one such tool may be very appealing for those interested in quickly getting a sense of a document and in extracting insightful knowledge. Obviously, defining whether a term is a relevant keyword is itself a complex problem that may depend on the use, case, user background or application, such that, as referred by Sterckx et al. (2016), reaching a consensual list of keywords by two different persons for the same document turns out to be a very difficult task. For this purpose,

several different techniques have been proposed over the years, from statistical methods that detect keywords based on statistical features (El-Beltagy and Rafea, 2009), to graph-based models (Mihalcea and Tarau, 2004) or machine-learning approaches (McCreadie et al., 2018), a supervised solution that learns from previous examples. A complete survey on keyword extraction algorithms may be found in a recent work of Papagiannopoulou and Tsoumakas (2020).

We grew to realise that for this experiment, Ricardo, a computer scientist, and Sarah, a literature professor, first needed to clarify what each meant by 'keyword'. Indeed, our research went in rather humorously opposite directions until we finally realised the need to back-up and start again by agreeing on definitions of the terms that would guide our research. Note to future interdisciplinary collaborators – define your terms at the outset! Do not assume each of you shares the same ideas about what words mean. To arrive at a shared meaning of 'keyword', Ricardo offered the helpful example of keywords that publishers ask us to assign to our published articles. These words are usually "subject" words that convey the topics of a text. In our corpus, these include names of important places and people in the article, such as 'Canal District', and 'Edward Augustus' (Worcester's City Manager), and important topics, such as 'revitalization', 'gentrification', and 'affordable housing'. Sarah pointed out that since we sought through a keyword extraction algorithm to create a 'summary' of a text, keywords needed to include not just the subject of the article but what was being said about that subject, which she termed the argument, or main point, of the text. Thus, if the article was about the fast rising cost of real estate in the Canal District, keywords would be 'subject' words like 'real estate' and 'Canal District' but also 'argument' words like 'booming' and 'properties are hot' and 'only just beginning'. We thus concluded that for the purpose of this experiment, keywords = subject + argument words.

In this analysis, our chosen tool was YAKE! (Campos et al., 2018a; Campos et al., 2020), an unsupervised statistical keyword extractor method that has demonstrated success in tackling documents from different languages, domains and length, without the need for prior knowledge. Our purpose is to understand whether this kind of algorithm may be used in the context of geography, which focuses on the study of place, to quickly create a flow of stories from a set of documents collected over time and if they help the reader survey the topic being discussed, without the need to acquire further knowledge. With this in mind, we resort to YAKE! Python package[3] to automatically extract the relevant keywords from the set of texts, where a keyword may be a single word or a group of *n* terms (known as keyphrases).

In this experiment, Sarah was offered five different lists of top-40 relevant YAKE! keywords of our corpus texts with different *n* settings, namely $n = \{1, 2, 3, 5, 10\}$. Ricardo aimed to offer Sarah the chance to compare the results for

[3] https://github.com/LIAAD/yake

the different *n*'s and to see if the effectiveness of their summaries varied to a significant extent in this varied array of texts. She concluded that by and large $n = 5$ and $n = 10$ did not significantly increase the information that $n = 3$ was able to convey about the text. This was a surprise to Sarah, since she had assumed that the more words one extracted, the better the chances that a good summary would be produced. Instead, she discovered that $n = 5$ and $n = 10$ increased the chance for more 'noise' – more little words – that made the summary less clear, and sometimes even inaccurate.

One example illustrates this well. The text is an excerpt from historian Roy Rosenzweig's 1985 study of industrial working-class life in Worcester during its industrial heyday. It describes the kitchen breweries run by Irish women in the Canal District. The excerpt concludes 'It is **unlikely** that these kitchen barrooms were especially lavish or spacious since they shared the physical limitations of most working-class dwellings of this period' (Rosenzweig, 1983). At $n = 5$ YAKE! extracts 'kitchen barrooms were especially lavish', which is the opposite point that the original sentence made. At $n = 10$ YAKE! extracts 'kitchen barrooms were especially lavish or spacious'. At $n = 3$, however, YAKE! extracts the important subject of the 'kitchen grog shops' but does not falsely couple them with the modifiers 'especially lavish [or spacious]'. That said, $n = 3$ does extract (from elsewhere in the passage) the words 'formal and elegant'. The unwary reader could make a false connection in her mind that the 'kitchen grog shops' were 'formal and elegant', which would misrepresent the text"s main point. But it is important that at $n = 3$ YAKE! does not itself falsely connect the two phrases into one keyword phrase as it does at $n = 5$ and $n = 10$.

Based on such results, we opt to define $n = 3$ as our safer and preferred setting, which is in line with the work of Campos et al. (2018, 2020), who pointed out that the most effective results are obtained when the number of grams, that is, the number of terms of a keyword, is set to a maximum of three terms (e.g., 'roads', 'Worcester railroad', 'large manufacturing city').

6.4 Interpretations and Results

After Ricardo processed the corpus texts using YAKE!, the results were then interpreted by Sarah, a literary geographer, who created a gold standard dataset for comparison. Sarah read the original text samples, identified what she considered to be keywords and phrases, and compared them to what YAKE! extracted as an automatic result. Each set of YAKE! extracted keywords was then classified by Sarah as having good, sufficient or insufficient results. In 'good' results, YAKE! extracted enough of Sarah's subject and argument keywords so that the main point of the article is conveyed. In 'sufficient' results, YAKE! extracts some subject keywords and one, or a few, argument words so that the main point of the argument is fairly clear. In 'insufficient' results, YAKE! extracts some subject keywords but no, or not enough, argument key words, so that the argument is not clear.

Good* = 15	Sufficient** = 8	Insufficient*** = 3
1837 Lincoln, History of Worcester[4]	1879 Marvin, History of Worcester County [re Canal and Railroad][5]	1998 Green Island Businesses say city help is killing them
1917 Washburn, Blackstone Canal formed[6]	1985 Rosenzweig, from Eight Hours for What We Will	2000 Green Island revitalization dropped
1917 Washburn, Worcester's entrepreneurial spirit[7]	1999 Vacant Industrial Sites of no use to neighborhood	2019 An Away Game for Businesses[8]
1917 Washburn, importance of steam power[9]	2011 Life in Green Island: We have hope[10]	
1983 Worcester Shedding Smokestack Image[11]	2018 Time to talk gentrification in Worcester[12]	
1984 Fell, The Prophecy[13]	2018 NPR Story Worcester the new It Town[14]	

Table 6.1: Sarah's Evaluation of YAKE! Results – sources linked, where available, online in footnotes.

How many of Sarah's key terms (or close variations of them) did YAKE! catch? Using the above-referred grading scale, Sarah concluded that YAKE!'s summaries of the corpus texts were good in fifteen cases, sufficient in eight and insufficient in three (Table 6.1).

---

[4] https://archive.org/details/historyofworcest00inlinc/page/290/mode/2up
[5] https://archive.org/details/historyofworcest03marv/page/83/mode/2up
[6] https://archive.org/details/historyofworcest03marv/page/83/mode/2up
[7] https://archive.org/details/industrialworces00wash/page/300/mode/2up
[8] https://www.telegram.com/story/news/local/worcester/2019/07/06/away-game-for-worcester-property-owners-facing-ballpark-redevelopment/4743703007/
[9] https://archive.org/details/industrialworces00wash/page/30/mode/2up
[10] https://incitytimesworcester.org/tag/millbury-stree/
[11] https://www.nytimes.com/1983/09/25/us/worcester-shedding-smokestack-image.html
[12] https://www.worcestermag.com/news/20181011/feature-time-to-talk-about-gentrification-in-worcester
[13] https://capa.conncoll.edu/fell.persistence.html#38
[14] https://www.npr.org/2018/10/23/658263218/forget-oakland-or-hoboken-worcester-mass-is-the-new-it-town

Good* = 15	Sufficient** = 8	Insufficient*** = 3
1989 New focus for an old area[15]	2019 Worcester pledges $3M to Green Island[16]	
1997 Bureau urges liability relief for brownfields	2019 Worcester Organizers hear from Nashville…[17]	
2007 Canal District Shapes Up[18]		
2018 WooSox Ball Park has a long history[19]		
2018 A City Reclaimed		
2019 A Totally Cool Place to Live[20]		
2019 New Shine for old building		
2019 Worcester gets brownfield funds[21]		
2019 Renee Diaz, WooSox killing [Canal] district dreams[22]		

Table 6.1: (*continued*).

[15] https://www.nytimes.com/1989/12/10/realestate/national-notebook-worcester-mass-a-new-focus-for-an-old-area.html

[16] https://www.telegram.com/story/news/local/worcester/2019/04/30/worcester-pledges-3m-to-green-island-neighborhood-vows-it-wont-be-overshadowed-by-ballpark/5307782007/

[17] https://www.worcestermag.com/story/news/2019/02/28/worcester-organizers-hear-from-nashville-buffalo-for-tips-on-woosox-cba-push/5802104007/

[18] https://www.telegram.com/story/news/local/east-valley/2007/09/25/canal-district-shapes-up/52786534007/

[19] https://www.bostonglobe.com/business/2018/08/17/new-home-for-woosox-has-long-history/oUqIGARpKusD1NrNItDIaI/story.html

[20] https://www.masslive.com/worcester/2019/04/a-totally-cool-place-to-live-allen-fletcher-offers-sneak-peek-inside-new-kelley-square-lofts-bringing-48-high-end-units-to-worcesters-canal-district.html

[21] https://www.telegram.com/story/news/local/worcester/2019/06/23/worcester-gets-federal-money-for-brownfields-cleanup/4846706007/

[22] https://www.wbjournal.com/article/construction-woosox-regulation-are-killing-canal-district-dreams

These positive results encourage us to think that YAKE! has the potential to serve as an effective summarising tool. Of course, Sarah's "algorithm" or grading rubric was not as precisely or clearly formulated as YAKE!'s; nevertheless, Sarah did find that in the majority of cases YAKE! did extract enough of what she deemed keywords to present a good or sufficient summary of a text. These findings pointed to several primary ways we might hope YAKE! results could be made increasingly informative by tweaking YAKE!s algorithm so it could more closely match Sarah's results. We discuss these briefly below.

As stated above, Sarah generated her list of keywords (subject + argument) for each text in the corpus. Table 6.2 shows one example that Sarah ranked 'good,' which YAKE! generated from Mary Fell's poem 'The Prophecy' (1984). The poem is written from the point of view of a long-term resident in the neighbourhood. Despite the dominant Worcester-narrative of the neighbourhood is in decline, the poem's speaker argues nevertheless that the 'neighborhood

Sarah's Keywords Extracted by YAKE!	**Sarah's Keywords Missed by YAKE!**
Jews settling	same old bars
Green island	big stories
Remembering the canal	immigrants
Learn polish prayers	
Irish laborers	
Patsy spoke	
Aggie	
Catholic school	
Whiskey	
Aggie brew	
The Neighborhood remains	
Kids	
Built by Irish	
Canal that cut	
Persistence of memory	
prophecy	
Millbury and Harding Streets	
made beer	
American born	
Kelly Square	

Table 6.2: 'Good' Rating: Subject + Argument Words Extracted.

remains'. As Table 6.2 shows, YAKE! extracted almost all of the keywords identified by Sarah.

YAKE!'s results sufficiently capture Fell's emphasis on the ethnic diversity of the canal district– 'jews', 'learn polish prayers', 'irish laborers', 'patsy', 'aggie' (both Irish names), 'catholic school' – as well as her focus on the neighborhood's drinking culture and the 'whiskey', which was 'aggie's brew'. The argument of the poem is also conveyed that 'kids' are still a part of the neighborhood, and 'learn polish prayers', which is in the present tense. A sense of the past is also present through the words 'built by irish', 'remembering the canal', the 'canal that cut' [through this neighborhood], and the 'persistence of memory' (the title of the collection in which the poem appeared). While it is unclear what 'prophecy' means in this context (it is the title of the poem), the word helps convey the weighty, assertive tone of the poem and the suggestion of an (unspecified) future for the neighborhood.

In Sarah's 'sufficient' examples, YAKE! extracted some subject and argument words but omitted keywords critical to the piece. Table 6.3 shows the results for the National Public Radio story 'Forget Oakland or Hoboken, Worcester, Mass. is the New "It" Town' (Schacter, Aaron, 2018a). As the title of that story suggests, this is exactly the sort of discourse that could fan powerful economic narratives that could overwhelm local efforts to achieve a green Worcester for everyone. YAKE!'s words suggest that Worcester is booming, but misses the article's argument that the new real estate boom is just beginning and that local residents are concerned about what this might do for the identity of Worcester.

To see if such minimal summarising happens with other keyword extractors, we ran the very same text under IBM Watson, one of the most well-known commercial solutions. Interestingly, we found that the top-10 keywords retrieved by the IBM system[23] also did not extract the keywords Sarah thought were critical to the piece, demonstrating that there is still much work to do text understanding.

Finally, Table 6.4 shows an example Sarah rated 'Insufficient'. The text is a *Worcester Telegram Gazette* article (7-6-2019) 'An Away Game for Businesses: Property Owners Near Ballpark Make Way for Redevelopment' (*An Away Game for Businesses: Property Owners Near Ballpark Make Way for Redevelopment* 2019). Here YAKE! extracts subject words, but no argument words. As a result, the main point of the article does not come through, which is that tenants have mixed feelings about being displaced.

From Sarah's perspective, two main problems emerged with YAKE!'s results. One was that YAKE! repeats keywords in a text, often several times. In the example from National Public Radio above (Table 6.3), city names (Worcester, San Francisco, Oakland) are repeated and dominate the results (similarly, 'city' was

[23] 'smaller cities', 'largest city', 'worst thing', 'expensive city prices people', 'small city', 'cities', 'larger city', 'piece of Worcester', 'City officials' and 'lot of times'.

Sarah's Keywords Extracted by YAKE!	Sarah's Keywords Missed by YAKE!
forget Oakland	New "it" town
Worcester booming industrial	properties are hot
Hoboken	incredibly cheap
beautifying Worcester common	destination for foodies
rise	celebrity chefs
second-largest city	master brewers
expensive city prices	former factories
smaller city	mill buildings
Boston	waiting to die
piece of Worcester	stay away
Massachusetts housing alliance	Union Station
fifth-generation Worcesterite	decrepit and roofless
	only just beginning
	still untapped potential
	sky's the limit
	see everyone succeed
	success begets success
	community spirit
	critical tipping point
	most vulnerable residents
	young families out
	elders out
	crisis
	cash investors
	flip it
	genuinely care about
	maintaining its character
	where immigrants come

Table 6.3: 'Sufficient' Rating: Subject + Some Argument Words Extracted.

repeated seven times out of 10 by IBM Watson's processing of this text, cited above). Sarah would have liked to have seen a list with a greater variety of keywords to create a fuller picture of the text's main point. This is acknowledged by Ricardo as a drawback of YAKE! that deserves further attention in the future. One of the possibilities is to apply a more elaborated deduplication algorithm.

Sarah's Keywords Extracted by YAKE!	Sarah's Keywords Missed by YAKE!
Worcester redevelopment authority	taking properties
properties	eminent domain
Washington street	moving out
canal district	complicated plan
Pawtucket red sox	business like transaction
tenant	highly contentious
	lawsuits
	not to sue
	awards well above
	reconstruct Kelley Square

Table 6.4: 'Insufficient' Rating: Subject but No Argument Words Extracted.

Another is to apply some topic modelling algorithm so that a more diversified set of keywords is shown to the users of YAKE!

The second problem was cited above, in the example from Rosenzweig's description of the kitchen grog shops of Worcester, where the key modifying phrase 'It is unlikely' was missed by YAKE!. Such omissions occurred only in a few instances but are significant. In this case, even at $n = 3$ YAKE!'s juxtaposition of keywords invites the reader to make a connection that the barrooms were 'formal and elegant', which misrepresents the text's main point. One of the reasons for this is that YAKE! is a purely statistical keyword extractor that uses any linguistic features such as part of speech. While this keeps the system mostly language-independent and easily adaptable to different languages (a plug and play feature), it makes it difficult to understand the structure and the idea of the text in a more principled way.

YAKE!'s greatest communicative power as a summarising tool is its ability to generate word clouds. The larger the word in the cloud, the higher YAKE! ranked it in importance. Figure 6.2 shows the word cloud YAKE! generated of Mary Fell's poem 'Prophecy' ($n = 3$).

Word clouds produce a pretty picture from a text, but do these have any real value? Given our research's emphasis on discourse and the importance of certain keywords in shaping development narratives, we appreciate that all word clouds communicate the importance of individual words and small phrases immediately. A word cloud is reductive, yet many important subject and argument words from Fell's poem are included here.

This ability of word clouds to concentrate the messages of a large number of texts, while highlighting the importance of repeated keywords that shape public discourse, encourages our initial aim to survey more voices on the Canal District's development. Figure 6.3 shows a fanciful experiment to present the

Figure 6.2: YAKE! word cloud (n = 3) Mary Fell, The Prophecy, from The Persistence of Memory (1984). Github filename: 7 - 1984_Mary Fell, The Prophecy (Poem).

word clouds of our entire corpus as a scroll to depict a narrative and image of some of Worcester's development discourse over time.

Given the fairly good results of YAKE!'s analysis, such a scroll might make a compelling tool for study. Further refinements to YAKE!'s algorithm could yield a still more useful tool. Key here is that such a scroll engages the reader in a more extended narrative of Worcester's change over time, that includes the city as a still agrarian landscape, the creation of its canal, railroad and factories, and its long efforts to clean up their industrial waste. It creates a broader understanding of Worcester's rich immigrant history to counter the shallow media buzz of it currently being 'hot'. 'Something good is happening in Worcester' is becoming a dramatic developer's mantra for the city (*Booming Worcester Real Estate* 2019). Ultimately the ability to digest and exhibit more sources and views through a supple keyword extractor might produce a more democratic and fuller understanding of just what that 'something' is.

6.5 Discussion of Implications, Limitations, and Links to Other Work

Recent scholarship has argued that in order to better assess and mitigate connections between revitalisation and gentrification, we need to analyse each

place and its particular situation and challenges. We need to counter the reductive booster narratives of 'green revitalization' generated by the market with more representative, fine-grained accounts to get closer to the truth of what is going on in the streets of these cities (Saunders, 2018). Such diverse local accounts might help undercut the reification of gentrification as an unstoppable economic narrative.

City planning best practice now acknowledges that resident input is essential to any development's success (Myerson, Deborah L., 2004). More and more venues both live and digital are opening up for local residents to speak their needs and visions and play a part in their town's redevelopment[24]. Such input can be strengthened by an increased awareness of the history of the conversation surrounding a neighborhood's development and key terms that have been used to shape it. Such input could also importantly be tracked, and amplified, by tools like YAKE!

Do our results show a more representative and complex picture of what Worcester thinks about 'revitalization'? We think so. Even from this small sampling, our YAKE! word clouds show that city leaders and local residents have been trying to revitalise the city for a long time, in many admirable homegrown ways. Most importantly, they show that the new baseball stadium increasingly dominates Worcester's public conversation about its revitalisation (see Figure 6.3. 2018, *BG*; 2018, *WM*; 2019-2-27 *WTG*, 2019 6-1-*WTG*, 2019-6-20, *WBJ*, 2019-24-6 *WTG*, 2019-24-6 *WTG* 2). Since this article was written the stadium has opened, suggesting that it will take up an increasingly large share of the discursive landscape. Our brief survey suggests too that this will prompt more positive and perhaps increasingly critical discourse from Worcester's residents about the ballpark's development (see Figure 6.3. 2018, *WM*; 6-20-2019, *WBJ*).

Given the insights registered by Shiller, Smith and Zukin, cited above, when it comes to public discourse that fuels economic narratives, we could conclude from our study that some regulation is in order to level the playing field of just who gets heard. Right now, Worcester – its Mayor and City Manager, local press and residents – seem to concur in the desire to 'keep Worcester Worcester'. But public discourse is like a busy traffic intersection. Without a smart traffic light in the center, the largest trucks will succeed in barrelling through, and the local pedestrians and bicyclists, whom many want to encourage, never get a chance to be seen and cross. The Ballpark's development makes good media copy, and it may signal a new phase of Worcester-making that comes, even more than in the past, increasingly from outside rather than local energies. In our furthest reaching reflections from this experiment, we wonder if regulating the flow of discursive traffic through an unsupervised keyword extractor could help Worcester better hear itself think and so navigate its growth through this critical time.

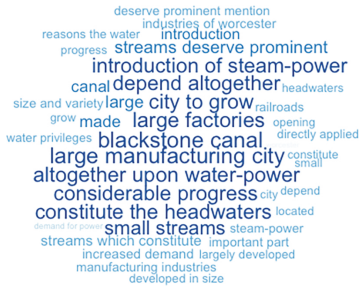
---

[24] CoUrbanize. OnLine Community Engagement. https://www.courbanize.com/

Our turn to an unsupervised keyword extractor in an effort to track more voices is reductive, eclectic and selective, certainly, but we believe these aspects are strengths as well as weaknesses. YAKE! word clouds produce a quick picture, but they also invite one to linger, fathom connections, pay attention to language and survey a discussion's key players and keywords. Thus this project's collaboration may suggest some hopeful first results for one possible approach to create a large engaging canvas of a community conversation that includes voices from the past and present and also highlights the role of language in the ongoing remaking of place. Furthermore, we believe that our collaboration stresses the need for computer scientists and humanists to continue working together to refine an unsupervised keyword extractor like YAKE! to consistently identify significant keywords and produce useful summaries from a wide array of sources on an important question. The potential for such a tool could be significant indeed. In the case of Worcester's Canal District, diversifying the narrative of development seems crucial in the public arbitration over the ethical preservation of history and the greening of place.

Lincoln, History of Worcester, 1862

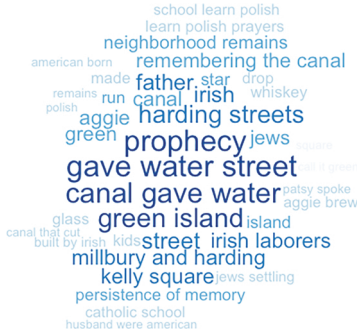
Washburn, History of Worcester, 1917



Washburn, History of Worcester, 1917

New York Times, 1983

Figure 6.3: YAKE! word cloud (n = 3) scroll of corpus.



Mary Fell, 1984

Roy Rosenzweig, 1985

New York Times, 1989

WTG, 1997

WTG, 1999

WTG, 2000

Figure 6.3: (*continued*). YAKE! word cloud (n = 3) scroll of corpus.

WTG, 2007

InCity Times, 2011

Boston Globe, 2018

Vitality Magazine, 2018

Worcester Magazine, 2018

National Public Radio, 2018

Figure 6.3: (*continued*). YAKE! word cloud (n = 3) scroll of corpus.

WTG, 2019-2-27

MassLive, 2019

WTG, 2019-6-1

WBJ, 2019-6-20

WTG, 2019-24-6

WTG, 2019-24-6

Figure 6.3: (*continued*). YAKE! word cloud (n = 3) scroll of corpus.

References

An Away Game for Businesses: Property Owners Near Ballpark Make Way for Redevelopment (2019).

Booming Worcester Real Estate (2019). URL: https://www.wcvb.com/article/booming-worcester-real-estate/25741906#.

Campos, Ricardo, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt (2018a). "A text feature based automatic keyword extraction method for single documents". In: *European Conference on Information Retrieval*. Berlin: Springer, pp. 684–691. DOI: 10.1007/978-3-319-76941-7_63.

Campos, Ricardo, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt (2018b). "YAKE! collection-independent automatic keyword extractor". In: *Advances in information retrieval*. Vol. 10772. Grenoble: Lecture Notes in Computer Science, pp. 806–810. DOI: 10.1007/978-3-319-76941-7_80.

Campos, Ricardo, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Celia Nunes, and Adam Jatowt (2020). "YAKE! Keyword extraction from single documents using multiple local features". In: *Information Sciences* 509, pp. 257–289. DOI: 10.1016/j.ins.2019.09.013.

Conway, Mike, Salomeh Keyhani, Lee M. Christensen, Brett R. South, Marzieh Vali, Louise C. Walter, Danielle L. Mowery, Samir E. AbdelRahman, and Wendy W. Chapman (2019). "Moonstone: A novel natural language processing system for inferring social risk from clinical narratives". In: *Journal of Biomedical Semantics* 10. DOI: 10.1186/s13326-019-0198-0.

El-Beltagy, Samhaa R and Ahmed Rafea (2009). "KP-Miner: A keyphrase extraction system for English and Arabic documents". In: *Information Systems* 34.1, pp. 132–144. DOI: 10.1016/j.is.2008.05.002.

Fell, Mary (1984). "The prophecy". In: *The persistence of memory*. New York: Random House.

Jorge, Alípio M, Ricardo Campos, Adam Jatowt, and Sérgio Nunes (2019a). *Information Processing & Management Journal Special Issue on Narrative Extraction from Texts (Text2Story): Preface*. DOI: 10.1016/j.ipm.2019.05.004.

Jorge, Alípio Mário, Ricardo Campos, Adam Jatowt, and Sumit Bhatia (2019b). "Second international workshop on narrative extraction from texts (Text2Story'19)". In: *Advances in Information Retrieval*. Vol. 11438. Cologne: Lecture Notes in Computer Science, pp. 389–393. DOI: 10.1007/978-3-030-15719-7_54.

Martinez-Alvarez, Miguel, Udo Kruschwitz, Gabriella Kazai, Frank Hopfgartner, David Corney, Ricardo Campos, and Dyaa Albakour (2016). "First international workshop on recent trends in news information retrieval (NewsIR'16)". In: *European Conference on Information Retrieval*. Berlin: Springer, pp. 878–882. DOI: 10.1007/978-3-319-30671-1_85.

McCreadie, Richard, Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis (2018). "Explicit diversification of event aspects for temporal summarization". In: *ACM Transactions on Information Systems (TOIS)* 36.3, pp. 1–31. DOI: 10.1145/3158671.

Meehan, J. (1976). *The Metanovel: Writing stories by computer*.

Mihalcea, Rada and Paul Tarau (2004). "Textrank: Bringing order into text". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411.

Myerson, Deborah L. (2004). *Involving the Community in Neighborhood Planning*. URL: http://uli.org/wp-content/uploads/2012/07/Report-1-Involving-the-Community-in-Neighborhood-Planning.ashx_.pdf.

Papagiannopoulou, Eirini and Grigorios Tsoumakas (2020). "A review of keyphrase extraction". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.2, e1339. DOI: 10.1002/widm.1339.

Pasquali, Arian, Vítor Mangaravite, Ricardo Campos, Alípio Mário Jorge, and Adam Jatowt (2019). "Interactive system for automatically generating temporal narratives". In: *European Conference on Information Retrieval*. Berlin: Springer, pp. 251–255. DOI: 10.1007/978-3-030-15719-7_34.

Rosenzweig, Roy (1983). *Eight hours for what we will: Workers and leisure in an industrial city 1870-1920*. Cambridge: Cambridge University Press.

Saunders, Pete (2018). "The scales of gentrification". In: *Planning* 84 (11), pp. 16–23. DOI: 10.1044/leader.BGJ.23112018.16.

Schacter, Aaron (2018a). *Forget Oakland or Hoboken, Worcester, Mass. is the New It Town*. https://www.npr.org/2018/10/23/658263218/forget-oakland-or-hoboken-worcester-mass-is-the-new-it-town. [Online; accessed 05-May-2020].

— (2018b). *It's Time to Talk About Gentrification in Worcester*. URL: https://www.worcestermag.com/news/20181011/feature-time-to-talk-about-gentrification-in-worcester.

Shiller, Robert J. (2019a). *Narrative economics*. Princeton, N.J.: Princeton U Press.

— (2019b). *What People Say about an Economy Can Set of a Recession*. [Online; accessed 05-May-2020]. URL: https://www.nytimes.com/2019/09/12/business/recession-fear-talk.html.

Smith, Neil (1996). *The New Urban Frontier: Gentrification and the Revanchist City*. New York: Routledge.

Sterckx, Lucas, Thomas Demeester, Chris Develder, and Cornelia Caragea (2016). "Supervised keyphrase extraction as positive unlabeled learning". In: *EMNLP2016, the Conference on Empirical Methods in Natural Language Processing*, pp. 1–6. DOI: 10.18653/v1/D16-1198.

Vossen, Piek, Tommaso Caselli, and Yiota Kontzopoulou (2015). "Storylines for structuring massive streams of news". In: *Proceedings of the first workshop on computing news storylines*, pp. 40–49. DOI: 10.18653/v1/W15-4507.

Whitmore, Bernard (2018). *A Mayor, A Manager, A City Reclaimed*. [Online; accessed 17-October-2019].

Why Worcester Works (2019). URL: https://www.wcvb.com/article/why-worcester-works/25741937.

Zukin, Sharon, ed. (1982). *Loft living*. Baltimore: Johns Hopkins University Press.