Text Visualization and Close Reading for Journalism with Storifier

Nicole Sultanum, Anastasia Bezerianos, Fanny Chevalier

Text Visualization and Close Reading for Journalism with *Storifier*

Nicole Sultanum*
University of Toronto

Anastasia Bezerianos†
Univ. Paris-Sud, CNRS, INRIA, Université Paris-Saclay

Fanny Chevalier‡
University of Toronto

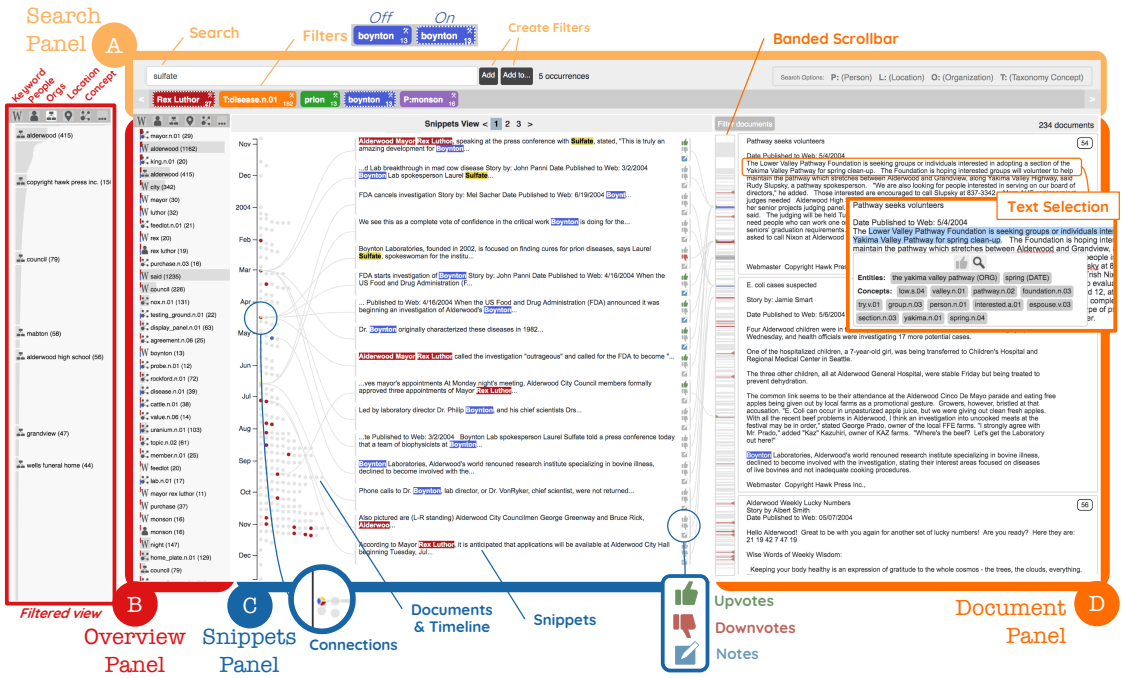

Figure 1: *Storifier*, a tool for journalistic text analysis focused on reading. The interface features (A) a *Search panel*, for keyword and entity search, (B) the *Overview panel*, listing prominent terms and entities ((B) with active filters on far left), (C) a *Snippets panel* listing a document timeline and search results, and (D) the *Document view*, listing full documents in a continuous scroll view.

Abstract

Journalistic inquiry often requires analysis and close study of large text collections around a particular topic. We argue that this practice could benefit from a more text- and reading-centered approach to journalistic text analysis, one that allows for a fluid transition between overview of entities of interest, the context of these entities in the text, down to the detailed documents they are extracted from. In this context, we present the design and development of *Storifier*, a text visualization tool created in close collaboration with a large francophone news office. We also discuss a case study on how our tool was used to analyze a text collection and helped publish a story.

1 Introduction

Data journalists often analyze large text collections when pursuing newsworthy stories about current relevant topics, such as presidential tweets [36] and political party manifestos [10]. Given the ever increasing scale of these collections and the laborious nature of text analysis, journalism practices have adopted automated text analysis to support investigative work [7, 17, 19]. While these methods constitute invaluable resources to generate quantitative insights and to guide qualitative analysis, they require significant technical expertise and do not *"replace the human judgement needed for fine-grained, qualitative forms of analysis"* [17]. This calls for a more careful emphasis and support of *reading* as a key step of the analysis workflow.

Several past works in text visualization sought to bridge this gap and provide overviews of large text collections to support journalistic practices. However, many focus on high-level summaries and on showcasing structured data extracted from text (e.g., entities, relationships, and trends), dedicating little space (or providing only indirect access) to original text [20, 26, 38]; other more text-centered tools, on the other hand, were designed for specific types of text collections, e.g., social media posts [13] and PDF documents [8], which limits their generalizability. Text-centered analysis tools in sensemaking [29, 33] and digital humanities [23, 24] could support journalists in this aspect; however, these tools tend to be fairly complex and not always amenable to the stringent timeframes and dynamic requirements of the newsroom [5]. Journalistic inquiry also has a unique focus on "newsworthy" (e.g., surprising or unusual) findings [3, 13], a notion which is hard to fully automate and therefore requires user intervention for searching and reading. We argue that a tool where both corpus-level overviews and detailed text content are readily accessible could help mitigate these issues.

In this work, we present *Storifier*, a text analysis tool for journalists centered on reading and the transitioning from corpus-level features to detailed text analysis, via a fully connected and simple-to-use interface. We report on our collaboration with journalists from a large francophone newspaper office to derive design requirements that guided our design and development. We also discuss findings of a case study where *Storifier* was used to explore (and publish a story on) a large text dataset of citizen engagement feedback.

*e-mail: nicolebs@cs.toronto.edu
†e-mail: anab@lri.fr
‡e-mail: fanny@dgp.toronto.edu

2 Background

We discuss works in text visualization for journalism and other fields that inspired our work, grouped by recurring **design themes (D1-D4)**.

(D1) Transitioning between levels of detail. In line with Shneiderman's information seeking mantra [32], an assessment of journalistic tasks around document collections [8] outlined two tasks — content summarization and hypothesis testing — which require not only the ability to identify patterns and generate overviews but also a need to effectively peruse details. This alludes to the *distant* vs. *close reading* spectrum for text analysis in the digital humanities, with many text visualization tools operating on either end of this spectrum [23,24]. We argue that a tool to support journalistic exploration of text collections should instead *bridge* the two ends — i.e., provide a high-level overview of the collection (to support discovery), along with fluid access to full-text details (for richer context).

This transitioning between levels of detail has been explicitly considered in the design of text analysis tools for literary studies that feature dedicated space for full document views plus several levels of progressive corpus-level summarization [3, 12, 25]. However, these tools are typically geared towards deep and detailed analysis of very long documents (e.g., books), whereas journalistic practices emphasize discovery, skimming, and retrieving of quotes. On the other hand, tools tailored for journalistic text analysis have only partially incorporated this notion of levels of detail: among those featuring dedicated spaces for text, they have either been designed for very short snippets (e.g., headlines, social media posts and user reviews) [13, 15, 20], or they support full document views but lack text browsing features (e.g., showcasing text snippets from multiple documents at a time) [8].

(D2) Overview and Summarization. Providing top-down awareness on a corpus level (i.e., *distant reading* techniques) enables not only to inform corpus trends and high-level patterns but also to provide entry points for further exploration [3,22]. Common approaches include topics [3,12,14,22,26], extracted entities [14,22,33,38], relevant keywords [13,15,27], and aggregate statistics over terms and metadata [15], sometimes organized over time [13,14,27]. These elements are often interactive and serve as content filters, linking back to particular text mentions. We provide similar views.

(D3) Navigation and Exploration. Reading is time-intensive, and having the means to efficiently and effectively search for relevant content is key. Past assessments found that journalists [5] and intelligence analysts [22] prefer "high recall" queries, i.e., starting with a broad search that retrieves a larger set of documents for quick browsing and later re-filtering. This implies that filters should be compoundable, allowing for multiple filtering criteria at once. This is often achieved via metadata filters [3,5,13,15,16] and entity/topic filters [12,14,22,27]. A less common but powerful approach to complement keyword search is to use ontology concepts [11,34], where a higher level hypernym may replace multiple keyword searches of alternative spellings of the same concept.

Another important step to guide discovery is lateral search, i.e., finding concepts and entities related to a topic of interest. This is often done by explicitly encoding relationships as graphs [9, 27, 33, 38], matrices [3], and proximity in scatterplots [28]. Another strategy is showcasing overlap of parallel search queries, such as associating different colours to each search filter [3, 8, 15, 25] and using small multiples [16]. In the end, the analyst still needs to have a closer look at related texts to qualify said relationships. We support both ontology-powered search and multiple search filters while allowing analysts to read text more closely.

(D4) Curation and tagging. User-defined structures are an integral component of sensemaking and intelligence analysis [29, 33] with particular relevance to journalism. From case studies with journalists, Brehmer et al. [8] found that tags and annotations were widely used to keep track of what documents were seen and what information was learned. Related strategies include tagging documents of interest [8, 22], outlining text snippets for classification [25], saving search configurations [12, 15, 28], and user-defined entities and connections [33]. In the end, few text analysis tools offer support to keep track and revisit user-defined segments of interest within documents. Our work also aims to fill this gap.

3 Formative Investigation

To help us contextualize past work and current journalistic needs, we conducted a formative investigation with staff from *Ouest France*[1], the most read francophone newspaper in the world (2.5 million daily readers). One data journalist, one information curator managing newspaper archives, and two support staff with journalistic training (who aid journalists with information search and fact-checking) participated in a half-day workshop. After a round table, participants were asked to discuss recent stories they wrote, particularly focusing on common data analysis practices and specific challenges they faced when working on these stories. Our collective discussion and reflections on their experiences led us to identify three **high-level tasks** to the exploration of text document collections **(T1-T3)**:

(T1) High-level overviews. This includes understanding the collection and identifying possible stories, keywords, or topics of interest. This task is most relevant for unfamiliar collections, e.g., when a new text corpus is released or updated. Journalists reported performing this task by skimming multiple articles of the collection, identifying important keywords/entities, and then explicitly searching for them in documents. They would then read the articles and identify further relationships or connections they may have, refining their search. Major challenges in this task were deciding where to start their research (i.e., which articles to read first), and how to narrow down the set of documents to read within a topic of interest.

(T2) Focus on entities of interest. The data journalist and support staff often investigate the trajectory of one or more entities (e.g., person, company, political party, movement) and their relationships in depth. Currently, the journalists perform this task by searching for the names of entities and related keywords and then looking over the articles to understand how they relate to each other and how they evolve over time. A major challenge is gaining such an understanding quickly, especially in face of recent emerging events that need to be commented on in a very short time frame. On the day of our visit, the data journalist was planning a piece on the suspects associated with an attack in Strasbourg that had happened a few hours before. To build a timeline for the sudden event, they only had a few hours to go over hundreds of documents via keyword search, which, apart from a handful, turned out to be uninformative.

(T3) Evidence to support hypotheses. The support staff explained that journalistic stories are often built around a starting hypothesis for which they acquire evidence to support or disprove. In the context of investigative journalism, this takes the form of *fact checking*, *quotes or statements* related to an event, *statistics*, and *understanding the context* for background information in the data and in external sources that led to the hypothesis. A collection of evidence forms a skeleton that helps scaffold the story. Currently, journalists and staff conduct multiple searches on terms related to the hypothesis and refer back to documents they maintain as content summaries for topics revisited often to collect evidence (*e.g.*, statistics on protests related to environmental issues). A major challenge is quickly rounding up possibly relevant content and then efficiently determining which ones merit further investigation.

We transform these findings and design guidance from related work into 6 **design recommendations (D1-D6)** for a tool that supports journalists exploring thematic text collections:

(R1) Support content overviews. Providing a high-level view of parts or the whole collection can contribute to a general awareness

[1] http://www.ouest-france.fr/

of the content, suggest paths for initial exploration, and potentially lead to serendipitous discoveries **(D1, D2, T1)**.

(R2) Acknowledge entities. Entities should be easily identifiable. This includes the *Whats*, *Whos* and *Wheres* of the 5Ws, but also higher level concept types (e.g., a search for *"dogs"* that returns mentions of *"poodles"* and *"huskies"*) **(D2, T2)**.

(R3) Showcase connections and temporal trends. This includes detecting and highlighting relationships between entities and how they evolve over time **(D2, T2, T3)**.

(R4) Flexible search. Content should be thoroughly searchable, not only by keywords but also by entities and concepts **(D3, T3)**.

(R5) Seamless connection to context. Understanding motives and settings can help contextualize relationships and events, information that is notoriously difficult to obtain without access to the full text. The system should therefore provide easy access to segments of text relevant to an entity or topic of interest **(D1, T2, T3)**.

(R6) Track content of interest. Curation actions (e.g., annotation, selection and highlighting of entities) can help externalize the construction of a story and organize evidence **(D4, T2, T3)**.

4 *Storifier*

Following the formative investigation and outlining of design goals, we created *Storifier*[2], a tool to support exploration and understanding of text collections. It was designed to expedite journalistic analysis of documents and to help users keep track of passages of interest.

Storifier leverages structured data automatically extracted from raw text documents. Its natural language processing (NLP) pipeline extracts sentences, paragraphs, word counts, entities (person, location, organization) via named entity recognition (NER), key phrases [6] and relations [30]. It also associates taxonomy tags to the text to support concept search via a WordNet-based word sense disambiguation (WSD) module [35]. If available, dates or other order attributes (e.g., chapter numbers) are also extracted.

The interface (Fig. 1) features 4 thoroughly integrated panels: a (A) *Search panel* where users can perform and manage search queries; an (B) *Overview panel* listing frequent terms and entities in the collection that can be searched; a (C) *Snippets panel* featuring short text segments that match search queries and filters, alongside a temporal axis and a collection of documents; and a (D) *Document panel*, providing full access to documents in the text collection.

Kickstarting the exploration. Users working on a brand new text collection have two starting points: search for content using the *Search panel*, or browse the *Overview panel* for prominent terms **(R1)**. There are 5 types of searchable elements: (a) keyword, (b) person, (c) location, (d) organization, and (e) taxonomy concept **(R2)**. The *Overview panel* lists corpus-relevant entities and keywords in a scrollable vertical bar chart sorted by frequency. Items can be selectively filtered by toggling the corresponding searchable type icon on top of the panel (Fig. 1(B)), for more convenient browsing. Users can also use the *Search panel*, either via regular keyword search or via named entity search by appending designated prefixes to the search term, e.g., ("`P:`" for person, "`L:`" location, "`O:`" organization, and "`T:`" for taxonomy concept) **(R2, R4)**. Entity search (person, location and organization) operate as contextual filters to a regular keyword search: for example, a search for "`anne`" will match to terms like "canned" and "banned", but a search for "`P:anne`" will match only to identified person terms, e.g., "Joanne" and "Leanne". Taxonomy search ("`T:`"), on the other hand, retrieve all matches and its hyponyms (children) for a chosen concept based on assigned Wordnet concepts: for example, "`T:district.n.01`" will match to terms like "town", "state", and "county". Entity and taxonomy search can also be triggered from existing content by selecting a text segment in the *Document panel* (Fig. 1(D)): a pop-up window lists all encompassed entities and taxonomy concepts **(R2, R4)**, which can then be selected to perform a new search.

[2]Demo available at `storifier.cs.toronto.edu`

Navigating the collection. Search results are shown in the *Snippets panel* **(R3)** and *Document panel* **(R5)**. The *Document panel* provides access to full documents in a text collection, displayed in a continuous scrollable view, whereas the *Snippets panel* shows a scrollable list of *snippets*, i.e., search matches plus surrounding context to quickly assess relevance **(R1, R5)** (Fig. 1(C)). Snippets are linked on their left to corresponding *document glyphs* placed in a timeline to support temporal awareness **(R3)** (or, in the absence of a date, any other orderable variable, e.g., chapter numbers). On their right, they are linked to corresponding segments of the *Document panel* for easier matching **(R3, R5)**; clicking on a *snippet* also redirects to the original document **(R5)**.

On the *Document panel*, a *banded scroll bar* provides a visual overview of the entire text collection **(R1)** via gray and white bands, each matching a document in the collection. Larger bands match longer documents, and colours are alternated to better denote where consecutive documents begin and end.

Keeping track of relevant information. Useful search queries can be saved as *filters* for later retrieval **(R3, R6)**. Filters serve as multifaceted content slices and can encompass several queries to represent complex concepts. For example, a `Places` filter could include a taxonomy search query ("`T:district.n.01`") as well as other keyword and named-entity queries found to correspond to cities in the collection, e.g., "`Alderwood`" and "`L:Washington`". Users can use the *"Add"* button in the *Search panel* to create a new filter for a search query. This adds a new entry to the *Filters list* (Fig. 1(A)) with an associated color (from ColorBrewer [21]) and a default title (to be updated as its meaning evolves). The *"Add to"* button, on the other hand, adds the search query to an existing filter, thus merging their search results. When a *filter* is turned on (Fig. 1(A)), search results are then displayed with the corresponding filter color **(R6)**. The *Overview panel* reorders its items based on active *filters*, re-ranking terms that co-occur frequently in proximity to active *snippets* **(R3)**. Multiple *filters* can be active at once, which helps reveal relationships between them **(R3)**.

Finally, useful text segments can be bookmarked for later revisiting **(R6)**. From the *Snippets view*, *snippets* can be *upvoted* (to mark important), *downvoted* (checked, but unimportant), and commented on with a *note*. *Snippets* are re-sorted to display upvoted snippets first and downvoted ones last, while untagged snippets appear in the middle; snippets are further sorted by match to active *search/filters* and by *time*. From the *Documents panel*, text selections can also be upvoted; *upvotes* and *downvotes* are shown on the *banded scrollbar* and can be navigated with a click (Fig. 1(D)). Nevertheless, only the Snippets view order gets affected by up and down voting.

5 Case Study

We revisited the data journalist a month later in two separate 40min sessions, one to validate findings and introduce *Storifier*, and one to get feedback on its use for composing a news article. Our goal was to assess *Storifier* on its ability to support journalistic inquiry.

5.1 Session 1: Preparation

In the first session, we obtained more information about the journalist's experience (20+ years in data journalism) and practices. We also discussed what stories he was currently working on and where *Storifier* could help. A project in course was to explore content generated in the French National Debate (*Grand Débat National*) [2]. Between January-March 2019 the French government put online a website to invite citizens to express their opinions on several major issues, both in the form of closed-form (multiple-choice) questions and open-ended responses, available under the government open data initiative [18]. The journalist was interested in a subset of responses on environmental policies and we collected answers from 3 open-ended questions selected by him. The total set of answers amounted to over 195K responses. To handle the large scale, we

split each collection into three separate sub-collections based on response length and aggregated them by hour into 5 daily digests (0h-8h, then 4h intervals).

The journalist explained his goal in using *Storifier* was threefold: understand the general trends in open-ended responses that allow citizens to elaborate on the topic; search for unexpected points of view; and collect supporting quotes and testimonials. Prior to this qualitative exploration, he had processed the closed-form (multiple choice) responses using OpenRefine [1]. The journalist then independently used *Storifier* for 2 days in preparation for a new story.

5.2 Session 2: Insights

The second session took place 3 days after the story was published. We discussed the process he followed and how this was different from his past work, findings from using the tool that contributed to the story he wrote, pros and cons, and possible additions to better support journalistic inquiry. Next, we detail the journalist's comments on tool use and the insights he reached using it, and we tie his feedback to our design recommendations (Sect. 3).

Initiating exploration. The journalist explained that one of the positive aspects of *Storifier* is that it helps find possible starting points. The *Overview panel* (Fig. 1(B)) provided awareness of the most important topics; this is where he found that a very common occurrence was that of *recycling* (more than 6,000 occurrences) and led him to investigate how people deal with waste management. He thus created a filter for "`recycle`" and added to it the term "`waste`" to see what contributors had to say about the two. He was able to find a list of other actions people take along with recycling that were not part of the multiple-choice questions, such as eating less meat, eating local and organic food, buying less clothes, and so on.

Even though the NLP of the tool was far from perfect, the journalist found it useful when looking for information on how people view air travel for the environment. The taxonomy suggested a group of terms (including plane, air travel, flight) and gave him a better understanding of how many people touched the topic (1,015 occurrences), more than each of the individual terms **(R2, R4)**.

He explained that finding what may be an interesting topic in a completely new corpus is still hard. But again, he emphasized the possible starting points **(R1)** and that the tool made it easy to turn them into search filters that could be combined **(R4)**. In the past, his alternative was to apply NLP using Python to identify emerging topics and then conduct a text search for them in the full text. He explained that having the entity/keyword identification integrated with search functionality was powerful in *Storifier*. He stated that he often focused on the *Overview panel* and snippets to (i) identify interesting relationships between more than one term — something that is hard to do with simple text search, and (ii) to quickly skim snippets and moving to the full text only when needed — contrary to his previous practice that returns search results on the full text. One of the benefits of *Storifier* is that the snippets view allows to quickly identify the context under which the different terms are mentioned and to decide if reading the full text is needed or not **(R5)**. He explained that without the tool he would have not been able to conduct this type of exploration in 2 days. His first article on the topic appeared on March 16, 2019 [4], with *Storifier* used for the qualitative analysis of the open-ended responses.

Serendipitous and unexpected discovery. While skimming through the snippets view, the journalist saw an emerging pattern that he then verified by reading the full text. Among people who mentioned they recycled, they often expressed the feeling that this is not very helpful, using different terms (e.g., "useless" or "futile" in the snippets). Following onto corresponding full text, he found several mentions of people stating they felt their efforts had little impact since larger organizations (e.g., companies) do not necessarily contribute to the cause. To the journalist, this is an insight that points to the public's desire for a broader and coordinated public policy on recycling and waste management. He mentioned that this would have been hard to do with other exploration tools he now uses, as he would have to read many articles in full before identifying the trend. He explained that snippets made it easy to quickly go through multiple responses to get an overview of other entities/terms mentioned together, i.e., identify relationships between entities **(R3)**. The use of snippets also helped collected quotes to use in the final story **(R6)** which he states is crucial for this type of qualitative reporting.

Hypothesis formation and evidence. The French National Debate initiative was the government's response to the then-emerging Gilets Jaunes movement. As such, the journalist was not surprised to see the name of the French president (Macron) appear as an entity a few times (16 occurrences). He hypothesized that the forum may have been used as a platform for members of the movement to express a political agenda. But by skimming associated snippets, he found that in most cases the name referred instead to the Macron law from 2015 that opened the public transport market to the private sector, providing more competitive prices for train and bus travel. He verified this by reading the full text next. The combination of overview and snippets aided in both forming a hypothesis and quickly finding evidence to disprove it **(R5)**.

5.3 Other Feedback & Opportunities for Design

The journalist saw true potential in *Storifier* as a tool to aid journalistic inquiry: he was enthusiastic about it, used the tool in practice to write and publish an article. But he also provided further comments raising several design opportunities to improve this process.

In particular, he raised two usage scenarios that would benefit text analysis on a wider variety of scenarios proposed in our formative phase. First, he stated that he often writes retrospective stories revisiting a topic seen in the past (e.g., articles on recurring events like the *Tour de France*, or investigating the health care system). The tool inspired him to think that being able to conduct the same analysis in a new or updated corpus and see differences in results would be beneficial to update his stories. He suggested being able to export a set of filters and re-apply them to a different corpus.

Second, he felt it would be valuable to maintain a history of his investigative process to better support his claims. A way to export and show his exploration path (e.g., all his search filters) could provide audiences with an understanding of his process and allow them to potentially try the exploration and see if they reach the same conclusions. He also suggested being able to export all *items* in the *Overview panel* to quantitatively show the frequent entities in the text, to serve as evidence supporting certain findings.

6 Conclusion

In this work, we presented the design, development and evaluation of *Storifier*, a tool that supports a fluid reading-centered exploration of text collections by thoroughly connecting overview-level information to the underlying text. We worked closely with journalists throughout this process, and the tool supported the uncovering of newsworthy findings on a large collection of citizen feedback submissions and led to a published story. Our case study also revealed areas for improvement, such as better starting points for exploration, and the reuse of search filters across multiple collections. In addition, we see potential in repurposing user bookmarks and notes as additional resources to find related topics. Finally, it remains future work to investigate how we can better support summary views that represent the outcomes of the investigation, going from journalistic analysis and story investigation to story expression and summarization. Another space for improvement is to use active learning and intent modeling strategies to leverage user curation (e.g., upvotes and downvotes) to improve information retrieval [31, 37].

References

[1] Open refine. `http://openrefine.org/`.

[2] Le grand débat national. http://granddebat.fr/, 2019. Online; accessed June 2021.

[3] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. Serendip: Topic model-driven visual exploration of text corpora. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 173–182. IEEE, 2014.

[4] E. Alix. Climat. Ce que les franais font déjà, ce quils se déclarent prłts à faire. https://www.ouest-france.fr/politique/grand-debat-national/climat-ce-que-les-francais-font-deja-ce-qu-ils-se-declarent-prets-faire-6264587, March 2019. Online; accessed June 2021.

[5] S. Attfield, A. Blandford, and B. Craft. Task embedded visualisation: the design for an interactive ir results display for journalists. In *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*, pp. 650–655. IEEE, 2004.

[6] F. Boudin. pke: an open source python-based keyphrase extraction toolkit. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pp. 69–73. Osaka, Japan, December 2016.

[7] J. W. Boumans and D. Trilling. Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital journalism*, 4(1):8–23, 2016.

[8] M. Brehmer, S. Ingram, J. Stray, and T. Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE transactions on visualization and computer graphics*, 20(12):2271–2280, 2014.

[9] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu. Facetatlas: Multifaceted visualization for rich text corpora. *IEEE transactions on visualization and computer graphics*, 16(6):1172–1181, 2010.

[10] S. Chinoy. What happened to americas political center of gravity? https://www.nytimes.com/interactive/2019/06/26/opinion/sunday/republican-platform-far-right.html, 2019. Online; accessed June 2021.

[11] C. Collins, S. Carpendale, and G. Penn. Docuburst: Visualizing document content using language structure. In *Computer graphics forum*, vol. 28, pp. 1039–1046. Wiley Online Library, 2009.

[12] M. Correll, M. Witmore, and M. Gleicher. Exploring collections of tagged text for literary scholarship. In *Computer Graphics Forum*, vol. 30, pp. 731–740. Wiley Online Library, 2011.

[13] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pp. 115–122. IEEE, 2010.

[14] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pp. 93–102. IEEE, 2012.

[15] C. Felix, A. V. Pandey, E. Bertini, C. Ornstein, and S. Klein. Revex: Visual investigative journalism with a million healthcare reviews. In *Proceedings of Computation+ Journalism Symposium (CJ)*, 2015.

[16] J. A. Fitzpatrick, J. Reffell, and M. Aydelott. Breakingstory: visualizing change in online news. In *CHI'03 Extended Abstracts on Human Factors in Computing Systems*, pp. 900–901. ACM, 2003.

[17] I. Flaounas, O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini. Research methods in the age of digital journalism: Massive-scale automated analysis of news-contenttopics, style and gender. *Digital journalism*, 1(1):102–116, 2013.

[18] R. Franaise. Données ouvertes du grand débat national. https://www.data.gouv.fr/en/datasets/donnees-ouvertes-du-grand-debat-national/, 2019. Online; accessed June 2021.

[19] E. Günther and T. Quandt. Word counts and topic models: Automated text analysis methods for digital journalism research. *Digital Journalism*, 4(1):75–88, 2016.

[20] A. Handler and B. O'Connor. Rookie: A unique approach for exploring news archives. *arXiv preprint arXiv:1708.01944*, 2017.

[21] M. Harrower and C. A. Brewer. Colorbrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

[22] E. Isaacs, K. Damico, S. Ahern, E. Bart, and M. Singhal. Footprints: A visual search tool that supports discovery and coverage tracking. *IEEE transactions on visualization and computer graphics*, 20(12):1793–1802, 2014.

[23] S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann. On close and distant reading in digital humanities: A survey and future challenges. In *Eurographics Conference on Visualization (EuroVis)-STARs. The Eurographics Association*, vol. 2, p. 6, 2015.

[24] S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann. Visual text analysis in digital humanities. In *Computer Graphics Forum*, vol. 36, pp. 226–250. Wiley Online Library, 2017.

[25] S. Koch, M. John, M. Worner, A. Muller, and T. Ertl. Varifocalreaderin-depth visual analysis of large text documents. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1723–1732, 2014.

[26] P. Laban and M. Hearst. newslens: building and visualizing long-ranging news stories. In *Proceedings of the Events and Stories in the News Workshop*, pp. 1–9, 2017.

[27] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):25, 2012.

[28] B. OConnor. MiTextExplorer: Linked brushing and mutual information for exploratory text data analysis. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 1–13, 2014.

[29] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, vol. 5, pp. 2–4. McLean, VA, USA, 2005.

[30] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 492–501. Association for Computational Linguistics, 2010.

[31] T. Ruotsalo, J. Peltonen, M. Eugster, D. Głowacka, K. Konyushkova, K. Athukorala, I. Kosunen, A. Reijonen, P. Myllymäki, G. Jacucci, et al. Directing exploratory search with interactive intent modeling. In *Proceedings of the 22nd ACM international conference on information & knowledge management*, pp. 1759–1764, 2013.

[32] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*, pp. 364–371. Elsevier, 2003.

[33] J. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, 7(2):118–132, 2008.

[34] N. Sultanum, D. Singh, M. Brudno, and F. Chevalier. Doccurate: A curation-based approach for clinical text visualization. *IEEE transactions on visualization and computer graphics*, 25(1):142–151, 2019.

[35] L. Tan. Pywsd: Python implementations of word sense disambiguation (wsd) technologies [software]. https://github.com/alvations/pywsd, 2014.

[36] K. Van Syckle. The journalists who read all of president trumps tweets. twice. https://www.nytimes.com/2019/11/02/insider/trump-twitter-data.html, 2019. Online; accessed June 2021.

[37] J. Wenskovitch, L. Bradel, M. Dowling, L. House, and C. North. The effect of semantic interaction on foraging in text analysis. In *2018 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 13–24. IEEE, 2018.

[38] S. M. Yimam, H. Ulrich, T. von Landesberger, M. Rosenbach, M. Regneri, A. Panchenko, F. Lehmann, U. Fahrer, C. Biemann, and K. Ballweg. new/s/leak–information extraction and visualization for investigative data journalists. *Proceedings of ACL-2016 System Demonstrations*, pp. 163–168, 2016.