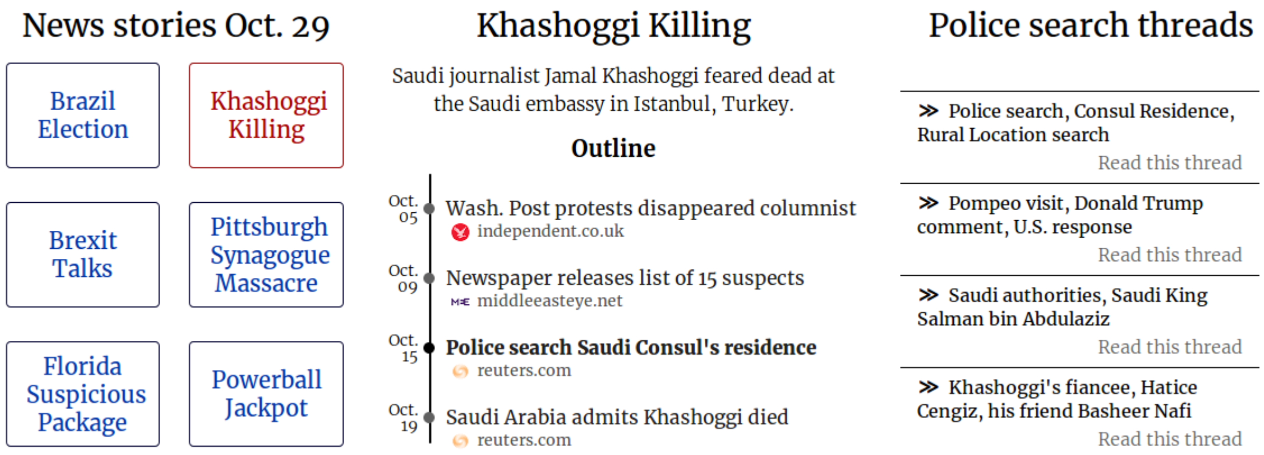
A framework for a text-centric user interface for navigating complex news stories

Philippe Laban
UC Berkeley
phillab@berkeley.edu

John Canny
UC Berkeley
canny@berkeley.edu

Marti Hearst
UC Berkeley
hearst@berkeley.edu



News stories Oct. 29

Brazil Election

Khashoggi Killing

Brexit Talks

Pittsburgh Synagogue Massacre

Florida Suspicious Package

Powerball Jackpot

Khashoggi Killing

Saudi journalist Jamal Khashoggi feared dead at the Saudi embassy in Istanbul, Turkey.

Outline

Oct. 05 Wash. Post protests disappeared columnist
independent.co.uk

Oct. 09 Newspaper releases list of 15 suspects
middleeasteye.net

Oct. 15 **Police search Saudi Consul's residence**
reuters.com

Oct. 19 Saudi Arabia admits Khashoggi died
reuters.com

Police search threads

≫ Police search, Consul Residence, Rural Location search
Read this thread

≫ Pompeo visit, Donald Trump comment, U.S. response
Read this thread

≫ Saudi authorities, Saudi King Salman bin Abdulaziz
Read this thread

≫ Khashoggi's fiancee, Hatice Cengiz, his friend Basheer Nafi
Read this thread

Figure 1: Sketch decomposition of the newsLens interface. News articles are composed into news stories (left). A story can be viewed at multiple levels-of-detail: a story name, an overall summary, an outline, and a detail view (center). The detail view is organized in parallel threads, which can be explored independently (right). This example is a simplified version manually curated from the newsLens interface for the purpose of making it understandable for this paper.

ABSTRACT

Many news articles are part of larger news stories that unfold over a period of time. Detecting these news stories, and presenting them to news readers is appealing, as it allows the reader to access the story's history freely, and see coverage of different sources on the same issues. It is however a technical challenge, as accumulating many news articles creates a large volume of potentially redundant content. We propose two principles, multiple levels-of-detail and threading, that form the basis of the newsLens interface, a text-centric intelligent interface to navigate complex news stories. We outline the principles with example use cases, and the paired technical challenges that arise when organizing large amounts of textual data.

1 INTRODUCTION

In September 2018, Reuters produced one news article every 7 minutes: news is produced continuously and relentlessly, as it is intended to relate in real-time the unfolding of our society. But news articles are not all isolated from each other, and it is common for some news articles to be updates, revisions of previously published articles. News articles can therefore form chains in time. We call these groups of news articles that relate to the same subject news stories. News stories can span several days (the release of the new Star Wars movie), several weeks (the murder of journalist Khashoggi), and even several years (MH370 plane disappearance).

The chain of news articles making a news story is enriched by the presence of many news sources covering the same stories, each adding perspective, expertise and providing a lens to see the story through.

This leads to complex news stories being composed of hundreds or thousands of news articles, even when only considering about 20 reputed news sources. For example, the news story about the Khashoggi murder contains 1200 news articles from 20 news sources, in 20 days of coverage. Reading all the article contents is not possible for a human, and yet each article can introduce new witnesses, experts, evidence, etc. **The objective of newsLens is to help humans navigate complex news stories, when there is too much content to be read exhaustively.**

Our challenge is to build an interface that helps news readers navigate the textual content of many news articles, by organizing and helping users what content to read. We

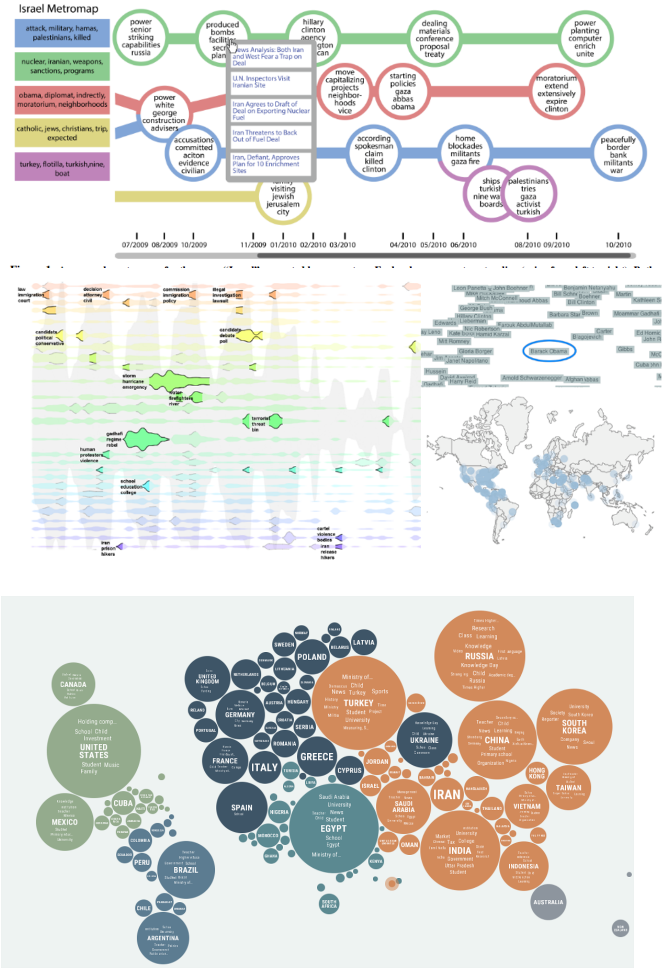

Israel Metromap
attack, military, hamas, palestinians, killed
nuclear, iranian, weapons, sanctions, programs
obama, diplomat, indirectly, moratorium, neighborhoods
catholic, jews, christians, trip, expected
turkey, flotilla, turkish,nine, boat
power senior striking capabilities russia
produced bombs facilities
hillary clinton agency
dealing materials conference proposal treaty
power planting computer enrich unite
move capitalizing projects neighbor-hoods vice
starting policies gaza abbas obama
moratorium extend extensively expire clinton
power white george construction advisers
accusations committed aciton evidence civilian
according spokesman claim killed clinton
home blockades militants gaza fire
peacefully border bank militants war
ships turkish nine wa boards
palestinians tries gaza activist turkish
visiting jewish jerusalem city
News Analysis: Both Iran and West Fear a Trap on Deal
U.N. Inspectors Visit Iranian Site
Iran Agrees to Draft of Deal on Exporting Nuclear Fuel
Iran Threatens to Back Out of Fuel Deal
Iran, Defiant, Approves Plan for 10 Enrichment Sites
07/2009 08/2009 10/2009 11/2009 01/2010 02/2010 03/2010 04/2010 05/2010 06/2010 08/2010 09/2010 10/2010
law immigration court
decision attorney civil
commission immigration policy
illegal investigation lawsuit
candidate political conservative
candidate debate poll
Barack Obama
CANADA
UNITED STATES
MEXICO
CUBA
BRAZIL
ARGENTINA
PERU
CHILE
UNITED KINGDOM
FRANCE
SPAIN
GERMANY
ITALY
GREECE
POLAND
SWEDEN
LATVIA
TURKEY
RUSSIA
UKRAINE
EGYPT
SAUDI ARABIA
IRAN
INDIA
CHINA
SOUTH KOREA
TAIWAN
HONG KONG
VIETNAM
INDONESIA
AUSTRALIA

Figure 2: User Interfaces to help users navigate complex news stories[4, 8, 11]. The interfaces are not text-centric, and a user must exit the interface to access news article content.

present two principles we believe are important to organize the volume of content:

- Each story is organized into multiple *levels-of-detail* that correspond to different news reading objectives,
- The content is organized into *parallel threads*, to allow readers to select which content to explore.

In particular, we present the algorithms, and principles we follow to keep our algorithms transparent. A live demo of our system is available at: https://newslens.berkeley.edu

2 RELATED WORK

Other research projects have attempted to build user interfaces to present complex news stories, made of underlying collections of news articles.

In the MetroMap project[11][1], groups of articles are filtered based on an initial text query and timeframe (Israel in 2009-2010). The news articles are organized into threads which can partially overlap. The resulting organization is presented as a "Metro map" visualization, whose x-axis represents time, and has the different threads (metro lines) unfolding on the y-axis. The text on a metro map is limited to headlines and keywords that identify each thread.

The Leadline interface[4] uses topic modelling to group news articles into stories, extracts and organizes the entities in each story. Visualizations are created for each story: a timeline based on volume on content over time, entity graphs, maps of places mentioned, and word clouds of keywords in the story. When interacting with the visual elements, the user obtains headlines of news articles, which can be followed to the article contents.

Unfiltered.news[8] is a project from the Jigsaw group at Google which presents world news on a map visualization. A user can choose a topic (such as Education), and see how the topic is discussed at certain times in different countries. Upon selection of a topic, the map is updated to show related entities and terms, and news articles that are the source of the visualization are made accessible to the user.

Projects by the Txt2Vis group, such as the Contextifier[7] or NewsViews[5] propose to use text from groups of news articles to produce or personalize visualizations. The visualizations can then be displayed alongside a news article.

In each of the projects presented, news articles are organized into story collections, but the interfaces don't allow the user to directly navigate the content of news articles: the interfaces are not text-centric, and the only way to access textual data (sentences) is to open external articles. The newsLens challenge is to build an interface which helps a news reader directly navigate the wealth of textual content in a news story, in an organized way.

3 NEWSLENS PRINCIPLES

We propose to build a text-centric interface to help users read complex news stories. The objective of the interface is to help the news reader navigate intelligently the content in the collection of news articles that make the story. This is challenging for several reasons:

(1) Not all the content can be shown to the user, and an algorithm is in charge of choosing a priority order in which to show paragraphs and sentences in. Some important facts might not be shown to the user if they are not ranked highly, which is a bias in the algorithm.
(2) Because several news sources are discussing the same facts, there can be redundant reporting, and the algorithm must take that into account to avoid showing

[1] A live demo is available at: http://metromaps.stanford.edu/

redundant content to the user several times, without dismissing relevant, similar updates.

(3) When extracting content from a news article, and presenting it independently, it is out of context, and can be misunderstood. For example, when pulling quotes by an actor in a story and presenting it separated from the article it is a part of.

(4) When presenting any content, it is important to keep a trace of the origin of content, as it was written by a journalist, for a news organization. The trace can be used to credit the content in the interface, and give access to original content. Keeping a trace of content origin is challenging with algorithms that combine and modify content, such as summarization algorithms.

Because of these challenges, many projects described in the Related Work Section do not directly give access to article content in their interface. They use the content to produce aggregate metadata (such as extracting keywords, entities, locations, events), and redirect the user to original news articles for the content.

We propose to tackle the challenge of building a text-centric news exploration interface by leveraging two principles: representing a story at multiple levels-of-detail, and building a story into multiple parallel threads. Our interface operates at the paragraph level: articles are split into paragraphs, and our challenge, is to organize, order and present the paragraphs to a news reader.

Multiple Levels-of-Detail

In order to attenuate the problem of dealing with a vast amount of text, we propose to represent the story as a hierarchical set of levels-of-detail. Each level of detail corresponds to an expected amount of content, and an expected read-time. We believe this to be particularly relevant in the context of news as readers might have different levels of interest in stories and can leverage an interface that explicitly allows them to choose a level-of-detail. We list the levels-of-detail we believe are relevant for news stories, with each we determine an expected read-time, outline an example use case of the level, and present related algorithmic challenges.

Level 0: Name of a story. We define the name of a story to be a noun phrase, which identifies the story at an abstract level. This name is intended to be an anchor the human uses to remember and access the story, and discuss with other news readers. In this paper, we've referred to stories using their names: Khashoggi Killing, Brazil Election, etc.

Expected read time: < 10 seconds

Example use case: Jane wakes up in the morning, and wants to spend a few minutes seeing what the main news stories are for the day. She goes to an interface where she is presented with a list of top news story names, and learns that the Chipotle E. Coli outbreak is an important story this morning. She decides she will tryutn to the story later.

Technical challenge: The field of labeling, and naming a story is challenging. It can be seen as similar to generating interpretable topic names in topic modeling. We describe the algorithm we use in Laban and Hearst [9].

Level 1: Overall story summary. Once a user knows the name of a story, the next step in the detail hierarchy is an overall story summary. The overall story summary is analogous to the first paragraph of a Wikipedia page, or lead section. According to the Wikipedia Manual of Style for the Lead section[12]:

> The lead should stand on its own as a concise overview of the article's topic. It should identify the topic, establish context, explain why the topic is notable, and summarize most important points, including any prominent controversies.

We adopt this definition for the overall story summary and choose to limit its length to 50 words, or about 4 sentences.

Expected read time: < 2 minutes

Example use case: Jane is now on the metro, commuting to work, and she wants to know more about the "Chipotle E. Coli Outbreak" updates. Particularly, she wants to know the number of affected and the regions affected by the outbreak.

Technical challenge: The field of summarization is well established in NLP. We choose to use extractive summarization[10], as it is compatible with our transparency objective (we display source links below the summary), such as in Figure 3.

Level 2: Story chronology. A story chronology is the a table of content of the news story. It is a chronological account of all major events and updates in the news story. We first use time-series analysis to determine the key moments in the story (bursts in amount of content) to partition the story into sections.

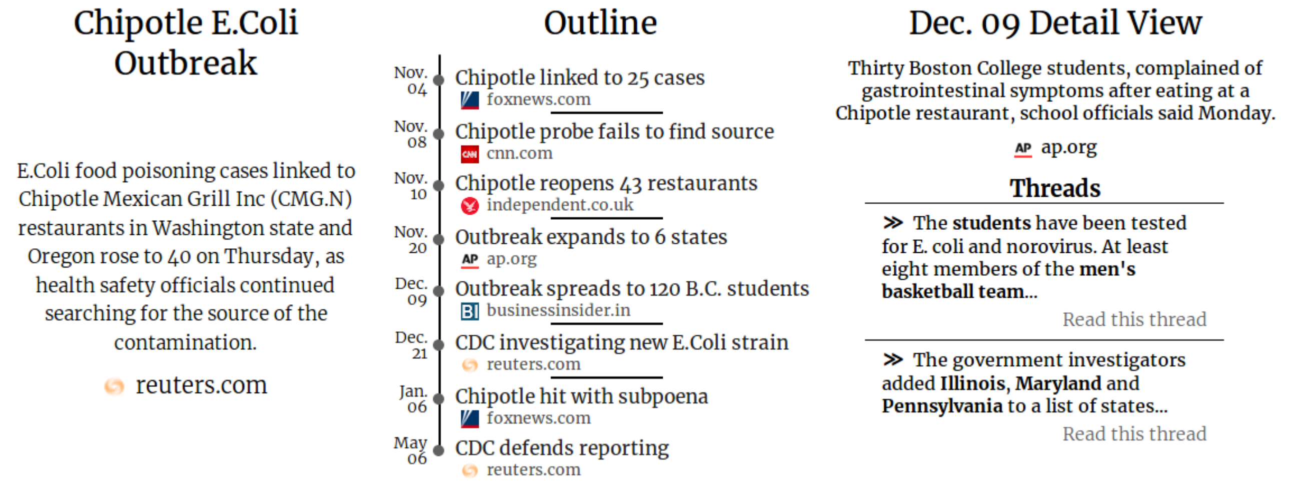
Within each section, articles are grouped into *highlights*, which represent one major event in the chronology of the story. Each highlight is represented in the chronology by a simple sentence. A simple sentence is a subject-verb-object sentence that is derived from a headline in the highlight. We propose to simplify sentences to avoid redundancy in the outline. For example, the sentence

> Chipotle reopens 43 restaurants after E. coli all-clear

is simplified to: Chipotle reopens 43 restaurants.

Expected read time: < 5 minutes

Example use case: Jane is reading about the "Chipotle E. Coli outbreak" news story, the outline, shown in Figure 3, gives quick access to the timeline of restaurants where the outbreak occurred, the CDC investigation, the closing and reopening of restaurants, etc.



Chipotle E.Coli Outbreak

E.Coli food poisoning cases linked to Chipotle Mexican Grill Inc (CMG.N) restaurants in Washington state and Oregon rose to 40 on Thursday, as health safety officials continued searching for the source of the contamination.

reuters.com

Outline

Nov. 04	Chipotle linked to 25 cases foxnews.com
Nov. 08	Chipotle probe fails to find source cnn.com
Nov. 10	Chipotle reopens 43 restaurants independent.co.uk
Nov. 20	Outbreak expands to 6 states ap.org
Dec. 09	Outbreak spreads to 120 B.C. students businessinsider.in
Dec. 21	CDC investigating new E.Coli strain reuters.com
Jan. 06	Chipotle hit with subpoena foxnews.com
May 06	CDC defends reporting reuters.com

Dec. 09 Detail View

Thirty Boston College students, complained of gastrointestinal symptoms after eating at a Chipotle restaurant, school officials said Monday.

ap.org

Threads

≫ The **students** have been tested for E. coli and norovirus. At least eight members of the **men's basketball team**...

Read this thread

≫ The government investigators added **Illinois**, **Maryland** and **Pennsylvania** to a list of states...

Read this thread

Figure 3: The Chipotle E. Coli Outbreak viewed through multiple levels-of-detail. The story name (level 0) and overall summary (level 1) are on the left, the story chronology (level 2) in the center, and the detailed event view (level 3) for Dec 09 is on the right. Terms in bold in the detailed view correspond to anchors of each thread.

Technical challenge: Partitioning a story into chronologically and semantically coherent sub-parts is a common challenge. We use a paraphrasing detector (common NLP task) to detect when headlines are about the same event to build highlights, and sentence simplification of headlines is a studied field as well.

Level 3: Detailed event view. The story chronology is only meant as an overview of major events in the story, it is a table of content of the story. Using this table of content, the user can access all the content around a given event, and explore it in a structured way. The detailed event view is organized into 2 sections: first an event summary is presented to the reader, which offers details specific to the event.

The second part of the detailed view is the thread section, which presents to the user several parallel paths of content the user can choose to read about, based on interest. Following the metaphor of a play, if the sections are the Acts of a play, the threads are scenes within the acts: each thread involves specific actors, locations and objects.

For example, when an industrial disaster occurs (such as the Chipotle outbreak), one thread might focus on Witness accounts, which would involve affected people, and their relatives. Another thread could be about lawsuits that followed, which would involve lawyers, plaintiffs, compensation amounts. Finally a final thread could be a political conversation about food safety standards, which would involve politicians, food safety scholars, and FDA representatives.

The threads are considered optional, in-depth content of the story. Based on interest, a user can choose which threads to spend time on and which to skip.

Expected read time: Variable, based on user interest.

Example use case: On the ouline, Jane notices that students at Boston College were affected around December 9th. Jane has friends in the Boston College, so she opens the detail event view around that time, and navigates the thread about student victims, to understand the chance that any of her friends were affected.

Technical challenge: Organizing the content into coherent threads is an open research problem, which can be seen as similar to the problem of topic modelling. The next section defines threads in a news context, outlines an algorithm to compute them, and proposes a thread navigation interface.

Threading of content

As explained in the Levels-of-Detail section, threads are part of the last level of detail of a story, where the user gets to actively choose what content to navigate based on interest. We first define the threads, then propose an algorithm to compute threads in a news setting, and showcase how the threads can be explored in our user interface.

Definition of a thread. In theater, the transition into a new scene is marked by the entrance or exit of a new character, a change of location, or both. Similarly in news, a thread is created upon the appearance of a new important actor or location in the news story. The thread is then composed of actions and statements made by the actors interacting with the new element.

This concept is similar to narrative anchors[3] in narrative theory, where a piece of content is defined by the *narrative* *space* it is projected on. In news, the narrative space is the entrance of a new important actor.

Threads are similar to topic modelling[1], but there are key differences. First of all, topic modelling makes a bag of word assumption (all words are judged equal), whereas a thread must be composed actors, locations, and objects. Secondly, a thread is characterized by some of its actors being anchors that must be new to the story.

In short, a thread is defined by a set of entering actors in a news story, and the content that involves them.

How to compute the threads. We assume that each news article can cover several threads, by addressing each one in a sequential manner. This assumption is similar to TextTiling [6]. For instance in the Chipotle E. Coli outbreak story, an article could be composed in the following way: first a 2 paragraphs talk about findings of E. Coli in a new state, followed by 3 paragraphs introducing and interviewing two witnesses from that state. Following this, the article can have 2 paragraphs about the ongoing CDC investigation, interviewing a CDC spokesperson. The article can conclude with public statements made by Chipotle in 2 concluding paragraphs.

Leveraging this assumption, we propose a simple algorithm to compute threads:

1. First determine the narrative anchors of the story. Although this is a challenging problem, our initial approach is to select a narrative actor corresponding to a person, organization, location or phrase that enters the story at a given point in time. In the Chipotle story, the CDC is a narrative anchor.
2. When a narrative anchor occurs in a paragraph, it is likely (although not always the case) that other terms in the paragraph are in a thread with the narrative anchor. For example, when the CDC is mentioned, the term "investigation", and the entity "Shiga toxin E. Coli O26", which the CDC was investigating. Build a co-occurrence count of terms to narrative anchors.
3. Cluster the narrative anchors with frequently co-occurring terms and actors, which forms the basis of the thread.
4. Assign paragraphs to threads when they match a narrative anchor and some of the basis terms in the thread. Some paragraphs might not be attributed to any thread, and some can be attributed to several.

Navigating a thread. A thread is made of one or several narrative anchors, involved actors and terms (that form the basis), which are represented by a set of paragraphs from news articles. To introduce the thread to the user, a single sentence is chosen which contains one of the narrative anchors of the thread. Upon interest, the user can open the thread and see the basis of the thread: terms, actors and locations mentioned in the thread. The user can then click on each element of the basis, which will append a paragraph to the thread involving that element. Below each paragraph, a link to the article can be used by a user wanting more information.

4 CONCLUSION

We have described the interaction framework for newsLens, our attempt at building a text-centric interface to navigate complex news stories. The two principles we rely on to organize large volumes of text content are organization into multiple levels-of-detail, and threading of content.

We are in the process of deploying our system on thousands of large news stories. In the future, we want to propose methods to evaluate effectiveness of our algorithms and principles. We believe quality of the multiple levels-of-detail should be tested in user studies, establishing the value of each level of detail, given a limited amount of time, whereas thread quality can be studied computationally, following work on topic modelling interpretability.[2]

REFERENCES

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[2] Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *NIPS*.

[3] Barbara Dancygier. 2007. Narrative Anchors and the Processes of Story Construction: The Case of Margaret Atwood's The Blind Assassin. *Style* 41, 2 (2007), 133–151.

[4] Wenwen Dou, Xiaoyu Wang, Drew Skau, William Ribarsky, and Michelle X Zhou. 2012. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on.* IEEE, 93–102.

[5] Tong Gao, Jessica R Hullman, Eytan Adar, Brent Hecht, and Nicholas Diakopoulos. 2014. NewsViews: an automated pipeline for creating custom geovisualizations for news. In *Proceedings of the SIGCHI conference on human factors in computing systems.* ACM, 3005–3014.

[6] Marti A Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics* 23, 1 (1997), 33–64.

[7] Jessica Hullman, Nicholas Diakopoulos, and Eytan Adar. 2013. Contextifier: automatic generation of annotated stock visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 2707–2716.

[8] Google News Jigsaw. [n. d.]. Unfiltered.news. http://unfiltered.news/. Accessed October 26, 2018.

[9] Philippe Laban and Marti Hearst. 2017. newsLens: building and visualizing long-ranging news stories. In *Proceedings of the Events and Stories in the News Workshop.* 1–9.

[10] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing.*

[11] Dafna Shahaf, Jaewon Yang, Caroline Suen, Jeff Jacobs, Heidi Wang, and Jure Leskovec. 2013. Information cartography: creating zoomable, large-scale maps of information. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 1097–1105.

[12] Wikipedia. 2018. Wikipedia Manual of Style Lead section. https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section