Nar-A-Viz: A Methodology to Visually Extract the Narrative Structure of Text

Dante Gama Dessavre[a], Jose E. Ramirez-Marquez[a,*]

[a]*School of Systems and Enterprises, Stevens Institute of Technology, Castle Point on Hudson, Hoboken, NJ 07030, USA*

Abstract

Being able to tell and understand stories is a key component for efficient communications. The narrative structure is the structural framework that underlies the order and manner in which this story is presented to a reader, listener, or viewer, this work presents a novel visualization methodology that can be used to extract the narrative structure of texts and present in a visual, immediate and intuitive manner. This tool efficiently sumarizes the content of a text, and it also allows in depth analysis of the narrative structure. The characteristics and use of the tool are exemplified using a TED talk by Hans Rosling. Its potential then is shown using two historical speeches: President Donald Trump's inaugural speech and I Have a Dream by Dr. Martin Luther King Jr.

Keywords: Visual text analytics, exploratory analytics, document summarization, topic analysis, interactive visualization

1. Introduction

Storytelling is an integral aspect of the human perception of reality. Narratives are a version of reality that is easier to accept, and as such they serve as instruments for our minds to construct reality[1].

Stories consist of a sequence of events, facts and observations linked together by unifying themes or arguments. Since ancient times, people have tried to understand

*Corresponding author

Email addresses: dgamades@stevens.edu (Dante Gama Dessavre), jmarquez@stevens.edu (Jose E. Ramirez-Marquez)

and formalize the elements of storytelling[2]. This elements compose the narrative structure of a text: the order and manner in which this story is presented to a reader, listener, or viewer. This definition of narrative structure can be extended to be applied to any type of textual message: *the structural framework that underlies the order and manner in which a message is presented to a reader, listener, or viewer*. This work presents a novel, automated methodology to visualize the narrative structure of textual and oral data.

The first contributions of this work is proposing an adaptable approach to modeling the narrative structure of a text:

1. Based on the chronological appearance of the words in the text. This reveals any temporal features of the narrative such as repetition.
2. Based on the meaning of words. This reveals the progression through groups of words of similar/different meaning, revealing the themes and progression of the text.
3. Based on the topic that each word refers to. This reveals what topics were mentioned in the text.

The other major contribution of this paper is the visualization of the narrative structure. T his visually can summarize the full contents of the text, showing the tremendous potential of the tool for both time savings as well as deep

We show the value of the narrative structure models and tool using three case studies, showing different aspects of how a message is delivered. To accommodate for intrinsic inaccuracies of the algorithms and visuals proposed, all of the modalities allow users to modify and adjust the results to improve the visual with expert/contextual knowledge.

Chapter 2 shows a survey of the existing literature on narratives and text visualization techniques. The methodology is defined in Chapter 3. It is exemplified using the TED conference: *Global Population Growth, Box by Box* by Hans Rosling in Chapter 4. Two case studies are presented to show the usefulness of the tool: President Donald J. Trump's inaugural speech and Dr. Martin Luter King Jr's *I Have a Dream* speech in Chapters 4 and 5. Finally, Chapter 7

2. Literature Review

2.1. Existing Work Regarding Narratives and Storytelling

Narratives have been defined as spoken or written accounts of connected events[1], with these accounts helping people to understand and make sense of the world. This definition of narrative implies a linear structure in the transference of messages, and has been called *event-telling*[3]. This concept can be applied to all forms of communicating a message: text, video, audio or combination of them.

Narratives can take many different forms: novels, short stories, fairy tales, newspaper articles and scientific journal articles.[4]. The structural characteristics that distinguish this genres is part of the narrative structure. Specific details distinguish one narrative from another, even when they are part of a same genre. The value of using narrative for content design and delivery has been intuitively understood for long time, and recent studies have shown scientific empirical results of its effectiveness[5, 6]. Several narrative features have been explored to better understand how they help culture and personal identity creation[7].

The value of narrative texts has been researched to understand what qualities are valuable for representation and communication. The first feature used for understanding is *temporality*, referring to the chronological appearance of the elements in the text[1] and how they can help a temporal understanding of the events[7]. The intrinsic differenced between genres also affect perception [1]. Genres influence our thoughts and perception. Among the genre defining features is referentiality[8], which refers to the use of external references from within one narrative. Other important features include: concreteness, vividness of description [9], and stylistic representations, such as linguistic features [10] or lexical forms [11]. It has been shown that there is a direct correlation between argument coherence and reading comprehension[12, 13]. Coherence has been decomposed into the following characteristics: referential, temporal, locational, causal, and structural. The objective of our work is to create a methodology that can expose, if not all, some of these important characteristics of the narrative structure.

2.2. Current Approaches to Text and Storytelling Visualization

There have been large amounts of work done in the areas of text visualization and using visualizations in the context of storytelling. A very important work, done by Segel and Heer[14], has caused a big increase in the amount of research being done in the area, and their paper made the idea of narrative visualization become popular. Their influence can be seen in multiple works. For example, GeoTime Stories and sense.us[15] were done to create visuals with annotations containing stories. Tableaus graphical histories[16] were developed so users can use their results for story telling style analysis.

Numerous approaches have been proposed for text visualizations. There are multiple tools to visually explore the content of text linked to metadata [17, 18, 19]. Some very recent works explore the idea of visualizing the evolution of citations in document collections, like cite2vec[20]. Some tools automatically annotate news articles[21, 22]. Multiple tools do extraction of the temporal structure, typically accomplished with regular expressions and pattern recognition. Examples include TempEx Tagger [23] , SUTime[24] , Heidel- Time [25] , and TERNIP [26]. There have been environments created to allow summarization of structured and unstructured datasets, allowing query-style operations[27, 28]. Non traditional sources of text, such as online forums and micro-blogging platforms, have also been explored[29, 30]. Overall, there has been a significant amount of work for visualizing patterns in collections of documents.

To understand the contextual meaning of the content of texts, topic modeling has been a very active area of research for the past 20 years[**?**]. Latent semantic indexing (LSI)[31] is one of the earliest topic modeling methods. Probabilistic topic modeling methods have been developed to overcome LSI's difficulty of interpretation,. Topics and documents are modeled as probability distributions over keywords. Probabilistic latent semantic indexing [31] and latent Dirichlet allocation (LDA)[32] are two of the most popular popular methods in this category. Two disadvantages that this methods have are: the relative high requirement of documents needed to create a high quality consistent topic model and the (relative) high amounts of processing time needed.There are others that have focused on visualizing the evolution of topics in time and/or space among multiple documents [33, 34, 35, 36, 37]. TIARA[38] in particular shows the

topical evolution of streaming document data, using an area graph style of visualization inspired by[39]. Other works have also used this type of visualization[40, 41]. FacetAtlas[42] utilizes graph layout-based visualization, similar to cite2vec, to show relationships between topic clusters. TopicPanorama[43] that relates topics from multiple different document corpora. TopicNets[44] and iVisClustering[45] continuosly create new topic models on subset of data to allow user navigation and interactivity.

The main inspirations for the visual component of our proposed methodology are timeline visualizations and lexical episode plots[46]. TimeSlice[47] uses faceted timelines containing many events; these timelines are generated from structured event data. LifeLines[48, 39, 49] and its descendants are interactive timelines that allow to summarize and compare medical patient treatment events. Law enforcement tools such as Criminal Activities Network[50] were created to identify crime patterns. Social media is another field where timelines are commonly used for trend analysis and anomaly detection.[51]. Lexical episode plots was developed as automated text-mining and visual analytics approach for exploratory analysis of a single document. Its core concept is the use of lexical chaining to identify significant words that whose frequency of appearance is significantly larger than expected relative to the document's average.

Overall there is a lot of work that has been done for text and topic visualization, yet to our knowledge there is no visualization methodology that can do wither of two things: extract the manner that a message is delivered in a single text, and show the meaning and topics of the contents of a single text. The closes, lexical episode plots, was developed mainly as an annotation tool to accompany the text instead of creating a visual representation, or summary, of the document. Our methodology aims to fill this gap.

3. Methodology

The Nar-A-Viz methodology was developed to accommodate multiple manners to model the narrative structure of a text. It is a three step design, with each step composed of various independent components that perform individual algorithms. This design can be seen in figure 1.

The general description of the steps is:

1. **Data Transformation:** Consists of obtaining the textual data and extracting the significant words (usually nouns and verbs).
2. **Narrative Structure Extraction:** Consists of transforming the data into the formats needed for the user interface. Then, analyzing the distribution of the significant words to see which are mentioned significantly more to emphasize them as important words. Finally, performing algorithms that cluster the words, based either on when they are mentioned or on their meaning.
3. **User Interface:** Consists on visually presenting the results of the Narrative Structure Extraction.

Each of the components can be created to accommodate a particular corpus. The next subsections describe the exact implementation of the components used to create the visuals shown in this manuscript.

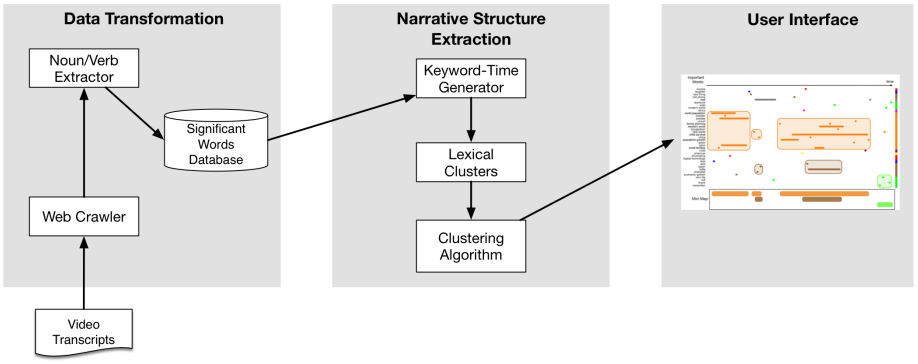

Figure 1: Nar-A-Viz System Description: A modular system has been implemented, with three steps: Data Transformation, Text Analysis and User Interface.

3.1. Step 1: Data Transformation

The text corpus analyzed consisted of transcripts of speeches. To obtain those transcripts, the following components were implemented:

1. A web *crawler* gets transcripts and articles from the web. It was implemented in Python using the *Scrapy*[**?**] package. Two sources were used to get documents:

- 3000 TED Talk transcripts were obtained from tedtalk.com.
- Google.com was used to obtain the transcripts of various important historical speech transcripts. They included speeches by historical political figures such as: *Donald J. Trump, Barack Obama, John F. Kennedy* , and *Dr. Martin Luther King Jr.*.

This created a collection of documents D of size n. Each document d contains the full text.

2. For each document d, the *keyword extractor* analyzes each sentence and extracts a list of the *significant keywords* $w_1, w_2...w_h$ (either nouns or verbs) $W_d = [w_1, w_2...w_h]$. This was done using Python and the TextBlob package[52].

3.2. Step 2: Narrative Structure Extraction

To extract the narrative structure based on the keywords, the following components were implemented:

1. For each document d, the *keyword-time generator* creates a dictionary T_d (symbol table)

$$T_d = \{keyword : times\}$$

For each mention of a keyword w in document d, T_d contains the times $t_w^1, t_{w2}^2, \ldots, t_w^m$ (number of sentence) where it is mentioned:

$$T_d = \{w : [t_w^1, t_w^2, \ldots, t_w^m]\}$$

.

2. For each entry $w : [t_w^1, t_w^2, \ldots, t_w^m]$ in T_d the *lexical clusters* component creates a set of lexical episodes

$$L_d = \{w : [t_w^i, t_w^{i+1}]\}$$

, where there is an entry $w : [t_w^i, t_w^{i+1}]$ if the distance $\delta = t_w^{i+1} - t_w^i$ is less than a parameter λ.

3. Three *clustering algorithms* for extracting the narrative structure were implemented for this work.

3.2.1. Chronological Narrative Clustering

The idea behind the *chronological narrative clustering* algorithm is that the temporal structure of when the important words are mentioned in a text reveals part of the narrative structure used to present the fundamental story. Particularly, repetition and return to previously used terms can reflect important strategies of communicating a message.

A clear example of this is poetry, where the temporal structure that creates rhyme is so important that it is the main characteristic that distinguishes it over any other form of writing.

The algorithm works as follows:

1. For each lexical episode l, create a narrative cluster

$$c = \{[t_i, t_j], [w_k, w_l \ldots]\}$$

   where t_i is the time where l begins and t_j is where it ends. $[w_k, w_l \ldots]$ are all the words that were mentioned between times t_i and t_j.
2. Create a cluster for consecutive words where there are no lexical episodes in clusters.
3. Join clusters that contain intersecting lexical episodes.
4. For each cluster with multiple lexical episodes remove the episode that are at the beginning and end of the cluster, and repeat steps 1 and 2.

3.2.2. Semantic Narrative Clustering

The idea behind the *semantic narrative clustering* algorithm is that a the meaning of words of a coherent text should follow a certain logical progression, and this algorithm tries to estimate what is that progression. The dictionary used can be defined by the users. In particular, we used a WordNet based implementation.

Before performing the clustering algorithm, the words are semantically sorted using the following algorithm [53]:

1. Given an unsorted list of words U, remove words that have the same meaning according to WordNet.

2. Given that unsorted list of cores U, create an empty list S that will store the sorted words.
3. Calculate the WordNet path distance between all the words.
4. Choose the two closest words and move them from U to S.
5. From the remaining words in U, choose the closest word W to either the first or the last word in S.
6. If W was closest to the first word in S, move it from U to S as the new first word in S, otherwise add it as the last word in S.
7. If there are remaining words in U go to step 4, otherwise end.

Afterwards, given a parameter λ_s, the algorithm works as follows:

1. For each lexical episode l, create a narrative cluster

$$c = \{[t_i, t_j], [w_k, w_l \ldots]\}$$

where t_i is the time where l begins and t_j is where it ends. $[w_k, w_l \ldots]$ are all the words that were mentioned between times t_i and t_j.

2. For each word w mentioned at time t, create a cluster c (or expand the cluster if the word is already in one) if in the next time $t + 1$ the word is less than λ_s semantic distance away.

The parameter $lambda_s$ allows users to control how lax the algorithm is in creating clusters. Experimentation has shown us that a lambda value of around 1/10 of the number of keywords gives a good first semantic narrative visual.

3.2.3. Topic Narrative Clustering

The idea behind the *topic narrative clustering* algorithm is to extract the chronological structure of the topics mentioned in a text. It is a manner to overcome the main limitation of the *semantic narrative clusters* as proposed in this work: a word can have a completely different meaning in different contexts.

The algorithm works as follows:

1. Create a topic model T_D based on the collection of documents D. In particular, we used two models: Latent Dirichlet Allocation[**?**] and manual annotation survey of 10 users (for the Donald Trump and Martin Luther King Jr speeches).
2. For each lexical episode l, create a narrative cluster

$$c = \{[t_i, t_j], [w_k, w_l \ldots]\}$$

where t_i is the time where l begins and t_j is where it ends. $[w_k, w_l \ldots]$ are all the words that were mentioned between times t_i and t_j.
3. Join consecutive words in the same topic where there are no lexical episodes in clusters.
4. Join intersecting clusters in the same topic.

3.3. Step 3: User Interface

The user interface consists of two components: *keyword sorting* and the *interactive visualization*.

3.3.1. Keyword Sorting

Based on the data of time and topic of words in each conversation, a new timeline/heat map hybrid visualization was developed. First, the keywords are sorted depending on what narrative cluster algorithm was performed before:

1. For the chronological narrative clusters, the keywords are ordered based on the first time they were mentioned, starting from the top.
2. For the semantic narrative clusters, the following sorting algorithm.
3. For the topic narrative clusters, the words are ordered so that they are contiguous to other words in the same topic.

3.3.2. Interactive Visualization

Using *d3.js*, this step takes the sets of results T_d, L_d and C and generates the narrative visualization: A table of size

$$number\_of\_unique\_key\_words \times number\_of\_total\_words$$

is defined. The color of cell (i, j) is defined based on the following rules:

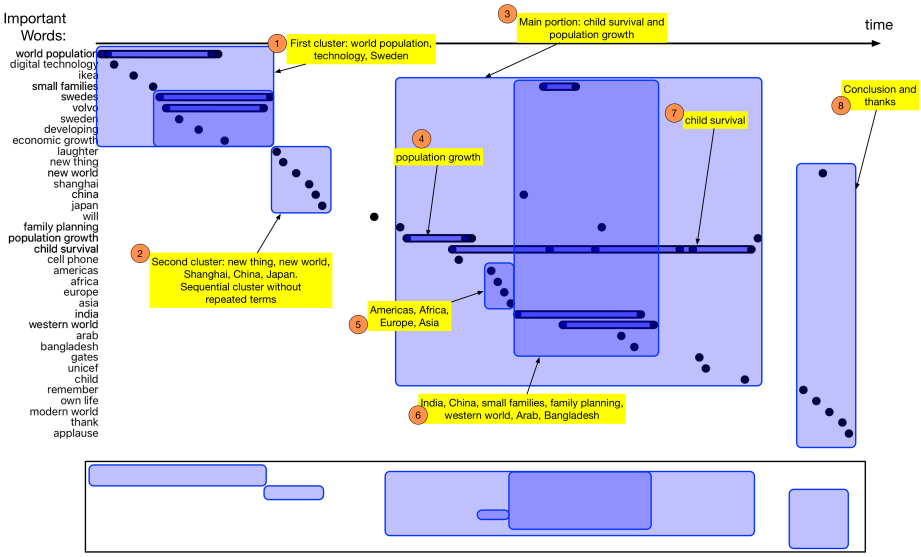

Figure 2: Chronological Narrative Visualization of the Speech *Population Growth, Box by Box* by Hans Rosling.

- cell (i, j) is colored if word i if was mentioned at time j.
- cell (i, j) is colored if it is part of a lexical episode $w : [t_w^i, t_w^{i+1}]$.
- cell (i, j) is colored with 50% transparency if it is part of a narrative cluster c.

, otherwise the cell is colorless.

A *mini map* representation then is build beneath the table, showing only the narrative clusters. The mini map can be used to zoom in and out in the main visual.

4. Nar-A-Viz Example: Han's Rosling - *Global Population Growth, Box by Box*

4.1. Chronological Narrative Visual

As was discussed in the previous section, the chronological visual is the first simple mode of Nar-A-Viz. The visual hows the structure of how an article progresses through different keywords, as can be seen in the Figure 2.

The chronological clusters show 4 main parts of the speach:

- A cluster, 1, that serves as introduction with three lexical episodes: *world population, Swedes* and *Volvo*. In general we can see how the author uses this terms of positive connotation to setup an introduction to the topic.

- Cluster 2 serves as a transitional portion before the main cluster. with no repetition of terms (essentially "advancing the plot"). The terms include: *new thing, new worlsd, Shangha, China, Japan.*

- The biggest cluster, 3, shows the main portion of the talk. The main lexical episones are *child survival* and *population growth*, showing what were the main concepts of the talk. Other lexical episodes included : *western world, small families* and *arab*. The cluster ends up in a higher part than the first word inside it, that means that the author mentioned a word that appeared before this cluster, showing an important component of the narrative structure: return to prior themes. In this case the term is *small families*, that was not mentioned since the introduction.

- Cluster 8 mainly shows the end of the talk, thanking the audience.

An important aspect of the chronological visual is the structure it reveals inside the main cluster (3):

- A first small cluster, 4, talking about *population growth*, serving as a segue to the largest portion of the talk, talking about *child survival*.

- A sequential cluster, 5, with no repeated words talking about different regions: *Americas, Africa, Europe, Asia*

- A big cluster, 6, with three lexical episodes: *India, western world* and *small families*.

4.2. Semantic Narrative Visual

For the semantic visual, Figure3 the keywords are ordered such that each word is adjacent to its closest words in meaning. This is a visual that shows the structure of how an article progresses through the meaning of keywords, exposing when an article

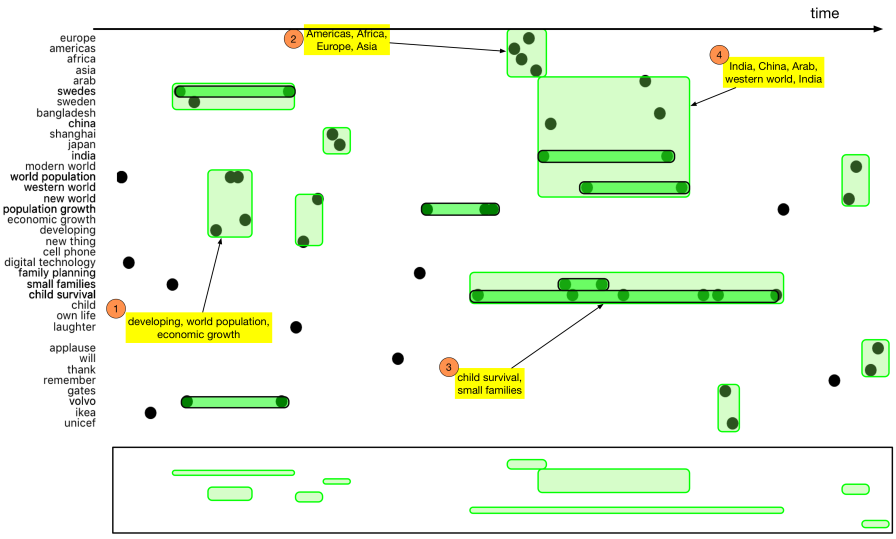

Figure 3: Semantic Narrative Visualization of the Speech *Population Growth, Box by Box* by Hans Rosling.

jumps to different groups of related words, exposing the idea of the narrative of the text and showing whether there is a continuity in it

A description of the semantic narrative structure is:

- A first small cluster, 1, talking about *developing world, population* and *economic growth*.
- A sequential cluster, 2, with no repeated words talking about different regions: *Americas, Africa, Europe, Asia*
- A big cluster, 3, with three lexical clusters: *western world, India, Bangladesh*.
- Shows the main lexical episodes in cluster 4: *child survival* and *small families*

4.3. Topics Narrative Visual

The final modality of the tool allows using it to understand the evolution of topics over time in the talk. Unlike the other two visuals, it requires a sizable corpus for creating the topic model, but it extracts different aspects of the narrative of text, based on the context of the text (politics, technology, etc.).

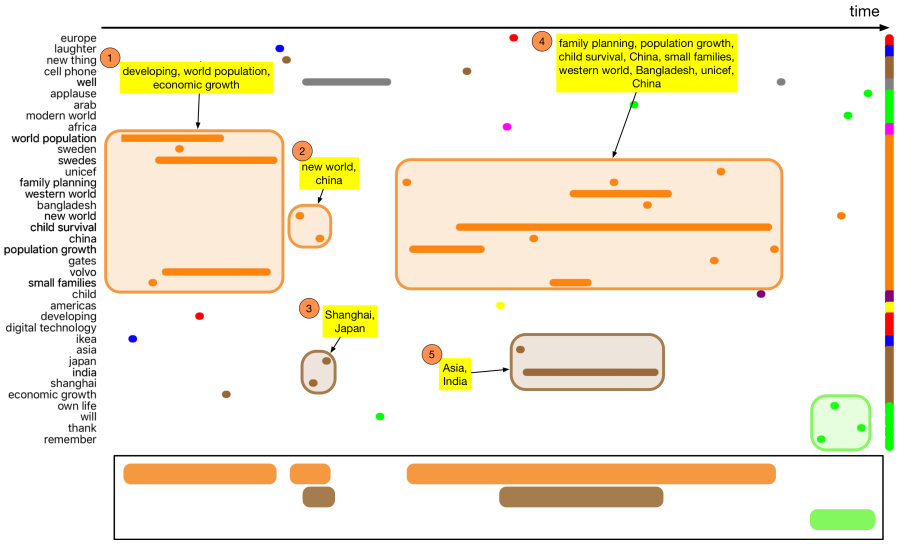

Figure 4: Topic Narrative Visualization of the Speech *Population Growth, Box by Box* by Hans Rosling.

Two main topics were present in the talk, represented bu orange and brown colors. The orange topic included population related terms mainly:*child survival, population growth, family planning*. The brown topic consisted mainly of asian geographical locations. It followed the following structure:

- During the intro the orange topic was completely dominant.
- then a connective part where both orange and brown were mentioned but not as predominantly.
- Then the main part with the orange topic being the main one, but the brown one being present as well.

5. Case Study 1: Donald J. Trump - *Inauguration Speech*

The first case study is President Donald Trumps inauguration speech. It is a very interesting speech that was written with a very clear objective: a straight forward speech to give a clear, simple, not politician like, message. In that sense all three visuals reflect that philosophy.

5.1. Chronological Narrative Visual

The main aspects of the chronological visual, Figure 5, are: A clear distinction between the introduction, main part and conclusion. Very few repeated keywords, except for one: *America*. The repetition of *America* and the constant come back to this term show its position as the central pivot of the speech. There is clarity and simplicity of the speech in its chronological structure., while using *America* as a framing device.

- The first cluster, 1, is the introduction with lexical episodes *obama, america, washington, nation's capital*.
- The main thing that characterizes all the other clusters is the lexical episode: *America*, which is constant and Trump returns constantly to.
- The conclusion cluster, 3, shows no new keywords and only consists of *god, together* and *America*.

This speech was written to reflect a straightforward, easy to understand and emotional talk by putting one term in the center at almost every single time: *America*. This manner of delivering a speech shows there was one objective: presenting a not political sounding, easy to relate, straightforward speech.

5.2. Semantic Narrative Visual

The semantic visual, Figure 6, shows the following characteristics: Very few parallel semantic clusters, coherent with the overall structure of the speech. Also a clear distinction between the beginning and the main part is shown, clusters 1 and 2 being the introduction. There is a clear progression through semantic cluster: 3 and 4 refer to infrastructure and education. 5 and 6 lead to 7, all talking about national and foreign factors that relate to industry and economy.

- Cluster 1 being the introduction thanking the people present.
- Cluster 2 shows him talking about Washington D.C. and Detroit.

- Clusters 3 and 4 show the first main keywords refered to infrastructure: *great schools, good jobs, safe neighborhoods, education system, beautiful students*. It also ended up in one of the main key terms: *american carnage stops*.

- Clusters 5 and 6 are parallel reflecting an ongoing theme for the main part of the speech: comparing foreign and national aspects. It ends in cluster 7 with: *new national pride*.

- Cluster 8 shows another major element in the narrative, Christianity against Islam. Terms include: *islamic, total allegiance, bible, god, almighty creator, special meaning*, showing terrorism as religion as major components of the speech.

5.3. Topics Narrative Visual

The topic narrative structure, Figure 6, is very similar to the semantic one, once again consistent with the other two visuals. The main aspect is that it reveals the amount of importance the topic of national against foreign had in the speech, seen in the big amount of the purple topic. Additionally, the yellow topic, referring to religious terms, is not very long in the amount of terms, but it has a significantly big cluster regarding the amount of time.

6. Case Study 2: Dr. Martin Luther King Jr - *I Have a Dream*

The second case study for Nar-A-Viz is Dr. Martin Luther King Jr.s speech: I Have a Dream. A widely studied speech whose complexity resembles more a poem than a simple speech. It is used to show how Nar-A-Viz can be used to show the multiple literary tools used by Dr. King and to contrast a very different text to President Trumps speech.

6.1. Chronological Narrative Visual

The chronological visual, Figure 8, shows the following characteristics:

- Constant referral to the main subject (note 1): *negro*.

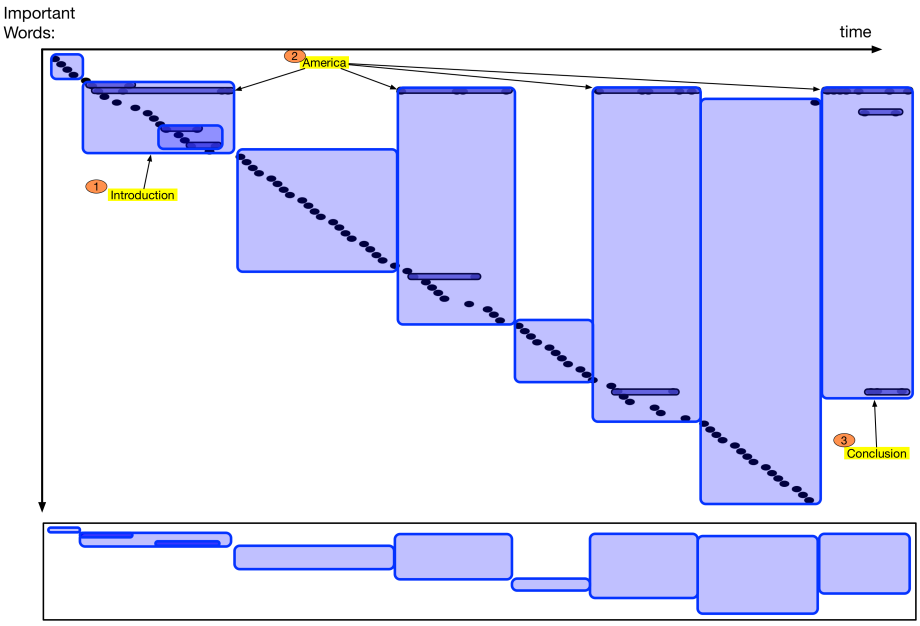

Figure 5: Chronological Narrative Visualization of the Inauguration Speech by Donald J. Trump. The Important Words are omitted due to space constraints, the most important elements are highlighted with notes.

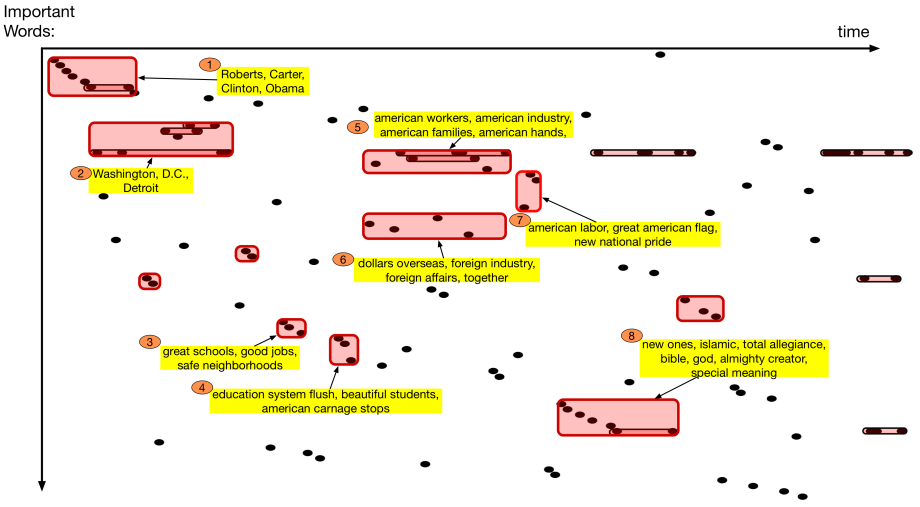

Figure 6: Semantic Narrative Visualization of the Inauguration Speech by Donald J. Trump. The Important Words are omitted due to space constraints, the most important elements are highlighted with notes.

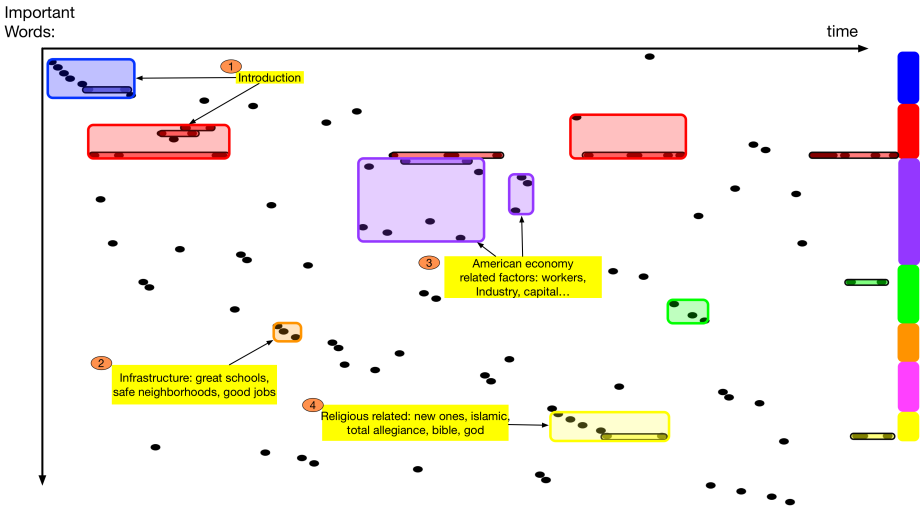

Figure 7: Topic Narrative Visualization of the Inauguration Speech by Donald J. Trump. The Important Words are omitted due to space constraints, the most important elements are highlighted with notes.

- The first big cluster has multiple different clusters of repeating structures (note 2).
- Eventual return in the second half to themes of the first half (note 3).
- The use of anaphora, repeating a sequence of words at the beginnings of neighboring sentences, shown towards the end in the big clusters: *freedom ring* and multiple geographical places (note 4).

Overall the speech has a very complex structure with clusters inside clusters showing multiple layers of composition.

6.2. Semantic Narrative Visual

The semantic visual, Figure 9 shows the following literary devices used by Dr. King:

- The use of contrasting terms next to each other: *black men-white men*, *racial justice-racial injustice*, among others shown in clusters with the note 1.

- Multiple figurative terms used in conjunction with geographical terms: *mighty stream, sunlit path, warm treshold*, among others shows in cluster 2 and 5.
- Use of metaphors such as: *emancipation proclamation-momentous decree, declaration-promissory note-bad check-insufficient funds*, among other shown in clusters 3 and 4.

The amount of literary devices used in this speech is very impressive, particularly given how easy it is to understand and relate to the speech.

6.3. Topics Narrative Visual

The topic narrative visual, Figure 10, shows multiple parallel topics being used at every moment. In particular it shows the following important characteristics:

- Clusters marked with 1 show the prominent references to the declaration of independence in the beginning.
- The purple clusters, marked with 2, show the prominent use of figurative terms all over the speech, but particularly in the beginning.
- Clusters marked by 3 show how the speech references the main subject, segregation of the *negro*, constantly.
- Orange clusters, marked with 4, show the moment when literal description of problems was mentioned.

7. Discussion and Conclusions

7.1. Contributions and Implications

A novel visual exploratory text analytic system called Nar-A-Viz was presented in this works. It can help users rapidly view, explore, and summarize the contents of a text. Additionally, it extracts the narrative structure in three ways: chronological, semantical and topical. This allows for a deep yet intuitive analysis of the manner that messages are constructed, which has a huge potential in its use for creating effective communications. Nar-A-Viz integrates this text analytics with interactive visualization technologies to support easy to use exploratory text analysis.

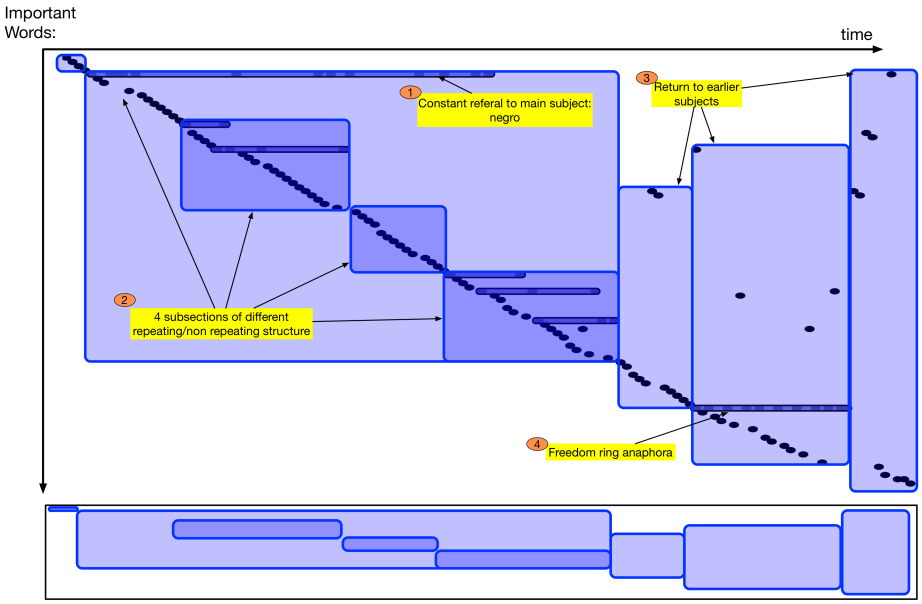

Figure 8: Chronological Narrative Visualization of the speech *I Have a Dream* by Dr. Martin Luther King Jr. The Important Words are omitted due to space constraints, the most important elements are highlighted with notes.

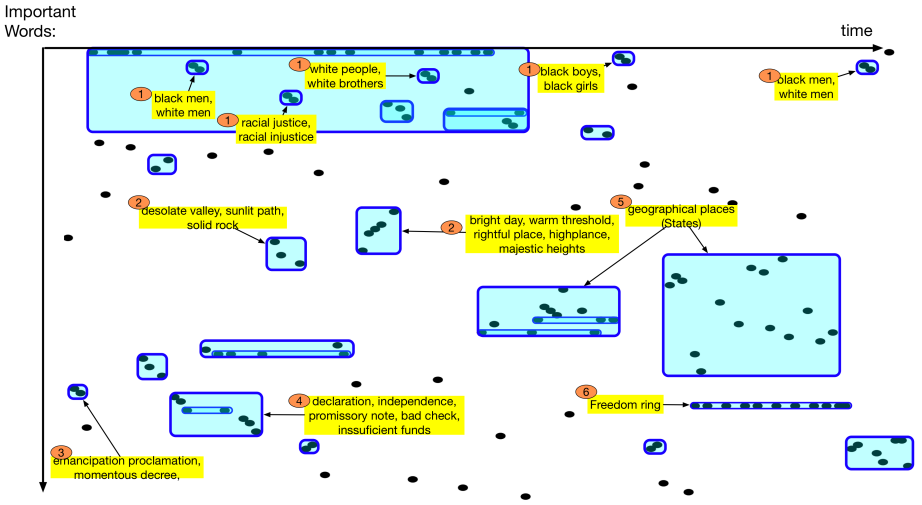

Figure 9: Semantic Narrative Visualization of the speech *I Have a Dream* by Dr. Martin Luther King Jr. The Important Words are omitted due to space constraints, the most important elements are highlighted with notes.

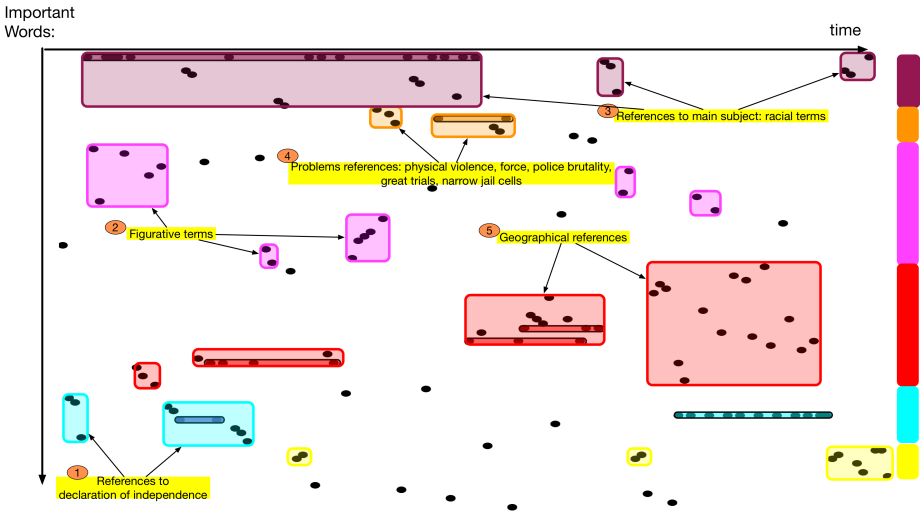

Figure 10: Topic Narrative Visualization of the speech *I Have a Dream* by Dr. Martin Luther King Jr. The Important Words are omitted due to space constraints, the most important elements are highlighted with notes.

7.2. *Future Developments*

Improvements to the tool that are currently being explored and implemented include:

- Reducing the list of words to only the most important words.
- Creating contextual dictionaries that can replace WordNet for better semantic clustering.
- Improving the selection of significant keywords. Currently we use off the shelf algorithms provided by TextBlob, yet some important feature were missed. For example, Dr. King famously starts his speech with the words *Five score years ago*, referencing Abraham Lincoln.

Since the framework of the current Nar-A-Viz system is easily extensible to any kind of communication, support other forms of differently structured communications, such as interviews and lectures (such as Chinese and Japanese) is being explored.

The research group is building an extensive text corpus database to have a baseline of topics to analyze new text pieces at the moment they are produced.

Finally, we plan to optimize the performance of the clustering components to reduce the processing time for the topic narrative structure. Currently the processing time of the tool is measured in hours for the topics visual, using parallel computing technologies this time could be reduced significantly, and it would allow analyzing the feasibility of a real time topic analysis system that can be used in the field for communication monitoring, which could be very useful in scenarios like crisis management and disaster response.

References

[1] J. Bruner, The narrative construction of reality, Critical inquiry 18 (1) (1991) 1–21.

[2] R. Parkinson, History of storytelling, Retrieved June 11 (2001) 2011.

[3] S. B. Chatman, Story and discourse: Narrative structure in fiction and film, Cornell University Press, 1980.

[4] M. Bal, Narratology: Introduction to the theory of narrative, University of Toronto Press, 2009.

[5] D. Herman, Narratologies: New perspectives on narrative analysis, Ohio State Univ Pr, 1999.

[6] D. Herman, Storytelling and the sciences of mind: Cognitive narratology, discursive psychology, and narratives in face-to-face interaction, Narrative 15 (3) (2007) 306–334.

[7] A. D. Brown, A narrative approach to collective identities, Journal of management Studies 43 (4) (2006) 731–753.

[8] L. Stanley, The knowing because experiencing subject: Narratives, lives, and autobioraphy, in: Women's Studies International Forum, Vol. 16, Elsevier, 1993, pp. 205–215.

[9] J. Steuer, Defining virtual reality: Dimensions determining telepresence, Journal of communication 42 (4) (1992) 73–93.

[10] J. J. Gumperz, J. Cook-Gumperz, Introduction: Language and the communication of social identity, Language and social identity (1982) 1–21.

[11] A. Duranti, C. Goodwin, Rethinking context: Language as an interactive phenomenon, no. 11, Cambridge University Press, 1992.

[12] D. S. McNamara, M. M. Louwerse, A. C. Graesser, Coh-metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension, Tech. rep., Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN (2002).

[13] D. S. McNamara, M. M. Louwerse, P. M. McCarthy, A. C. Graesser, Coh-metrix: Capturing linguistic features of cohesion, Discourse Processes 47 (4) (2010) 292–330.

[14] E. Segel, J. Heer, Narrative visualization: Telling stories with data, IEEE transactions on visualization and computer graphics 16 (6) (2010) 1139–1148.

[15] R. Eccles, T. Kapler, R. Harper, W. Wright, Stories in geotime, Information Visualization 7 (1) (2008) 3–17.

[16] J. Heer, J. Mackinlay, C. Stolte, M. Agrawala, Graphical histories for visualization: Supporting analysis, communication, and evaluation, IEEE transactions on visualization and computer graphics 14 (6).

[17] B. OConnor, Mitextexplorer: Linked brushing and mutual information for exploratory text data analysis, Sponsor: Idibon (2014) 1.

[18] E. Hoque, G. Carenini, S. Joty, Interactive exploration of asynchronous conversations: Applying a user-centered approach to design a visual text analytic system, Sponsor: Idibon (2014) 45.

[19] A. Muralidharan, M. A. Hearst, Supporting exploratory text analysis in literature study, Literary and linguistic computing (2012) fqs044.

[20] M. Berger, K. McDonough, L. M. Seversky, cite2vec: Citation-driven document exploration via word embeddings, IEEE Transactions on Visualization and Computer Graphics 23 (1) (2017) 691–700.

[21] M. Verhagen, R. Knippen, I. Mani, J. Pustejovsky, Annotation of temporal relations with tango, in: Proceedings of LREC, 2006.

[22] M. Verhagen, I. Mani, R. Sauri, R. Knippen, S. B. Jang, J. Littman, A. Rumshisky, J. Phillips, J. Pustejovsky, Automating temporal annotation with tarsqi, in: Proceedings of the ACL 2005 on Interactive poster and demonstration sessions, Association for Computational Linguistics, 2005, pp. 81–84.

[23] I. Mani, G. Wilson, Robust temporal processing of news, in: Proceedings of the 38th annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2000, pp. 69–76.

[24] A. X. Chang, C. D. Manning, Sutime: A library for recognizing and normalizing time expressions., in: LREC, Vol. 2012, 2012, pp. 3735–3740.

[25] J. Strötgen, M. Gertz, Multilingual and cross-domain temporal tagging, Language Resources and Evaluation 47 (2) (2013) 269–298.

[26] C. Northwood, Ternip: temporal expression recognition and normalisation in python, Ph.D. thesis, Masters thesis, University of Sheffield (2010).

[27] C. Felix, A. V. Pandey, E. Bertini, Texttile: An interactive visualization tool for seamless exploratory analysis of structured data and unstructured text, IEEE Transactions on Visualization and Computer Graphics 23 (1) (2017) 161–170.

[28] S. Fu, J. Zhao, W. Cui, H. Qu, Visual analysis of mooc forums with iforum, IEEE Transactions on Visualization and Computer Graphics 23 (1) (2017) 201–210.

[29] F. Beck, S. Koch, D. Weiskopf, Visual analysis and dissemination of scientific literature collections with survis, IEEE transactions on visualization and computer graphics 22 (1) (2016) 180–189.

[30] E. Graells-Garrido, M. Lalmas, R. Baeza-Yates, Data portraits and intermediary topics: Encouraging exploration of politically diverse profiles, in: Proceedings of the 21st International Conference on Intelligent User Interfaces, ACM, 2016, pp. 228–240.

[31] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 1999, pp. 50–57.

[32] D. M. Blei, Probabilistic topic models, Communications of the ACM 55 (4) (2012) 77–84.

[33] S. Koch, M. John, M. Wörner, A. Müller, T. Ertl, Varifocalreaderin-depth visual analysis of large text documents, IEEE transactions on visualization and computer graphics 20 (12) (2014) 1723–1732.

[34] W. Dou, X. Wang, D. Skau, W. Ribarsky, 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, IEEE, 2012, pp. 93–102.

[35] W. Dou, X. Wang, R. Chang, W. Ribarsky, Paralleltopics: A probabilistic approach to exploring document collections, in: Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on, IEEE, 2011, pp. 231–240.

[36] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, X. Lian, Interactive, topic-based visual text summarization and analysis, in: Proceedings of the 18th ACM conference on Information and knowledge management, ACM, 2009, pp. 543–552.

[37] M. Kim, K. Kang, D. Park, J. Choo, N. Elmqvist, Topiclens: Efficient multi-level visual topic exploration of large-scale document collections, IEEE Transactions on Visualization and Computer Graphics 23 (1) (2017) 151–160.

[38] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, Q. Zhang, Tiara: a visual exploratory text analytic system, in: Proceedings of the 16th

ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2010, pp. 153–162.

[39] S. Havre, E. Hetzler, P. Whitney, L. Nowell, Themeriver: Visualizing thematic changes in large document collections, IEEE transactions on visualization and computer graphics 8 (1) (2002) 9–20.

[40] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, X. Tong, Textflow: Towards better understanding of evolving topics in text, IEEE transactions on visualization and computer graphics 17 (12) (2011) 2412–2421.

[41] J. Leskovec, L. Backstrom, J. Kleinberg, Meme-tracking and the dynamics of the news cycle, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2009, pp. 497–506.

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

[43] S. Liu, X. Wang, J. Chen, J. Zhu, B. Guo, Topicpanorama: A full picture of relevant topics, in: Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on, IEEE, 2014, pp. 183–192.

[44] B. Gretarsson, J. Odonovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, P. Smyth, Topicnets: Visual analysis of large text corpora with topic modeling, ACM Transactions on Intelligent Systems and Technology (TIST) 3 (2) (2012) 23.

[45] H. Lee, J. Kihm, J. Choo, J. Stasko, H. Park, ivisclustering: An interactive visual document clustering via topic modeling, in: Computer Graphics Forum, Vol. 31, Wiley Online Library, 2012, pp. 1155–1164.

[46] E. Bertini, J. Kennedy, E. Puppo, Exploratory text analysis using lexical episode plots.

[47] J. Zhao, S. M. Drucker, D. Fisher, D. Brinkman, Timeslice: Interactive faceted browsing of timeline data, in: Proceedings of the International Working Conference on Advanced Visual Interfaces, ACM, 2012, pp. 433–436.

[48] C. Plaisant, B. Milash, A. Rose, S. Widoff, B. Shneiderman, Lifelines: visualizing personal histories, in: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, 1996, pp. 221–227.

[49] D. Archambault, D. Greene, P. Cunningham, N. Hurley, Themecrowds: Multiresolution summaries of twitter usage, in: Proceedings of the 3rd international workshop on Search and mining user-generated contents, ACM, 2011, pp. 77–84.

[50] H. Chen, H. Atabakhsh, C. Tseng, B. Marshall, S. Kaza, S. Eggers, H. Gowda, A. Shah, T. Petersen, C. Violette, Visualization in law enforcement, in: CHI'05 extended abstracts on Human factors in computing systems, ACM, 2005, pp. 1268–1271.

[51] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, T. Ertl, Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition, in: Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on, IEEE, 2012, pp. 143–152.

[52] S. Loria, Textblob: simplified text processing, Secondary TextBlob: Simplified Text Processing.

[53] C. Lipizzi, D. G. Dessavre, L. Iandoli, J. E. R. Marquez, Towards computational discourse analysis: A methodology for mining twitter backchanneling conversations, Computers in Human Behavior 64 (2016) 782–792.