A Simplified Topological Representation of Text for Local and Global Context*

Ishrat Rahman Sami†
Goldsmiths, University of London
London
isami001@gold.ac.uk

Dr. Katayoun Farrahi‡
Goldsmiths, University of London
London
k.farrahi@gold.ac.uk

ABSTRACT

Topological data analysis (TDA) is a branch of mathematics that analyzes the shape of high-dimensional data sets using geometry and algebra. TDA is used for data visualization which represents the relationship among elements using a network. Traditionally, TDA is quadratic in complexity and not commonly used for natural language processing. In this research, we visualize the relationship among words in a text block, words in a corpus and text blocks in a corpus. Text block represents a unit of a corpus such as, a web page in a web corpus, a chapter or section in a book corpus or a document in media corpus. This research proposes circular topology for representing words both for Local Context (LC) and Global Context (GC). Each text block is a set of sentences forming the LC. We found that feature words are extracted successfully from our LC analysis. The occurrence of extracted featured words in the corpus formed the GC. We evaluate this proposed simplified topological analysis on 3 different corpora: a single book corpus, a book corpus consisting of 7 books having 6020 narrations and a web corpus consisting of 990 web pages. The peripheral nature of the LC reduced the vocabulary size of the corpus significantly in O(nm) time where n is the number of text blocks and m is number of nouns in a sentence. GC analysis of featured words reflected useful properties of featured word movement which can be used to analyze topic evolution. GC analysis of text block points is aimed to find closely related text blocks in a radius. This reflected interesting results that need further supervised investigation. Research on topology driven natural language processing is in its infancy. This article contributes to this research field by introducing a method motivated by TDA to represent and visualize the peripheral nature of text block and corpus, by achieving success in dimensional reduction using local analysis and by simplifying the approach of complex topological analysis through localization.

CCS CONCEPTS

• **Mathematics of computing → Geometric topology**; • **Information systems → Content analysis and feature selection**; • **Computing methodologies → Natural language processing**;

KEYWORDS

Topology; Local Context; Global Context; Text Mining

1 INTRODUCTION

Data visualization of natural language can reveal new knowledge about a corpus. The high dimensional nature of language makes visualization a challenging task. Graph theoretical approaches are often useful for data visualization and analyzing similarity between documents [9]. Topology is a branch of mathematics that uses geometry and algebra to analyze the shape of complex data. Topological data analysis (TDA) is used for data visualization which represents the relationship among elements using a network. The topology of data does not depend on coordinate representation, rather distance between 2 points in the data reflects proximity and nearness [2] [8]. TDA methods are used to extract structural information before supervised and unsupervised analysis in the areas of image analysis, bioinformatics, finance analysis and astrophysics [8]. Due to the complexity of classical TDA, it is unpopular in relational analysis for big corpus.

This research attempts to introduce a novel method of representing relationships among words in a text block, words in a text corpus and text blocks in a corpus for knowledge discovery. This approach simplifies the complexity of TDA by assuming that a text block and text corpus have a peripheral structure where core words characterizing the text block or corpus remains close to the center and the explanatory or new feature words remains close to the periphery. We reflected the editorial writing pattern of a corpus in Local Context (LC) which is different from classical statistics based topic models and bag of word models. In an editorial writing pattern, the most important concepts are presented in introductory and concluding parts of a text block. Therefore, in our research, the order of the word appearance plays a major role in representation. We assume that, this is particularly true for chapters or sections of a book or web page article or news. In Global Context (GC), featured words identified for the text blocks in LC are converted to a point cloud. Text blocks are represented by the center of the polygon created by the corresponding featured word points for analysis. This article contributes by introducing

- Topological representation of text blocks for LC and GC

*Produces the permission block, and copyright information
†PhD Research Student, Department of Computing
‡Lecturer, Department of Computing

MM '17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/10.1145/3123266.3123330

- New dimension reduction technique by filtering peripheral words of LC
- Peripheral nature of GC which can be useful to monitor topic evolution

2 RELATED WORK

Data visualization enhances the understanding of underlying information more efficiently and quickly by using human visual capabilities [1]. Various plotting formats, for example scatter plots, pie charts, bar charts, heat maps, histograms etc. can be used to identify various aspects of data. The difference between these graphical representations and TDA is that TDA represents data using a connected network of nodes [1]. It reflects closer points are more similar than distant points. As each node can represent multiple dimensions of data, it can represent high dimensional data [1]. Topology is often used to represent output of various machine learning algorithms [1][9].

Research performed by Wagner et al. [9] reflects the use of TDA for analyzing structural similarities in a text corpus. In this research, 10-50 words have been extracted for each document. Each document is represented as a term vector in vector space. The term entries in the term vector reflects the tf-idf value of the term. Term vectors are used to project points in a point cloud. The researchers computed homology using discrete Morse theory and persistence homology of flag complex build from the point cloud [9]. They used cosine similarity to compute the similarity between documents. The experiment found lack of efficiency as the algorithm of persistent homology exhibits quadratic behavior [9].

Research performed by Luong et al. [6] used topology to analyze the syntactic organization of language and extract narrative structure related to author. They considered text as a topological space where topology of text is defined as incidence, distribution and density of individual linguistic features. Grammatical category and word occurrence had been used to identify linguistic features. Individual linguistic events are plotted as weighted points in a textual chain. Qualitative and quantitative analysis of the space exposed syntactic features associated with an author. They provided a comprehensive graphical representation of finite and non-finite verbal forms of the English and Latin languages.

Jo et al. [5] analyzed topic evolution on a collection of time stamped documents. They defined topics as a quantize unit of evolutionary change in the content and scanned documents chronologically. They captured rich topology of topic evolution in their inherent corpus. Their graphs revealed unknown information about the connected corpus.

Research by Oesterling et al. [7], analyzed high dimensional document point clouds generated by tf-idf weighting of words. They worked on classified documents and took a supervised approach for dimension reduction and point cloud generation. They identified uniform distance in higher dimension makes it difficult of identify similarity and dissimilarity among documents in a point cloud [7].

Research by Choudhury et al. [3] analyzed common linguistic networks of seven different languages which reflects interesting topological characteristics. They mentioned word occurrence to be governed by the syntactic property of a language. Their work reflected the peripheral structure of a language where the core

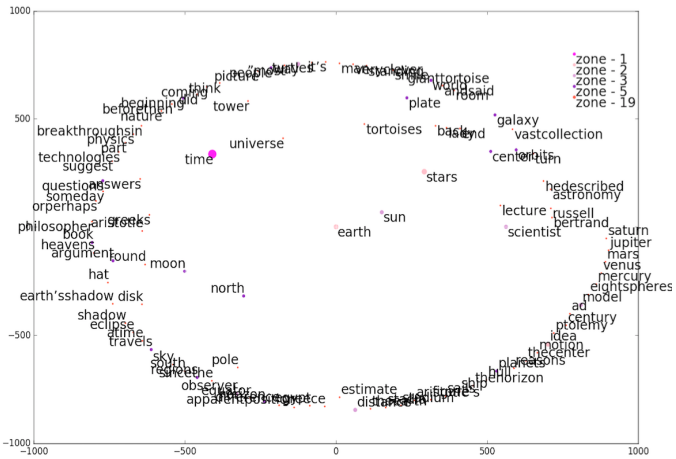

Figure 1: The LC of stem words in chapter 1 of the book "A Brief History of Time". The size of the points reflects the number of chapters the word appears. Colors of the points reflects the zones as presented in Table 1 used in GC. Core concept words of chapter 1 ('earth', 'sun', 'orbit', 'center', 'star', 'univers', 'begin', 'time', 'argument', 'moon', 'north', 'planet', 'idea', 'model', 'theori', 'forc', 'newton') are in the core and explanatory words are in the periphery of Chapter 1. This visualization appears to reflect the peripheral structure of LC.

hardly evolves with time and new words appear in the periphery of the network. A language grows slowly as new words creep in over time [3].

This article aims to address the complexity challenge of TDA for natural language processing by localization. It introduces the peripheral nature of text block, books corpus and web corpus. Data visualization of LC and GC interesting knowledge about the text block and the corpus.

3 TOPOLOGY FOR TEXT ANALYSIS

Topology was first introduced by Eular in 1974 [2]. Coordinate invariance, deformation invariance and compressed representation are the three main properties of topology [1][2]. In this research, we represented a particular text block in a corpus for LC and GC. A corpus is a set of related text blocks characterizing GC. A text block is a unit (document / web page/ chapter / section / subsection / article) of a corpus which is a sequence of paragraphs combined together under a title or file name depending on the type of corpus. A paragraph is a set of sentence contexts reflecting the LC of a text block. A sentence context is a set of nouns used in a particular sentence. A word is the unit of text analysis.

For simplicity, rather than representing each text block as a term vector and projecting it in a point cloud and analyzing the shape of the text data, we are assuming that a corpus has a circular topology in 2D space where words characterizing the corpus will remain close to the center and words characterizing a set of text blocks will remain close to the periphery of the circle depending on the set

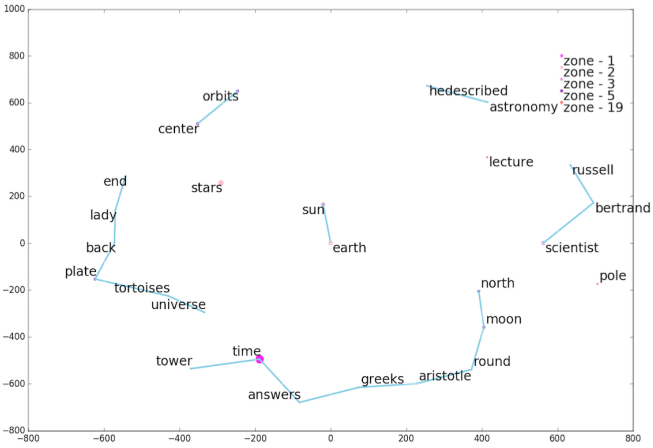

Figure 2: Filtered feature words after removing peripheral words from LC of chapter 1 in the book "A Brief History of Time". Connected edges are reflecting (scientist, bertrand, russell) simplicial complexes exposing context of the word. Edges represents that connected nodes are in pairwise distance ϵ.

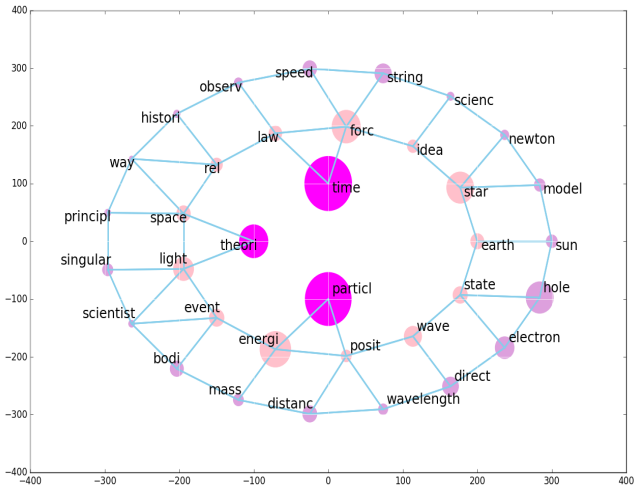

Figure 3: Feature words GC of the book "A Brief History of Time" where the corpus contains 12 chapters. Size of the nodes reflects average LC importance and color of the nodes reflects the zone as represented in Table 1. Edges represents that, connected nodes are in pairwise distance ϵ.

size. Similarly, we are also assume that a text block has a circular topology and words characterizing the text block will remain close to the center and words used to explain the core concept of the text block will remain closer to the periphery of the circle. Our assumption triggered an investigation of a linear weight of words in a text block such that it reflects the text block characteristics in a circular pattern. We stem all words for experiments.

3.1 LC analysis

The LC refers to a text block in a corpus. During the representation of LC of a text block as a point cloud, we aim to capture the editorial writing pattern. A word is an atomic unit of text processing. In the editorial writing pattern, most important words are reflected in the introduction and conclusion. So, the order of words in the text contributes to its importance. For extracting feature words, the text block is split into sentences. Parts of speech tagging is performed using the standard python natural language toolkit. Characteristics of the text block is reflected in its nouns. So, dimension is reduced by considering only nouns. A text block can be represented as,

$$T = \{w_1, w_2, \cdot\cdot, w_m\} \tag{1}$$

Where w_i represents the weight of noun i in a text block LC. m represents the number of nouns in a text block. Lets k represent number of sentence contexts in the text block. $i, m, k \in \mathbb{Z}$

$$w_i = \sum_{j=1}^{k} \Theta_p * p_{ij} + \sum_{j=1}^{k} \Theta_o * o_{ij} \tag{2}$$

Here j is the sentence index and $j \in \mathbb{Z}$. Θ_p is position contributing random constant 0.8. p_{ij} is the position contribution of the word w_i in sentence j. If a word appears in the first sentence, it gets highest p_{ij} set as number of sentences in a text block. p_{ij} is decreased by 1 as j increases. Θ_o is occurrence contributing constant 1. o_{ij} is the occurrence contribution of the word w_i in sentence j which is 1.

According to our assumption, max weighted feature words should be core to the text block. We project words for the LC in chronological order in a polar plot starting from 0 to 360 at a radius R_i where

$$R_i = \max_{1 \le i \le m} (T) - w_i \tag{3}$$

In TDA, persistence homology is a multi-scale algebraic geometric tool to analyze point clouds to identify clusters, holes and voids [10]. Persistence homology can be useful to identify the context of a point in a point cloud. Research on persistence homology for artificial intelligence is in its infancy due to the lack of applications [10]. It is beneficial to represent a family of simplicial complexes indexed by proximity parameter [4]. Cech Complex or Rips Complex theory can be used to build a family of simplicial complexes [4]. Given a collection of points in an Euclidean space, a Rips complex is the abstract simplicial complex whose k simplex corresponds to unordered (k+1) tuples of points which are with in pairwise distance ϵ [4] . We used the notion of Rips complex to connect nodes. If we build Rips complexes based on Euclidean pairwise distance on points on local context point cloud as in Figure 2, it reflects, word context. Persistence homology can be used to identify contexts and their representative main words based number of documents the word appears. For example in Figure 2 the complex [scientist, bertrand, russell] can be represented as scientist. Replacing the word contexts with representative word can further reduce dimension. Localizing topological analysis reduces the complexity of analyzing same thing in GC.

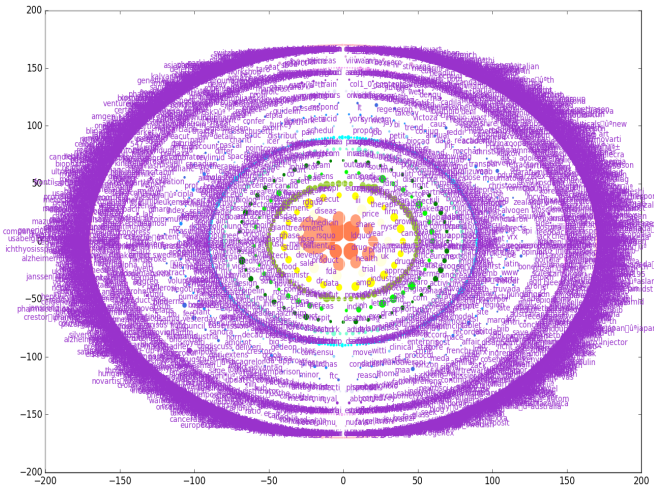

Figure 4: Featured words GC of a website. 990 web pages related to specific industry were analyzed. This reflects the density of featured words per zone. Only few words in the core describes the domain. The periphery has most amount words. New words appear in the periphery.

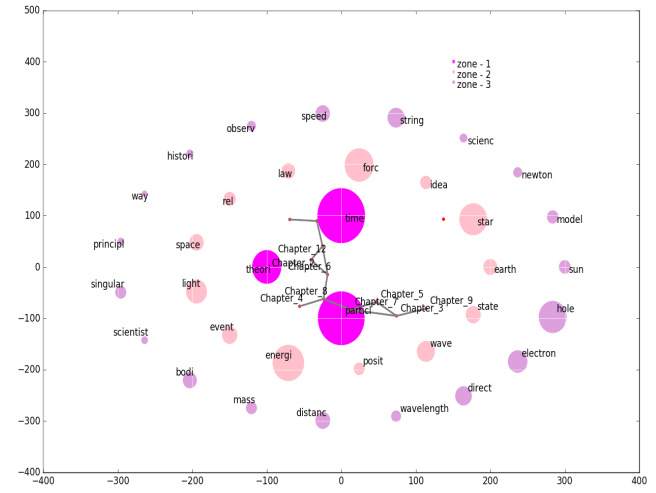

Figure 5: GC text blocks analysis of the book "A Brief History of Time". The text block point cloud is generated on top of GC word point cloud. Nodes prefixed by Chapter represents text blocks. A text block point represents the center of a polygon whose nodes are feature words of the block. This reflects, the relationships between the chapter similarity. Introductory chapter 1 is distant from others as it is more related to the general theme than specific concept. If a corpus is a book, GC text nodes doesn't reflect peripheral structure.

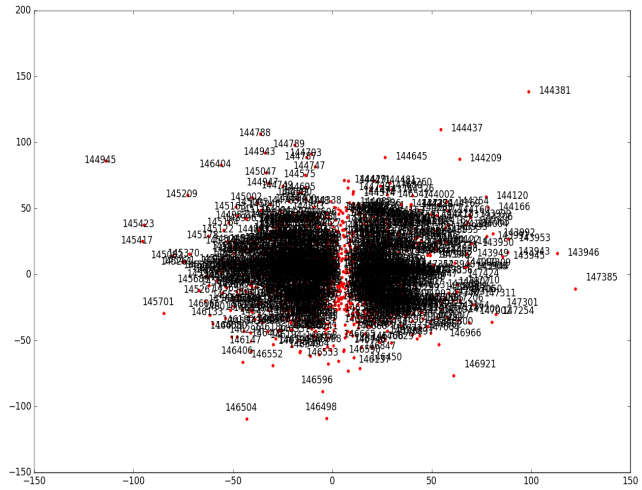

Figure 6: GC text blocks analysis of 990 web pages related to specific industry. The numbers represents the text block id. Discrete text block does reflect peripheral nature of corpus. It does reflects similar contents in neighborhood. It needs more supervised research.

3.2 GC analysis

GC refers to the relationships among featured words and text blocks in a corpus. For GC analysis, all the featured words identified for all text blocks in the corpus are plotted in a polar manner. As we reduce dimension significantly, we use number of blocks the featured word appeared for radius calculation. Assuming there are M corpus featured words. Radius R_i of a word w_i with b_i block appearance, is calculated as

$$R_i = \max_{1 \le i \le M} (b_i) - b_i \tag{4}$$

As the density of featured words near the core is less and the density of featured words is high near the periphery, to simplify observation, we divide the occurrences of word in text blocks in 19 zones as presented in table 1. As the density of words occurrences is high near the periphery, peripheral zones (11 to 19) are less distant than others. We calculate local average contribution for each feature word in the corpus. We tried local average contribution, tf-idf and number of blocks word appears as radius during generating global word point cloud. Using number of blocks as a radius gave us as peripheral structure where core words describe a domain and peripheral words or topics describes the blocks more. Topics or words of the corpus are plotted chronologically starting from 0. Incrementally adding text block reflects that new feature words appear in the periphery reflecting topic birth and as more related text node added topic starts moving to the direction of core. When the featured word move into the core radius it defines the domain as shown in figure 3 and figure 6. As the domain gets bigger, computing simplecies to identify word context becomes expensive. Therefore, localization of data becomes more important in GC.

Table 1: Zone distribution

Zone	Occurred more than % blocks
1	40%
2	30%
3	20%
4	15%
5	10%
6	5%
7	4%
8	3%
9	2%
10	1%
11	0.90%
12	0.80%
13	0.70%
14	0.60%
15	0.50%
16	0.40%
17	0.30%
18	0.20%
19	0.10%

For projecting text blocks in point cloud, the center of the polygon created by featured words of the text block was taken. For example, chapter 1 of the book "A Brief History of Time" is represented by the featured words ('earth', 'sun', 'orbit', 'center', 'star', 'universe', 'begin', 'time', 'argument', 'moon', 'north', 'planet', 'idea', 'model', 'theory', 'force', 'newton'). So, chapter 1 is projected at the center of this polygon calculated by average x and y co-ordinates of the featured words. The aim of this experiment was to localize related text blocks in a radius. For a small corpus it reflects related nodes in neighborhood. For example, chapter 4 and chapter 8 have 4 featured concepts (universe, time, energy, particle) common so they are in a neighborhood. But it is evident that for bigger corpus, our current chronological distribution doesn't project maintaining semantic similarity. So, in some cases related text blocks do not reside in neighborhood. So, a supervised projection of words is required for better results.

4 RESULTS

We evaluate this proposed simplified topological analysis on 3 different corpora: a single book corpus, a book corpus consisting of 7 books having 6020 narrations and a web corpus consisting of 990 web pages.

4.1 LC results

LC analysis reduces the dimension efficiently as shown in table 2. LC identified the featured words or topics accurately. Figure 1 visualizes LC of chapter 1 of the book "A Brief History of Time". In this example, the corpus is the book "A Brief History of Time" and each chapter represents a text block. Figure 1 reflects our claim that a text block can be represented meaningfully as a circular / peripheral pattern and the core word reflects the features of the chapter. If we further filter the periphery words, we get the core representatives or valid topics of the chapter as shown in Figure 2. This reduces the dimension of corpus vocabulary massively. This can be achieved in O(nm) times where n is the number of text blocks and m is number of nouns in a text block. We got similar results for text blocks in corpus 2 and corpus 3.

Table 2: Vocabulary size

Sample	Total blocks	Total words	Total nouns	Total topics
Corpus 1	12	14120	5760	106
Corpus 2	990	31105	16865	5176
Corpus 3	6020	33273	15814	6089

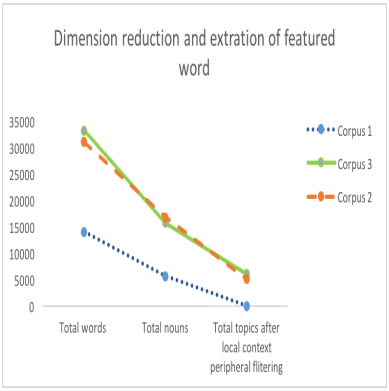

Figure 7: Vocabulary size of the corpus at various stages

4.2 GC results

GC reflection of corpus words for all corpora reflects peripheral nature of data. Topic word distribution in central zone reflects the core concept of domain. Topic word distribution in the periphery zones reflects the core classification of the blocks in the corpus. Many of the topics in the middle zones are an explanation or a connection between the topics in the core zones and peripheral zones. New topics are arriving in peripheral zone triggering topic birth. Migration of a topic through zones can be further analyzed to identify the needs of topic split and merge based on movement direction. When a topic enters core zone, it is not a good candidate for classification and can be considered as a dead topic based on the application of interest. It is evident that dimension of topic vocabulary can be further reduced by exploring featured word co-occurrence. Further research can be performed on exploring extraction of taxonomy using radial information of zones. Context exploring is computationally expensive. So, efficiency and performance will also be a big challenge for it. Results for GC analysis for text blocks shows that single book corpus doesn't reflect peripheral structure but discrete text block corpus 2 and corpus 3 do reflect peripheral nature. It will be interesting to investigate auto projection of correlated words and analysis the related text nodes closeness.

5 CONCLUSION

Topological analysis of natural language can be much efficiently performed in LC than in GC. Unsupervised LC analysis is efficient in feature extraction and dimension reduction. Dimension can be further reduced by persistent homology or using co-occurrence statistics. GC word point cloud visually reflects topics of the corpus and can be used for tracking topic evolution. GC text block point cloud of single book doesn't have peripheral nature but discrete text blocks corpus does reflect peripheral nature. Next stage of this research will be to investigate better point projection of words in

GC words context based on word co-occurrence statistics and verbs to get better text blocks relationship. This may create a platform for extracting taxonomy based on topological structure of featured words.

REFERENCES

[1] 2016. *White Paper: Topology and topological data analysis.* Technical Report. Ayasdi Inc.

[2] Gunnar Carlsson. 2014. Topological pattern recognition for point cloud data. *Acta Numerica* 23 (2014), 289.

[3] Monojit Choudhury, Diptesh Chatterjee, and Animesh Mukherjee. 2010. Global topology of word co-occurrence networks: Beyond the two-regime power-law. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 162–170.

[4] Robert Ghrist. 2008. Barcodes: the persistent topology of data. *Bull. Amer. Math. Soc.* 45, 1 (2008), 61–75.

[5] Yookyung Jo, John E Hopcroft, and Carl Lagoze. 2011. The web of topics: discovering the topology of topic evolution in a corpus. In *Proceedings of the 20th international conference on World wide web*. ACM, 257–266.

[6] Xuan Luong, Michel Juillard, Sylvie Mellet, and Dominique Longrée. 2007. Trees and after: The concept of text topology. Some applications to verb-form distributions in language corpora. *Literary and linguistic computing* 22, 2 (2007), 167–186.

[7] Patrick Oesterling, Gerik Scheuermann, Sven Teresniak, Gerhard Heyer, Steffen Koch, Thomas Ertl, and Gunther H Weber. 2010. Two-stage framework for a topology-based projection and visualization of classified document collections. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*. IEEE, 91–98.

[8] Václav Snášel, Jana Nowaková, Fatos Xhafa, and Leonard Barolli. 2017. Geometrical and topological approaches to Big Data. *Future Generation Computer Systems* 67 (2017), 286–296.

[9] Hubert Wagner, Paweł Dłotko, and Marian Mrozek. 2012. Computational topology in text mining. *Computational Topology in Image Context* (2012), 68–78.

[10] Xiaojin Zhu. 2013. Persistent Homology: An Introduction and a New Text Representation for Natural Language Processing.. In *IJCAI*.