InfraNodus: Generating Insight Using Text Network Analysis

Dmitry Paranyushkin
Nodus Labs, Paris, France, dmitry@noduslabs.com

ABSTRACT

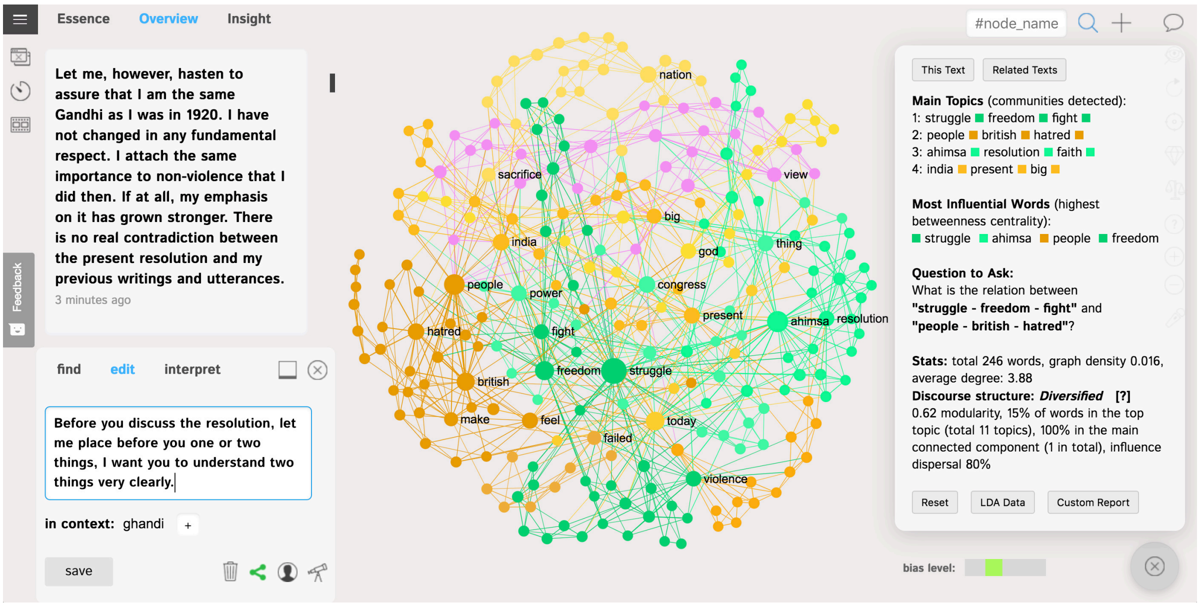
In this paper we present a web-based open source tool and a method for generating insight from any text or discourse using text network analysis. The tool (InfraNodus) can be used by researchers and writers to organize and to better understand their notes, to measure the level of bias in discourse, and to identify the parts of the discourse where there is a potential for insight and new ideas. The method is based on text network analysis algorithm, which represents any text as a network and identifies the most influential words in a discourse based on the terms' co-occurrence. Graph community detection algorithm is then applied in order to identify the different topical clusters, which represent the main topics in the text as well as the relations between them. The community structure is used in conjunction with other measures to identify the level of bias or cognitive diversity of the discourse. Finally, the structural gaps in the graph can indicate the parts of the discourse where the connections are lacking, therefore highlighting the areas where there's a potential for new ideas. The tool can be used as stand-alone software by end users as well as implemented via an API into other tools. Another interesting application is in the field of recommendation systems: structural gaps could indicate potentially interesting non-trivial connections to any connected datasets.

CCS CONCEPTS

• Human-centered computing~Visualization systems and tools • Applied computing~Document analysis • Information systems~Information extraction • Networks~Topology analysis and generation

KEYWORDS

text mining; mental maps; text network analysis; discourse bias; bias; network analysis; graph theory; TNA; information interfaces; network topology; cognitive science; topic modelling; ideation; comprehension; creativity; research; insight;



Let me, however, hasten to assure that I am the same Gandhi as I was in 1920. I have not changed in any fundamental respect. I attach the same importance to non-violence that I did then. If at all, my emphasis on it has grown stronger. There is no real contradiction between the present resolution and my previous writings and utterances.
3 minutes ago

Before you discuss the resolution, let me place before you one or two things, I want you to understand two things very clearly.

Main Topics (communities detected):
1: struggle 🟩 freedom 🟩 fight 🟩
2: people 🟧 british 🟧 hatred 🟧
3: ahimsa 🟩 resolution 🟩 faith 🟩
4: india 🟨 present 🟨 big 🟨

Most Influential Words (highest betweenness centrality):
🟩 struggle 🟩 ahimsa 🟧 people 🟩 freedom

Question to Ask:
What is the relation between **"struggle - freedom - fight"** and **"people - british - hatred"**?

Stats: total 246 words, graph density 0.016, average degree: 3.88
Discourse structure: ***Diversified*** **[?]**
0.62 modularity, 15% of words in the top topic (total 11 topics), 100% in the main connected component (1 in total), influence dispersal 80%

Figure 1: InfraNodus interface visualizing the text, the main topics inside, discourse structure, and bias index.

1 Introduction and Background

Mental models, mental maps and mind maps are often used to represent complex ideas [1], [2] and have a positive impact on teaching and learning [3], [4], [5]. Mental maps can be extracted from texts to produce cognitive maps, which can then be used for sentiment analysis [6], [7]. In the recent years this field has been enhanced with the network metaphor: the nodes are the concepts and the connections are the relations between them [8], [9]. The relations between the concepts can be plotted on the basis of their co-occurrences, which, in turn can facilitate the process of semantic priming — the process by which words tend to be recognized faster when used with the words that have close semantic proximity to them [10], [11], [12], [13]. Text network analysis methods based on this approach have been shown to be successful for better comprehension of texts and topic modeling [14], [15], [6].

In this paper we introduce an open-source tool for text network analysis InfraNodus[1], which can be used to enhance visual representation of ideas, their consequent analysis, and to generate insight. InfraNodus implements a method for text network analysis based on plotting a text as a network graph. The nodes on this graph represent the words and the edges are their co-occurrences [16], [17], [18]. Once a text is represented as a graph in this way, a wide range of tools from network analysis are then applied to detect communities of closely related concepts or topical clusters (topic modeling), identify the most influential nodes (top keywords), perform both quantitative and qualitative analysis of text, evaluate the discourse structure and bias, identify structural gaps, and perform comparative analysis of several texts [16], [19], [20]. Moreover, visual representation of a text as a network used in the process of ideation helps towards a more coherent expression of ideas [21], so the tool may find useful application in creative writing, research, in educational context, and in therapy.

State-of-the-art and its limitations. Similar tools have been presented in scientific literature: iVisClustering [22], TopicNets [23] and VisiRR [24] among others. However, those tools are not available to general public, their source code is closed, they are mainly focused on studying a large corpus of texts, and their application is limited to topic modelling and keyword-based text retrieval. InfraNodus is available both online and as a standalone open-source version. Its interface is specifically designed for real time text processing, providing text network visualization and analysis live as the new data is added. Finally, InfraNodus offers discourse structure analysis providing the measurement of the bias level in discourse and identifying the structural gaps in discourse, which is not available in other tools and methods, to our knowledge.

Relation to other text mining approaches. The approach implemented in InfraNodus can enhance the existing methods of finding structure within text that employ graphical models and topic modeling: latent semantic analysis or LSA [25], pLSA [26], Pachinco allocation [27], latent dirichlet allocation or LDA [28], relational topic models [29], word2vec algorithm [30], [31] and its extension lda2vec [32]. These methods are based on retrieving the topics from text by identifying the clusters of co-occurrent words within them, based on the bag-of-words and skip-gram models. This data can then be used to classify similar documents, improve text indexing and retrieval methods, and to identify evolution of certain topics overs a period of time within a specific text corpus [33].

The method used in InfraNodus is different from the probabilistic methods outlined above (LDA, pLSA, word2vec, lda2vec etc) even though its underlying assumption is similar: identifying the words that occur closer to each other in text can be used for topic modelling. InfraNodus uses graph theory instead of probability distribution to identify the related words and assign them into topical clusters. This application of graph theory helps gain a better understanding of the textual discourse structure and to provide advanced visualization tools, which can be useful both for quantitative and qualitative analysis.

The main difference is that methods like LDA or word2vec will take co-occurrences into account, but they do not provide insight into the structure of a narrative. These methods may ascribe words to the same topic (following bag-of-words model), but there will be no information provided about the relation of these topics to each other: how connected they are, how equally represented they are in a text, etc. The method implemented in InfraNodus provides this insight both qualitatively and quantitatively. Another issue with probabilistic methods is that they often require prior training on a document corpus, and they may be quite difficult to understand, to implement, and to interpret. The complex math behind LDA or word2vec may be difficult to access for the people who cannot afford to go into the technical details. Text network analysis approach used in InfraNodus, on the other side, does not require any training, it can work in real-time during the process of writing on a simple document, it is very intuitive and the math behind it is relatively simple. It is also much easier to understand intuitively as it provides direct visual representation.

At the same time, InfraNodus also offers LDA topic modelling and keyword extraction functionality, so that users can enhance the data obtained using LDA with additional information about the narrative structure and the topical cluster data obtained using network analysis (and vice versa). The method provided in InfraNodus does not compete with other topic modelling methods and does not attempt to outperform them. It provides the ability to get a better understanding of the text network topology, which can yield many other insights about the discourse structure that are not available in LDA or word2vec: how connected a discourse is, whether it's biased towards a certain set of concepts, how dominant a specific topic is towards the whole discourse, whether there are structural gaps that may indicate a potential for new ideas.

2 The Method and Its Implementation

InfraNodus is an open-source software written in JavaScript (Node.Js) using Sigma.Js, Cytoscape and Graphology libraries in the front-end and java-based Neo4J graph database for the backend[2]. It is platform-independent and can run both on desktop (hosted) and mobile (cloud) devices. Its source code is open and is available under the GPL open-source license on www.github.com/noduslabs/infranodus. Users can either install InfraNodus on their own machine or use the cloud-based version on www.infranodus.com. The resulting graphs and results can be embedded on other websites and exported into popular graph formats for further processing and analysis (e.g. Gephi). Some of the functionality is also available via an API, so InfraNodus can be used in conjunction with other text mining and analysis software.

In this section we will demonstrate step-by-step implementation of the text network analysis method as it is applied in InfraNodus. We will also demonstrate the qualitative and quantitative insights that can be obtained by the researchers using this method. For this demonstration we will use a well-known speech by Mahatma Gandhi "Quit India" [34].The interactive network visualization with all the analytics presented in this paper can be accessed via www.infranodus.com/texts/ghandi.

[1] Online: www.infranodus.com, source: github.com/noduslabs/infranodus, video tutorials: www.infranodus.com/#tutorials & https://youtu.be/bMmTRBZpjWw

[2] www.nodejs.org is a JavaScript run-time environment, www.sigmajs.org is a network visualization library created by Alexis Jacomy, graphology.github.io is a library for graph analysis created by Guillaume Plique, Cytoscape.org is a platform for complex network analysis, Neo4J.org is a graph database

2.1 Step 1: Text Normalization

First, all the words in the text are converted into their lemmas to reduce redundancy — keeping the morphological root of each word, bringing the different variations of the same word to the same common denominator (e.g. “books” become “book”, “contacted” becomes “contact”). The syntax information (“,”, “.”) is also removed. The paragraph structure is retained.

2.2 Step 2: Stop words Removal

The words that function as liaisons and that do not carry any additional meaning are removed from the text (e.g. “is”, “are”, “a”, “the”, etc.) The users can adjust the list of stop words used at that point using the Settings pane in the software or apply the tf–idf method to automatically generate the stopwords relevant to a document corpus.

The result of applying the steps 1 and 2 above is a word sequence, e.g. a sentence that reads:

Before you discuss the resolution, let me place before you one or two things, I want you to understand two things very clearly.

becomes

discuss resolution place thing want understand thing clear

2.3 Step 3: Text-to-Network Conversion

The text is then converted into a directed network graph. The normalized words (lemmas) are the nodes in the network graph and their co-occurrences are the edges.

Two consecutive scans are performed. The first scan creates the connections (graph edges) between the lemmas (nodes) that appear next to each other (bigrams) with the edge weight value of 3. The second scan uses a window of 4 lemmas (4-grams) to add a second layer of connections. The connections between the lemmas that are are only 1 word apart get the edge weight value of 2, while the lemmas that are separated by 2 words get the edge weight value of 1:

[discuss → (weight: 3) → resolution], [discuss → (weight: 2) → place], [discuss → (weight: 1) → thing],
[resolution → (weight: 3) → place], [resolution → (weight: 2) → thing], [resolution → (weight: 1) → want], etc.

These connections are encoded as edges on a directed text graph and are given the proper weight during analysis and visualization. Optionally, the links between the paragraphs can be made, where the last word of the previous paragraph links to the first word of the next one with the edge weight value of 1.

All the connections are unique and are saved even if they appeared earlier in the text. Later, for the purposes of graph visualization and community detection, the weights of the parallel edges are summed up.

2.4 Step 4: Extracting Most Influential Keywords Using Betweenness Centrality

We then apply a ranking algorithm to identify the nodes (the words) with the highest betweenness centrality – these are the nodes that appear most often on the shortest paths between any two randomly chosen nodes in the network [35], [36]. Those words are central for meaning circulation and can be seen as the meaning junctions within the discourse — they are shown bigger on the graph.

In the example Gandhi’s text, we obtain these most influential words that have the highest betweenness centrality[3]:
struggle - ahimsa - people - freedom (see Figure 1)

2.5 Step 5: Topic Modelling Using Community Detection and Force-Atlas Layout

We then apply community detection algorithm [37], [38] based on modularity. This is an iterative algorithm that detects the groups of nodes that are more densely connected together than with the rest of the network. As a result, we obtain the groups of nodes (words) which tend to appear together in the text: topical clusters. We then apply Force-Atlas algorithm [39], which aligns densely connected clusters together while pushing the most connected nodes apart, so that the network structure is more visible on the graph. We then get a visual network representation of the text with a clearly defined community structure (using both color and network topology) and the specific topical clusters (see Figure 1).

The topics identified[4] are displayed on the separate Analytics pane:
1 (green): struggle - freedom - fight
2 (brown): people - british - hatred
3 (light green): ahimsa - resolution - faith
4 (orange): india – present - big

2.6 Step 6: Summarization, Visual Text Search and Nonlinear Reading

The graph visualization is interactive. After the main topics are identified, InfraNodus highlights the parts of the text that contain the highest concentration of topics identified (using the Analytics or Essence panes).

The user can also select any nodes on the graph to find the part of the text that mentions the topics selected (see Figure 2). The user can also compare two or more different text graphs using the text “diff” feature.

Therefore, the graph can be used to read through the text in a non-linear way, following the concepts that are relevant — using the graph as a heuristic device to make sense of the text or to look at it from a new perspective.

[3] LDA method (also available in InfraNodus / Analytics pane) yields similar results for the top 4 keywords in this text: “struggle – people – british - ahimsa”.

[4] According to LDA, the top 4 topics are “ahimsa – resolution - present”, “british – people - hatred”, “struggle – freedom - today”, “power – india - congress”.

2.7 Step 7: Discourse Structure and the Measure of Discourse Bias

InfraNodus also identifies the structure of discourse based on the structure of its text network graph. The algorithm takes into account (1) the graph's modularity (M — a value higher than 0.4 indicates a highly pronounced community structure [38]); (2) the percentage of words in the topical cluster with the most nodes, or in the main connected (giant) component (C), if there are several connected components in the graph; (3) the dispersal of influence among the topical clusters (communities) measured in terms of Shannon entropy (E) for a string that contains a 1-letter code of the community assigned to each of the top 4 most influential keywords (e.g. if each of the top 4 nodes belongs to a different community A, B, C and D, the string will be 'ABCD' and the entropy in bits per symbol will be 2).

Based on these criteria a score is given to the text graph, which can be used to measure the level of bias in discourse:

- **Dispersed**: high modularity ($M > 0.65$) and ($C < 50\%$ and $E >= 1.5$) - the community structure is highly pronounced, the influential words are distributed among the communities. Therefore, the discourse has several topics, which are weakly related.
- **Diversified**: high modularity ($0.6 >= M > 0.4$) and ($C < 50\%$ and $E >= 1.5$) - pronounced community structure, the most influential words are distributed among the communities. Therefore, the discourse has several topics, each of them has a relatively high number of nodes in the graph, the topics are connected.
- **Focused**: medium modularity ($0.4 > M >= 0.2$ and $E > 0.5$) or ($M > 0.4$ and $C >= 0.5$ and $0.25 < E < 0.75$) - communities are present but not easily detectable, influential words may be concentrated around one topic, the discourse is focused on a certain perspective, the other perspectives are weakly presented.
- **Biased**: low modularity ($M < 0.2$), no detectable communities, influential words are concentrated around one subject - the discourse is biased

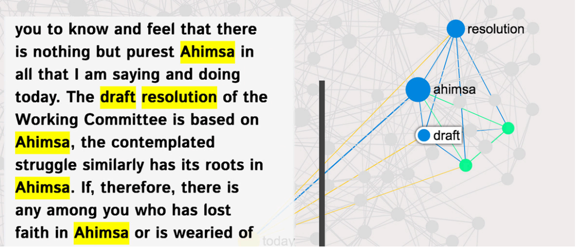
For example, the structure of the discourse in our example is "Diversified" meaning that it has several distinct topics and that those topics are still connected on the global level, the bias level is low (see Figure 1).

2.8 Step 8: Insight Generation using Structural Gaps

Graph visualization is also used to identify the structural gaps in the graph. This can be done both in a qualitative way by the researchers or automatically by the software's algorithm (detecting the distinct communities that are not well connected). It has been demonstrated [40], [41] that structural gaps are the parts of a graph that indicate the potential for new ideas. Text network visualization can be used to identify those gaps and to propose new connections between different ideas, topics and topical clusters [5].

In the example we use those are
"british" - *"people"* - *"hatred"* and *"struggle"* - *"freedom"* - *"fight"*

A quick analysis of the text snippets that include those topics and of the graph shows that there is a structural gap between the nationalist sentiment that Ghandi is talking about (in relation to the "hatred" for "British") and at the same time about the fact that, in his opinion, revolutions that are based on violence do not live up to democratic ideals. The algorithm shows that there is an important link between these two ideas (see Figure 3).



you to know and feel that there is nothing but purest Ahimsa in all that I am saying and doing today. The draft resolution of the Working Committee is based on Ahimsa, the contemplated struggle similarly has its roots in Ahimsa. If, therefore, there is any among you who has lost faith in Ahimsa or is wearied of

Figure 2: Visual graph text search feature.

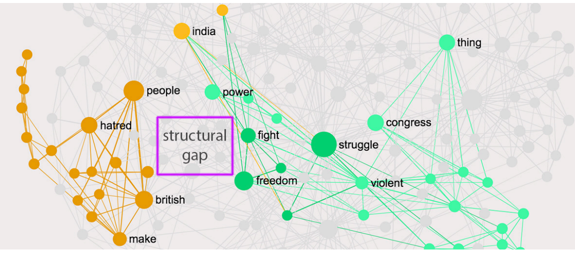


Figure 3: Structural gaps point out potential insights.

3 Conclusion and Practical Applications

In the previous section we demonstrated the implementation of text network visualization and analysis algorithm in InfraNodus software. We have shown how it can used to obtain a visual representation of text, to perform topic modeling, to extract relevant parts using the interactive interface and the visual search, to compare different texts, to analyze the structure of discourse, and to identify the structural gaps to generate new ideas in order to further develop the discourse.

The tool is currently used by researchers, marketing professionals, students, lawyers, artists and activists worldwide (20000 users a year according to Google Analytics for the online version as of December 2018) and it became first available in its beta version in 2014. The range of its practical applications is quite diverse: text categorization, search engine optimization, measure of bias, sentiment analysis, computer-assisted research and creative writing.

The structural gap identification feature can be potentially useful for enhancing recommender systems with non-trivial suggestions based on bridging the different clusters of disconnected datasets together, providing interesting insight for the users.

ACKNOWLEDGMENTS

This research was supported solely by the contributions made by the users of the online version of InfraNodus.

[5] We could also apply the same methodology to the topical clusters identified using LDA topic modelling (see footnote 4)

REFERENCES

[1] M. Minsky, "A Framework for Representing Knowledge," *Artificial Intelligence, Memo no. 306,* 1975.

[2] J. Sowa, "Semantic Networks," in *Encyclopedia of Cognitive Science*, 1992.

[3] Y. Liu, G. Zhao, G. Ma and Y. Bo, "The Effect of Mind Mapping on Teaching and Learning: A Meta-Analysis," *Standard Journal of Education and Essay,* vol. 2(1), p. 017– 031, 2014.

[4] M. Davies, "Concept mapping, mind mapping and argument mapping: what are the differences and do they matter?," *Higher Education,* vol. 62, no. 3, p. 279–301, 2011.

[5] J. Redford, K. Thiede, J. Wiley and T. Griffin, "Concept mapping improves metacomprehension accuracy among 7th graders," *Learning and Instruction,* vol. 22, no. 4, pp. 262-270, 2012.

[6] K. Carley, "Extracting Team Mental Models through Textual Analysis," *Journal of Organizational Behavior,* 1997.

[7] D. Jonassen and Y. Cho, "Externalizing Mental Models with Mindtools," in *Understanding Models for Learning and Instruction*, Boston, Springer, 2008.

[8] S. S. Sonawane and P. A. Kulkarni, "Graph based Representation and Analysis of Text Document: A Survey of Techniques.," *International Journal of Computer Applications,* vol. 96, no. 19, 2014.

[9] E. Castillo, O. Cervantes and D. Vilariño, "Text Analysis Using Different Graph-Based Representations," *Computación y Sistemas,* vol. 21, no. 4, 2018.

[10] R. Popping, "Knowledge Graphs and Network Text Analysis," *Social Science Information,* 2003.

[11] T. Heyman, B. Van Rensbergen, G. Storms, K. A. Hutchison and S. De Deyne, "The influence of working memory load on semantic priming," *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. 41, no. 3, pp. 911-920, 2015.

[12] S. Evert and G. Lapesa, "Evaluating Neighbor Rank and Distance Measures as Predictors of Semantic Priming," *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics,* p. 66–74, 2013.

[13] J. Bullinaria and J. Levy, "Extracting semantic representations from word co-occurrence statistics: A computational study," *Behavior Research Methods,* no. 39, 2007.

[14] F. D. Malliaros and K. Skianis, "Graph-Based Term Weighting for Text Categorization," *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '15,* p. 1473–1479, 2015.

[15] B. Paley, "TextArc text visualization software," 2002.

[16] S. Corman, T. Kuhn, D. Mcphee and K. Dooley, "Studying Complex Discursive Systems: Centering Resonance Analysis of Communication," *Human Communication Research,* vol. 28, no. 2, p. 157–206, 2002.

[17] D. Paranyushkin, "Identifying the pathways for meaning circulation using text network analysis.," *Nodus Labs,* 2011.

[18] F. Rousseau and M. Vazirgiannis, "Graph-of-word and TW-IDF: new approach to ad hoc IR," *Proceedings of the 22nd ACM international conference on information & knowledge management - CIKM '13,* p. 59–68, 2013.

[19] D. Paranyushkin, "Addresses to the Federal Assembly of the Russian Federation by Russian presidents, 2008–2012: comparative analysis," *Russian Journal of Communication,* vol. 5, no. 3, pp. 265-274, 2013.

[20] F. D. Malliaros and K. Skianis, "Graph-Based Term Weighting for Text Categorization," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*, 2015.

[21] D. Paranyushkin, "Direct Visual Feedback on the Process of Ideation using Text Network Graphs Encourages a more Coherent Expression of Ideas," *Nodus Labs,* 2018.

[22] H. Lee, J. Kihm, J. Choo, J. Stasko and H. Park, "iVisClustering: An Interactive Visual Document Clustering via Topic Modeling," *Computer Graphics Forum,* vol. 31, no. 3, p. 1155–1164, 2012.

[23] B. Gretarsson, J. O'Donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman and P. Smyth, "TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling," *ACM Transactions on Intelligent Systems and Technology,* vol. 3, no. 2, p. 1–26, 2012.

[24] J. Choo, H. Kim, E. Clarkson, Z. Liu, C. Lee, F. Li and H. Park, "VisIRR: A Visual Analytics System for Information Retrieval and Recommendation in Large-Scale Document Data," *ACM Transactions on Knowledge Discovery from Data,* vol. 20, 2018.

[25] T. K. Landauer, P. W. Foltz and D. Laham, "An introduction to latent semantic analysis," *Discourse Processes,* no. 25, pp. 259-284, 1998.

[26] G. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999.

[27] W. Li and A. McCallum, "Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[28] D. Blei, A. Ng and M. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* vol. 1, no. 03, pp. 993-1022, 2003.

[29] J. Chang and D. Blei, "Hierarchical Relational Models for Document Networks," *Annals of Applied Statistics,* vol. 4, no. 1, p. 124–150, 2010.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Advances in Neural Information Processing Systems,* no. 26, p. 3111–3119, 2013.

[31] Y. Goldberg and O. Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," 2014.

[32] C. E. Moody, "Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec," 2016.

[33] D. Blei, "Probabilistic Models of Text and Images," p. 2004.

[34] M. Gandhi, "Quit India," 1942. [Online]. Available: https://en.wikipedia.org/wiki/Quit_India_speech. [Accessed 15 09 2018].

[35] L. Freeman, "A Set of Measures of Centrality Based on Betweenness," *Sociometry,* vol. 40, no. 1, pp. 35-41, 1977.

[36] A. Brandes, "Faster Algorithm for Betweenness Centrality," *Journal of Mathematical Sociology,* vol. 25, no. 2, pp. 163-177, 2001.

[37] S. Fortunato, "Community Detection in Graphs," *Complex Networks and Systems Lagrange Laboratory,* 2010.

[38] V. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2008, 2008.

[39] M. Jacomy, T. Venturini, S. Heymann and M. Bastian, "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software," *PLoS ONE,* vol. 9, no. 6, 2014.

[40] R. Burt, Structural holes: The social structure of competition, 1992.

[41] L. Noy, Y. Hart, N. Andrew, O. Ramote, A. Mayo and U. Alon, "A quantitative study of creative leaps," in *International Conference on Computational Creativity*, 2012.