Visualizing Topic Models Generated Using Latent Dirichlet Allocation

Ashwinkumar Ganesan, Kiante Branley

Abstract—Topic Modeling is a set of statistical methods used to find ”topics” in a given document corpus, where a *topic* is defined as a word distribution. Topic Modeling opens the doors to a set of interesting questions such as the various topics in documents & which documents are associated which kinds of topics. One of the problems with statistical methods is that the word distributions generated may not interpretable by the user. Our project, *LDAExplore*, discusses the visualization designed, to look at topic models and how the user can interact with these statistical methods and provide feedback to improve the underlying model. Latent Dirichlet Allocation (LDA) is one of the basic methods to find such hidden topics that has been used in the project. To validate our design, we have used the *abstract*'s of 322 *Information Visualization* papers, where every abstract is considered a *document* and generated topics which are then explored by users.

Index Terms—DataViz, InfoViz, LDA, Latent, Dirichlet Allocation, Parallel Coordinates, Treemap, Topic Modeling.

1 Introduction

A Large body of human knowledge has traditionally been stored in the form of documents. These documents maybe in the form of books, magazines, journals. With the advent of the digital age, more and more content is stored digitally and with the cloud computing becoming increasingly common, we stored a lot of information online. We have websites that present this knowledge online in the form of HTML pages and are looking at this information being presented on mobile platforms such as IOS and Android in the form of apps. This growing knowledge base brings with it a unique set of challenges such as searching through a large set of documents for a specific piece of information, grouping similar documents together, looking at how these documents and their underlying content change over time. These changes could be the common theme or the language that is utilized in these documents.

Topic Modelling tries to automate the process of extracting topics from documents and annotate them with semantic information [2]. It is a set of statistical algorithms that extract correlated words from documents. These extracted word sets are called *Topics*. They are later annotated with semantic information, so that they are easily understood by people. For e.g. consider a a word set extracted such as *visualization, sets, clusters, infoviz, interfaces* from a document then we have a general notion that this word set represents the topic *Information Visualization*. In this example, the ”topic” generated using a topic modelling algorithm is the word set and *Information Visualization* is the semantic ”topic name” annotated by the user. Latent Dirichlet Allocation (LDA) [1] is one of the common methods to perform topic modelling on a given corpus of documents.

LDA generates viz. two types of distributions i.e. the topic distribution for each document in the set and the word distribution for each topic. These distributions can be changed by tweaking some of the underlying parameters. Our project *LDAExplore*, tries to give visual cues about how these distributions look, and how the topics and documents are interrelated at the corpus level and for groups or individual documents. Also, it gives the users the ability query documents for specific words based on topics and see how correlated a topic is to a document. Some of the main concerns while creating the design are that visual should be able to work with a large set of documents while providing the ability to see individual and group relations. The number of topics, though, is considered to be a smaller set as compared to the number of documents.

2 Related Work

There are quite a few ways of looking that the problem of how to represent distributions to show their interrelations. Some of the methods include looking at distributions graphs or sets [5]. In these visualizations, the distributions are converted to sets transforming them to a hierarchy of nodes. A standard technique to represent sets is to create an *Euler* diagram [5] or a variant of it. Region-based overlays can be used such as *Bubble Sets* and *Texture Splatting* or Line-based overlays like *LineSets* and *Kelp Diagrams* [5]. Some of the other methods include *Node-Link* diagrams and *Force-Directed* graphs. Standard Node-link diagrams include *Jigsaw*, Anchored maps and *PivotPaths*. These design patterns invariably limit the number of nodes that can be represented in a visual without applying some aggregation method which clusters nodes or edges together.

- *Ashwinkumar Ganesan is with the Department of Computer Science and Electrical, UMBC.*
  Kiante Branley is with the Department of Computer Science and Electrical, UMBC.
  E-mail: {gashwin1, bran4}@umbc.edu

2.1 Subsection Heading Here

Subsection text here.

2.1.1 Subsubsection Heading Here

Subsubsection text here.

3 Conclusion

The conclusion goes here.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

Acknowledgments

The authors would like to thank...

References

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

[2] Blei, David M. ”Probabilistic topic models.” Communications of the ACM 55.4 (2012): 77-84.

[3] Pan, Shimei, et al. ”Optimizing temporal topic segmentation for intelligent text visualization.” Proceedings of the 2013 international conference on Intelligent user interfaces. ACM, 2013.

[4] Yang, Yi, et al. ”Active Learning with Constrained Topic Model.”

[5] Alsallakh, Bilal, et al. ”Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges (Supplementary Material).”

[6] Cui, W., Liu, S., Wu, Z., & Wei, H. (2014). How hierarchical topics evolve in large text corpora.

PLACE
PHOTO
HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.