

**Geovisualizing concepts in texts**

**Transforming texts to maps:**

Mahesh Thapa

**Transforming Texts to Maps: Geovisualizing concepts in texts**

Dissertation Supervised by

**Francisco Ramos, PhD**

Associate Professor

Department of Computer Languages and Systems

University Jaume I,

Castellón, Spain

Dissertation co-supervised by

**Oscar Belmonte Fernández, PhD**

Associate Professor

Department of Computer Languages and Systems

University Jaume I

Castellón, Spain

**Roberto Henriques, PhD**

Assistant Professor

NOVA Information Management School

University of New Lisbon

Lisbon, Portugal

February 2017

# ACKNOWLEDGMENTS

I would like to thank all those without whom this dissertation would not have been possible.

I would like to express my sincere gratitude to Dr. Francisco Ramos, Dr. Oscar Belmonte Fernández and Dr. Roberto Henriques for supervising the dissertation work.

I am grateful to the Erasmus Mundus program for the scholarship to purse the course M.Sc in Geospatial Technologies without which it would be a far-fetched dream.

Finally, I would like to take this opportunity to express my love and gratitude to my family who have been supportive to me in all the choices that I have made.

**Transforming Texts to Maps: Geovisualizing concepts in texts**

# ABSTRACT

Unstructured textual data is one of the most prevalent form of communicating and preserving information. There has been a massive surge in the rate of generation of textual information especially after the adoption of Web 2.0, thanks to the popularity of blogs and social media. Techniques in text mining such as topic modelling have been developed to discover information from large collection of texts. Topic modelling is used to identify latent topics from such collections which is useful in acquiring information about the contents of the collection at a high level. While the generated topics are useful in acquiring a topic wise insight of the collection of texts, the contribution of location information to each topic could be exploited to get a spatial insight of the information in the collection of text. In this research, topic modelling of collections of unstructured text was done. The contribution of each of the location component on every topics generated were computed. These contributions were geovisualized which provides spatial insight to the contents of the text.

# KEYWORDS

Topic Modelling

Geoparsing

Natural Language Processing

Geovisualization

Spatial Insight

# ACRONYMS

**API -** Application Programming Interface

**NLP** - Natural Language Processing

**LDP -** Latent Dirichlet Allocation

**LSI –** Latent Semantic Indexing

Table of Contents

[ACKNOWLEDGMENTS iii](#_Toc504695382)

[ABSTRACT iv](#_Toc504695383)

[KEYWORDS v](#_Toc504695384)

[ACRONYMS vi](#_Toc504695385)

[INDEX OF FIGURES ix](#_Toc504695386)

[INDEX OF TABLES x](#_Toc504695387)

[1. INTRODUCTION 1](#_Toc504695388)

[1.1 Background 1](#_Toc504695389)

[1.2 Aim and objective 2](#_Toc504695390)

[2. Literature review 3](#_Toc504695391)

[2.1 Text Mining 3](#_Toc504695392)

[2.2 Topic Modeling 3](#_Toc504695393)

[2.3 Natural Language Processing 4](#_Toc504695394)

[2.4 Extracting Location Information from Text 4](#_Toc504695395)

[2.5 Relevant works in text visualization 5](#_Toc504695396)

[3. Methodology 6](#_Toc504695397)

[3.1 Building Corpus from text collections 6](#_Toc504695398)

[3.2 Extraction of Topic Model 8](#_Toc504695399)

[3.3 Identification of location information from the collection of text 9](#_Toc504695400)

[3.4 Geovisualization 9](#_Toc504695401)

[4. Results 10](#_Toc504695402)

[4.1 Overview 10](#_Toc504695403)

[4.2 Topic Models 10](#_Toc504695404)

[4.3 Geovisualization 11](#_Toc504695405)

[5. CONCLUSIONS 13](#_Toc504695406)

[6. FUTURE WORK 14](#_Toc504695407)

[7. References 15](#_Toc504695408)

[8. APPENDICES 16](#_Toc504695409)

# INDEX OF FIGURES

[Figure 1: Methodology 6](#_Toc504678651)

[Figure 2 : Visualization of topics (Topic 1 highlighted in red) using pyLDAvis 10](#_Toc504678652)

[Figure 3: Visualization of topics (Topic 10 in red) using pyLDAvis 11](#_Toc504678653)

[Figure 4: Screenshot of Geovisualization of Topic Number 0 11](#_Toc504678654)

[Figure 5: Screenshot of Geovisualization of Topic 9 12](#_Toc504678655)

# INDEX OF TABLES

# INTRODUCTION

## 1.1 Background

Out of many ways of expressing and storing information, textual format is by far the most dominating one. A query in google for number of books at the time of writing this document returns around 130 million. Besides books, textual information are also expressed in the form of newspapers, magazines, letters, etc. Even before the advent of internet technology, information in the form of text was already overwhelming. After the advent of internet technology, the generation and circulation of textual information has exploded. Besides the digitization of traditional form of textual information such as books into e-books, magazines into e-magazines and letters into emails, newer sources of textual information were devised such as web pages and blogs. A much bigger surge was seen in amount of textual information with the paradigm shift from Web 1.0 to Web 2.0. Web 2.0 allowed users not the opportunity to not only view and consume the information but also create and post their own content (O'Reilly, 2012). After the adoption of Web 2.0, the dramatic rise in textual information came along with the rise of social networking sites such as Facebook and microblogging sites as Twitter.

With the advent of internet technology, there has been massive surge in textual information. While information definitely is the key to proper decision making, huge amount of information can actually be detrimental (Buchanan & Kock, 2001). First, larger amount of information demands larger amount of resources for processing it. Secondly, humans only have a limited consumption capacity regarding information. While it was considered that good decisions came from considering all the information, it is no longer a rational choice given the vast amount of information (Etzioni, 1989). The vast amount of information and limited capacity to comprehend it, stimulated the development of tools and techniques to discover new information from unstructured text. This process of discovering information from text have matured to a new field, which is now known as text mining or text data mining. Text data mining can also be considered as exploratory data analysis which is useful in discovering unknown information from the texts (Hearst, 1999).

While text data mining has been applied in several fields for exploratory data analysis from unstructured text, a combination of text mining with emphasis on location component will add an additional dimension for discovering information. The location component in unstructured data is quite strong as captured by the common phrase in geospatial sector, “80% of data have location component”. With such a strong component of unstructured text being geographic, there is little doubt that text mining with a geospatial focus would prove useful. It is not a question of will it be useful but how can it be made useful. One of the strong benefits that could possibly be exploited by considering the location component is that the discovered information can be located on a map. As expressed by Russian writer Ivan Turgeney, “The drawing shows me at one glance what might be spread over ten pages in a book.” A combination of text mining and geovisualization provides more insight into the unstructured text for discovering information from it.

## 1.2 Aim and objective

The primary aim of this research is to discover hidden concepts in collections of texts and geovisualize these concepts which helps in providing spatial insight into the texts. Using topic modelling, one of the many techniques in text mining, a topic model is prepared from collections of texts. The topic model consists of number of topics which essentially represent the concepts in the texts. The contribution of location components that are present in the collection of the text are computed for each topic. These contributions are geovisualized using geovisualization libraries.

The objectives of the research are as follows:

1. Extract topic model from collection of unstructured texts
2. Compute the contribution of location components in the texts to each of the generated topics
3. Geovisualize the contribution of location components in each topic

# Literature review

## 2.1 Text Mining

The massive rate of text data generation along with already existing text data is one of the biggest impetus to the development of text mining. Text data constitutes a large amount of useful information. However, given the huge amount of text, manual extraction of information from texts is both costly as well as time-consuming. Also, humans have only so much capacity to consume information and can be overloaded with information leading to bad decision making (Buchanan & Kock, 2001). Hence, computer based automatic methods are advisable for extracting information from texts. However, information in texts based on natural language are in a form that it is anything but easy for machines to extract the information (Hearst, 1999). Understanding and making sense of natural language is a trivial matter to humans but making machines do the same is one of the most challenging problem.

Several text mining algorithms have been developed which discovers and extracts information. The algorithms that have been developed are oriented towards fulfilling the following functions: text categorization; text clustering; concept mining; information retrieval and information extraction(Ghosh, Roy, & Bandyopadhyay, 2012). The domain of application of text mining is limited only by the presence of text documents but the domains that has seen its largest application are as follows: media; telecommunications; finance; public administration; research and healthcare(Vishal Gupta, 2009).

## 2.2 Topic Modeling

Topic Modeling is used to discover the concepts in a collection of text documents. The basic premise in topic modeling is that a collection of text document contains various hidden topics. And each document contains one or more topics at varying proportions(Zhao et al., 2015). Several algorithms have been proposed for topic modeling. Some of the common algorithms used for topic modelling are as follows: Latent Semantic Indexing (LSA), Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). LDA is a generative probabilistic method for discovering topics in which each document in a collection is modelled as a finite mixture of topics(D. M. Blei et al., 2003). It is the most common method of topic modelling(Zhao et al., 2015). There are several variations of LDA that it has actually acted as the impetus for development of other topic model(D. M. Blei & Lafferty, 2009).

The popular topic modelling software Gensim has implementations for LSA and LDA. and HDP. In this study, LDA has been implemented to generate topic models from text collections. The Gensim library uses a variation of LDA based upon the paper Online Learning for Latent Dirichlet Allocation which allows for handling of large amount of document collection including data that arrives in stream(Hoffman, Blei, & Bach, 2010).

## 2.3 Natural Language Processing

Natural Language Processing is a computerized approach to analyze the texts. Some of the most common tasks in natural language processing are: word tokenization, sentence tokenization, part of speech tagging, named entity recognition, dependency parsing, coreference resolution, etc. Natural language processing provides tools and techniques for text mining as well as topic modelling.

Many software have been developed for natural language processing. Some of the most widely used software in natural language processing are listed below:

1. Natural Language Toolkit
2. Spacy
3. Stanford Core NLP

Natural Language Toolkit and Spacy are implement in Python. Stanford Core NLP is implemented in Java.

## 2.4 Extracting Location Information from Text

It is essential to extract location information from text in order to geovisualize the texts. While it is a mundane task for humans, automatic location extraction is a challenging field with a large amount of research work. One of the very active research and development field in this context is Named Entity Recognition (NER). NER identifies words that denote person, organization, location, object, etc. Different NER implementations have different classes of entities(Atdaǧ & Labatut, 2013).While NER identifies several classes of entities, only entities having location information are necessary for geovisualization. Almost all popular natural language processing software have facility for recognizing named entities. While named entity recognizes the names entities with locations, it is also necessary to extract the geographic location name of the entity. Digital gazetteers are specifically constructed to have unambiguous location information. Digital gazetteers contain structured information about geographic location. Digital gazetteers are particularly useful for automated georeferencing of location information in a text and termed as geoparsing(Goodchild & Hill, 2008). In this study, Geoparse.io is used for geoparsing the text. It is a RESTful web API that returns the information about the locations on the request text as GeoJSON. The Geoparse.io web API uses GeoNames geographical database as digital gazetteer.

## 2.5 Relevant works in text visualization

There are massive number of studies and developments in visualizing texts. (Cao & Cui, 2016) made a comprehensive review of more than 200 papers based upon text visualization techniques accumulated in Text Visualization Browser (*http://textvis.lnu.se/).* The paper has identified five categories of text visualization:

1. Visualization of document similarity.
2. Visualization for revealing content of the document
3. Visualization of sentiments and emotions in the text
4. Visualization of the corpus
5. Visualization of domain-specific rich-text corpus

These techniques cover a large and wide sectors of visualizing texts. However, even among such a large number of visualization techniques, texts are not visualized with focus on the location. All of these visualization techniques intend to answer “What “is in the text but fail to answer the “Where” portion of the text. There is a compelling motivation for development of text visualization technique that combines both the textual and spatial components of text data and can answer both “what” is in the text as well as “where” component of the text.

Studies that have considered both topic modelling and location component have primarily intended to improve the results of topic modelling mostly by segregating the texts based upon the location. (Hu & Ester, 2013) used locations of posts on social media to model user profiles using topic modelling and spatial location for improving location recommendation. (Pölitz, 2015) have

In this study, a new technique is demonstrated that visualizes large collection of texts by extracting the concepts (topics) covered in the collection and individually mapping the topics based upon the influence of spatial location on each topic.

# Methodology

The overall flow diagram of the methodology implemented in this study is shown in the figure below.

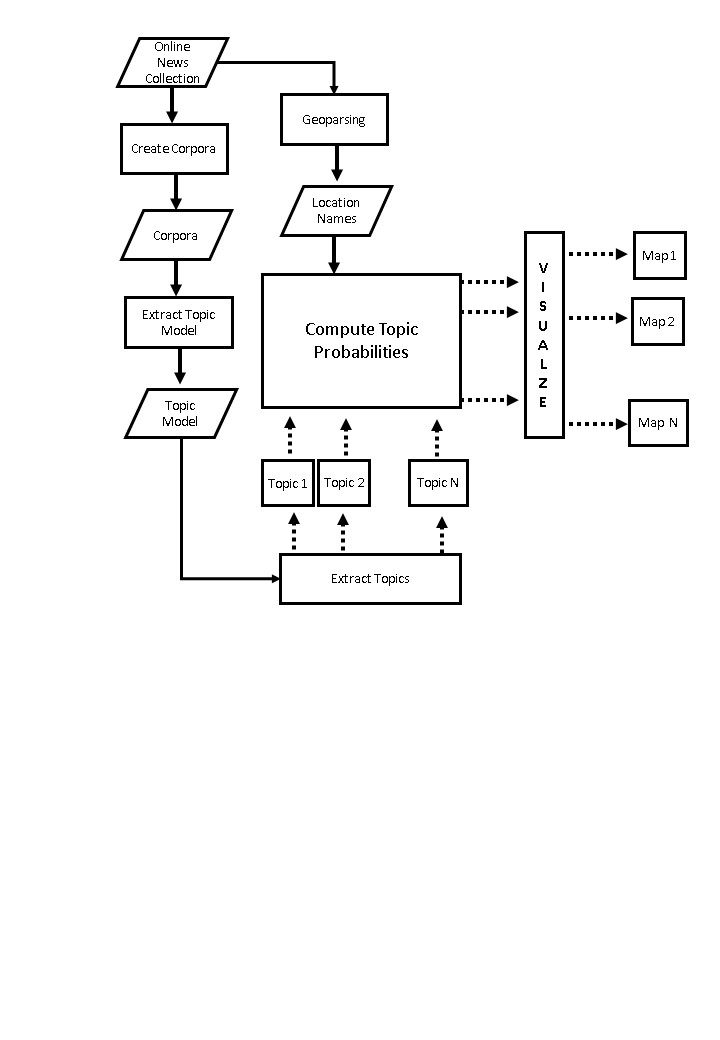


Figure 1: Methodology

## 3.1 Building Corpus from text collections

In this study, different type of collections of texts are considered for visualization. One of them is the collection of news articles related to earthquake in Nepal. These articles were collected manually from different news sources. Corpus is generally understood as a large collection of text documents. However a corpus can have other accompanying information. Such as annotations of named entities, part of speech, etc. While there exists thousands of corpus, these are created for certain research purposes and most are not available publicly(Xiao, 2010). The specific purpose of corpus creation means that although it is time consuming, specific research purpose demands creation of own corpus. In this study, news articles relevant to various fields are chosen to build corpus.

In this study, Gensim is used for corpus building. Gensim stands out among natural language processing software as it focuses particularly on topic modelling. It is implemented as a Python library. The strength of Gensim is that it can process large amount of text data even with limited computing power. The algorithms implemented in Gensim are memory independent. The source code of Gensim is hosted in GitHub and can be accessed at [*https://github.com/RaRe-Technologies/gensim*](https://github.com/RaRe-Technologies/gensim).

Turning text to corpus involves steps which are subject to the purpose of the library. However, the common task is to covert the text into machine readable form such that it is simpler and facilitates faster computation. Computation of texts in its native format is slow for computers. Hence, texts are converted into vectors for faster computation. There are several methods in which texts can be converted to vectors. Some of the methods consider the grammar and the word order while other methods only consider the frequency of words. Again, the choice of method depends upon the purpose of the study. In this study, bag-of-words (BOW) is implemented for converting texts into vectors. The bag-of-words method disregards the grammar as well as the order of the words. It only considers the frequency of words in the text collection. According to this method, every word in a collection of text is given a unique integer id and the number of appearances of the word in the collection is counted. It is a common method implemented in topic modelling and is also implemented in this study.

Before converting words to vectors, the text must be preprocessed. The importance of preprocessing cannot be emphasized more as it is what prevents an input being a garbage. For preprocessing, several natural language processing tools are required. In this study, Spacy is used for this purpose. Spacy is free and open source library for natural language processing. It is also implemented in Python. The source code of Spacy is also open in GitHub and can be accessed at [*https://github.com/explosion/spaCy*](https://github.com/explosion/spaCy). The choices of preprocessing steps again depend upon the purpose of the study. Preprocessing steps that are best fit for a purpose could produce garbage for other purposes. The following preprocessing steps were implemented in this study:-

1. Text Lemmatization

Lemmatization converts the words into its canonical or citation form(Bird, Klein, & Loper, 2009). Example: The lemmatization of swims and swam is swim.

1. Removal of stop-words

Stop-words are such words in a text which do not add any value to the purpose of the study. Again, this depends upon the purpose of the study. Each natural language processing software come with their own set of stop-words. Although, there is some overlap between them, there is no consensus in a universal set of stop-words. Many a times, defining stop-words is an iterative process. The results are examined and the words that add no meaning to the results are added to the list of stop-words.

1. Remove numbers, punctuation marks, symbols

Numbers, punctuation marks and symbols are removed from the text as these do not add any value to the topics.

1. Detect bigrams and trigrams

A combination of words such as climate change, tectonic plate, car race, ministry of education, etc., provide different meaning in combination than the individual words. These words, if treated as separate words, would provide different insight than what is intended in the texts. These words are called n-grams. Bigrams are combination of two words and trigrams are combination of three words. N-gram identification is especially important because the study has implemented bag-of-words model for transforming words to vectors as bag-of-words model do not preserve the order of the words. Bigrams and trigrams are highly prevalent in names of locations such as Pacific Ocean, Suez Canal, Kathmandu Valley, Grao de Castellon, etc. As these locations are to be mapped, bigrams and trigrams are identified in the texts and are processed as single entity. In this study, only bigrams and trigrams were considered. However, depending upon the language and the contents in the collection of texts, it is necessary to implement higher level of n-grams.

Once the process of preprocessing is complete, the text is converted to vectors based upon bag-of-words model and a corpus is built. In terms of Gensim, a corpus is a list of lists. The larger list represents the entire collection of texts. Each of the smaller lists represent individual document.

## 3.2 Extraction of Topic Model

Topic models are generated once the corpus are built. In this study, the topic model is generated using Latent Dirichlet Allocation (LDA) which is a generative probabilistic method of topic modelling. The basic assumption of LDA is that each document is a mixture of topics. And each topic is a probability distribution over words. The figure below conspicuously demonstrates the assumptions in LDA.

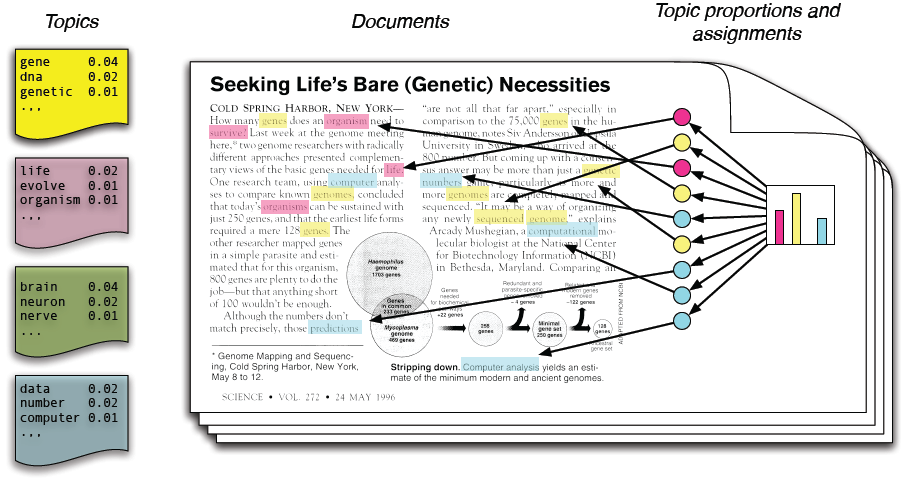


Figure 2: Intutitive digram for Topic Modeling using LDA (D. Blei, Carin, & Dunson, 2010)

The conceptual framework of LDA is shown in figure below.

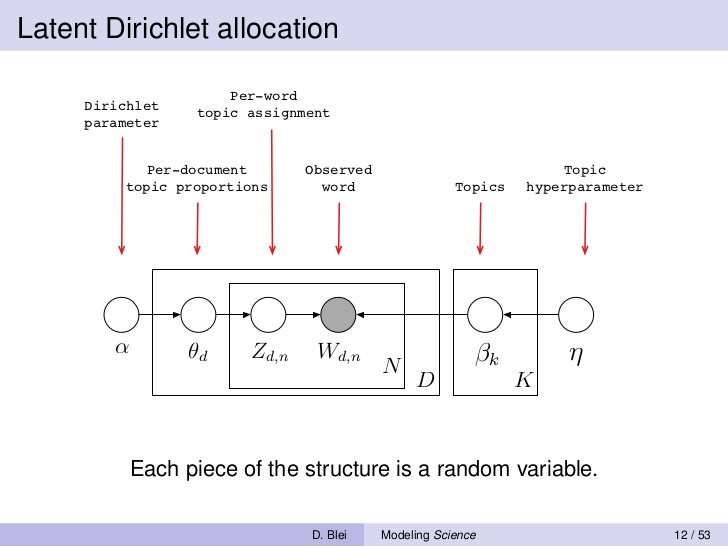


Figure 3: LDA Plate Notation(D. M. Blei & Blei, 2008)

The only model parameter for LDA in its implementation in Gensim is the number of topics to be generated. As of now, there is no straight forward and agreed upon method for specifying the number of topics. However, there are several measures of evaluating the topic models such as perplexity, topic coherence, human interpretability, etc. A python library, pyLDAvis is available for visualizing the topics. It can be accessed at [*https://github.com/bmabey/pyLDAvis*](https://github.com/bmabey/pyLDAvis)*.*

## 3.3 Identification of location information from the collection of text

The location information is critical in Geovisualizing the topics as it is used to georeferenced the topics. The location information is extracted by geoparsing the texts. In this study, Geoparser.io is used for geoparsing the texts. Geoparser.io is implemented as a RESTful web API. It is not free but provides 1000 free API calls per month. The response of the API call is in GeoJSON format. The response has the following information of interest: name of the location, country, state/province level administrative division, geographic feature type and coordinate.

## 3.4 Geovisualization

The mapping of topics is analogous to mapping of a thematic layer in cartography. The thematic layer is mapped according to its location. Then, it is symbolized according to one of its attribute value. In cases of continuous data, the attribute data is interpolated for visualization. For example a temperature map is visualized based upon finite number of measurements at various locations which are then interpolated. Similar approach is followed in mapping the topics generated from topic modelling. For this locations are the coordinates obtained by geoparsing all the text collection that were used for generating the topic model. To map a topic, say Topic A, the probability of the location name lying in Topic A was extracted from the topic. This probability is the attribute value of the location which is interpolated to geovisualize the topic. The geovisualization was done using leaflet geovisualization library which is implemented in JavaScript. The geovisualization was done as heatmaps.

# Results and Discussion

## 4.1 Overview

Topic models which essentially represent the hidden concepts in the collection of texts were generated using topic modelling in Gensim. The spatial locations in the texts were extracted using geoparsing. The probability of the spatial location lying in each topics were extracted. Using these probability as the attribute value, heatmaps were generated as geovisualization of the topics.

## 4.2 Topic Models

For the text collection of news related to earthquake in Nepal, topic model was generated using LDA in Gensim. The topic model is presented in the Appendix. A visualization of one of the topic generated using pyLDAvis is shown in the figure below.

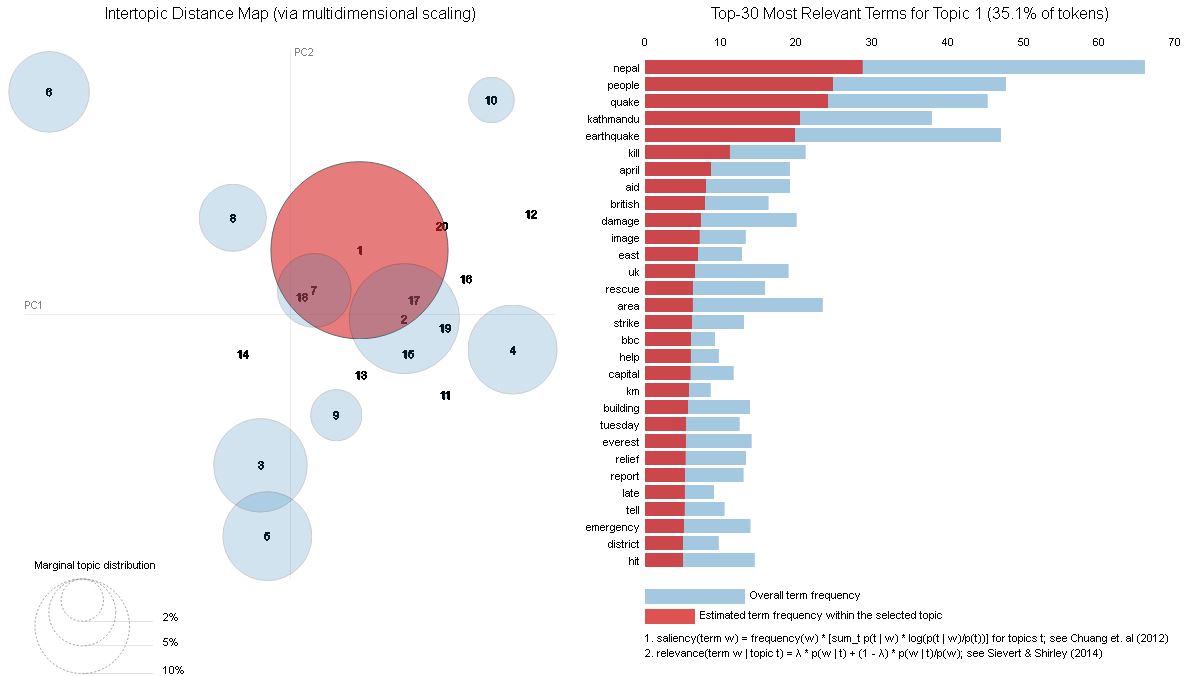


Figure 4 : Visualization of topics (Topic 1 highlighted in red) using pyLDAvis

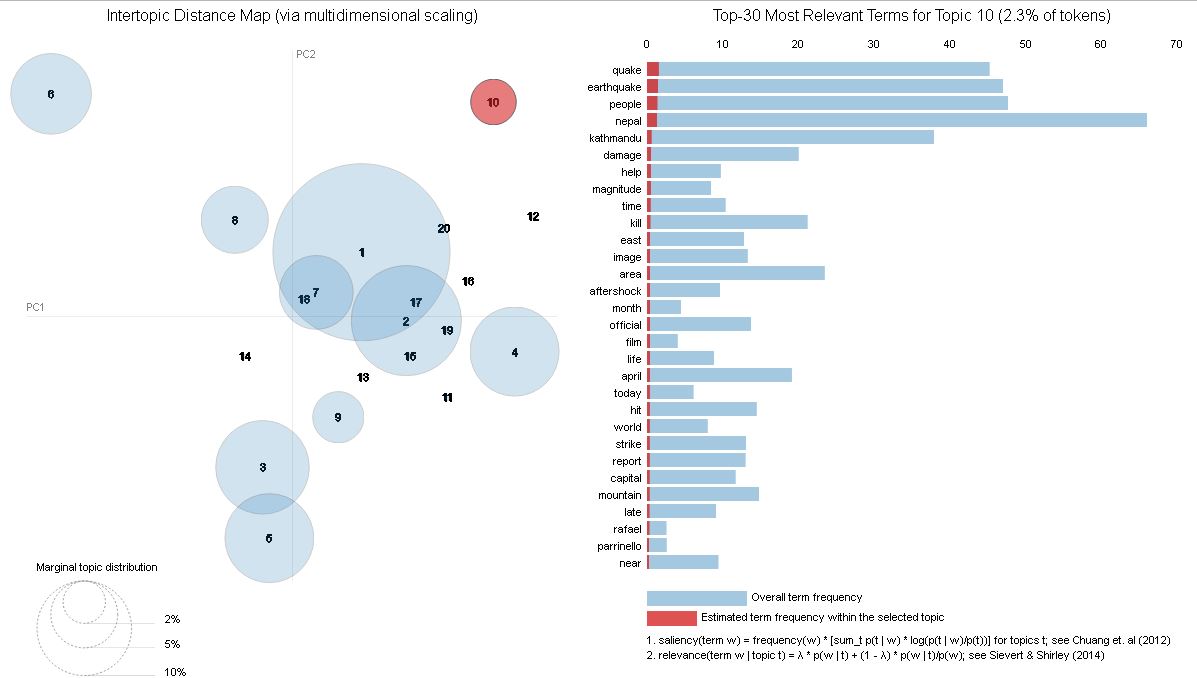


Figure 5: Visualization of topics (Topic 10 in red) using pyLDAvis

## 4.3 Geovisualization

The topics are geovisualized as heatmaps using leaflet geovisualization library. Screenshots of the geovisualization are shown below.

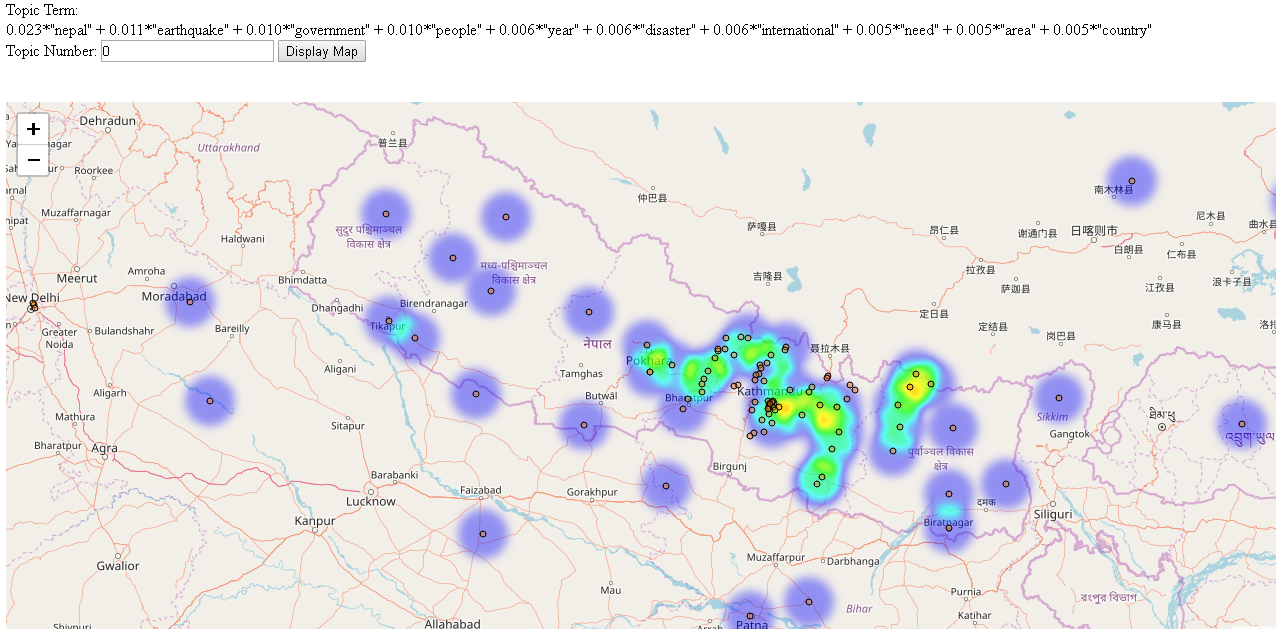


Figure 6: Screenshot of Geovisualization of Topic Number 0

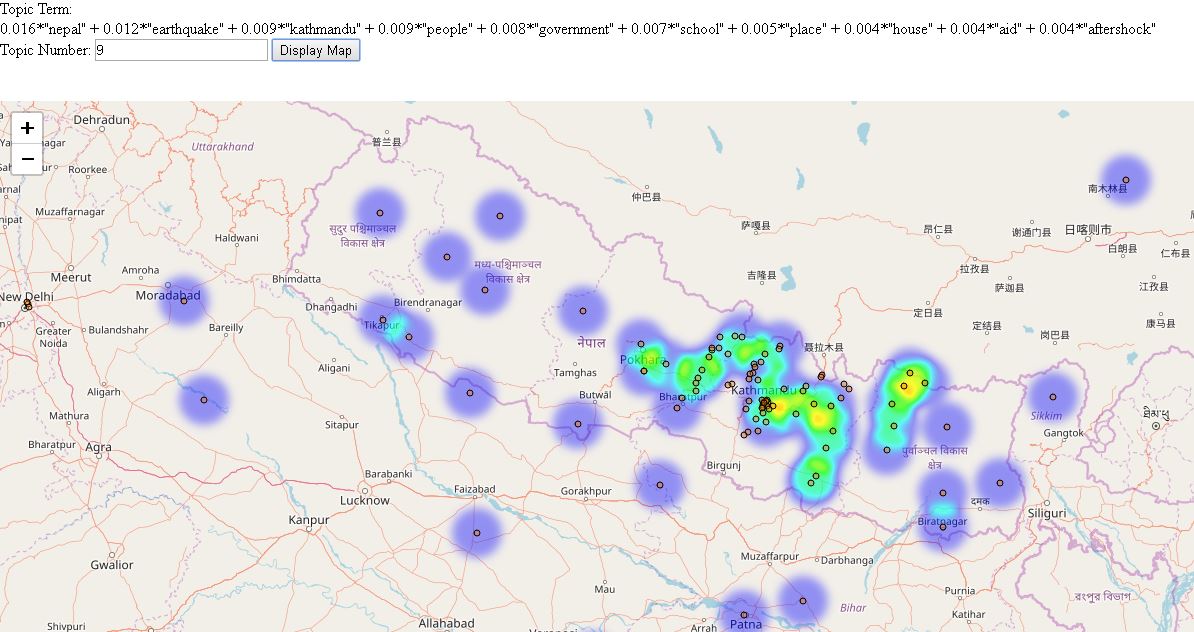
****

Figure 7: Screenshot of Geovisualization of Topic 9

# CONCLUSIONS

Topic models are useful in discovering topics from a large collection of texts. These topic models provide insights into the contents of the texts. However, they do not provide spatial insight into the texts. Since, textual data have a large spatial component, geovisualization of the topics have potential in providing insights of the text collection. In this study, topics extracted from topic models are successfully geovisualized based upon the probability of the location names lying in each subject.

# FUTURE WORK

# References

Atdaǧ, S., & Labatut, V. (2013). A comparison of named entity recognition tools applied to biographical texts. *2013 2nd International Conference on Systems and Computer Science, ICSCS 2013*, 228–233. https://doi.org/10.1109/IcConSCS.2013.6632052

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. *Text* (Vol. 43). https://doi.org/10.1097/00004770-200204000-00018

Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, *27*(6), 55–65. https://doi.org/10.1109/MSP.2010.938079

Blei, D. M., & Blei, D. (2008). Modeling Science Joint work with John Lafferty ( CMU ) Modeling Science. *Science*.

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., & Edu, J. B. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Blei, D. M., & Lafferty, J. D. (2009). Topic Models. *Text Mining: Classification, Clustering, and Applications*, 71–89. https://doi.org/10.1145/1143844.1143859

Buchanan, J., & Kock, N. (2001). Information Overload: A Decision Making Perspective (pp. 49–58). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-56680-6\_4

Cao, N., & Cui, W. (2016). *Introduction to Text Visualization*. https://doi.org/10.2991/978-94-6239-186-4

Ghosh, S., Roy, S., & Bandyopadhyay, P. S. K. (2012). A tutorial review on Text Mining Algorithms, *1*(4), 223–233.

Goodchild, M. F., & Hill, L. L. (2008). Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, *22*(10), 1039–1044. https://doi.org/10.1080/13658810701850497

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics -* (pp. 3–10). https://doi.org/10.3115/1034678.1034679

Hoffman, M. ~D., Blei, D. ~M., & Bach, F. (2010). Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems 23*, *23*, 856–864.

Hu, B., & Ester, M. (2013). Spatial topic modeling in online social media for location recommendation. *Proceedings of the 7th ACM Conference on Recommender Systems - RecSys ’13*, 25–32. https://doi.org/10.1145/2507157.2507174

Pölitz, C. (2015). Modelling Time and Location in Topic Models. In *Proceedings of the 2Nd International Conference on Mining Urban Data - Volume 1392* (pp. 95–96). Aachen, Germany, Germany: CEUR-WS.org. Retrieved from http://dl.acm.org/citation.cfm?id=3045776.3045791

Vishal Gupta, G. S. L. (2009). A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 17. https://doi.org/10.4304/jetwi.1.1.60-76

Xiao, R. (2010). 7 - Corpus Creation. *The Handbook of Natural Language Processing*, 147–165. https://doi.org/doi:10.1201/9781420085938-c7

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, *16*(13), S8. https://doi.org/10.1186/1471-2105-16-S13-S8

# APPENDICES