

# Masters Program in **Geospatial Technologies**



## Transforming texts to maps: Geovisualizing topics in texts

Mahesh Thapa

Dissertation submitted in partial fulfilment of the requirements  
for the Degree of *Master of Science in Geospatial Technologies*

# **Transforming Texts to Maps: Geovisualizing topics 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 opportunity to pursue the course M.Sc in Geospatial Technologies. It was also an opportunity to be among inspirational teachers and amicable and brilliant classmates.

Finally, I would like to take this opportunity to express my love and gratitude to my family.

# **Transforming Texts to Maps: Geovisualizing topics in texts**

## **ABSTRACT**

Unstructured textual data is one of the most dominant forms of communication. Especially after the adoption of Web 2.0, there has been a massive surge in rate of generation of unstructured textual data. While large amount of information is intuitively better for proper decision making, it also means that it becomes virtually impossible to manually process, discover and extract useful information from textual data. Several supervised and unsupervised techniques in text mining have been developed to classify, cluster and extract information from texts. While text data mining provides insight to the contents of the texts, these techniques do not provide insights to the location component of the texts. In simple terms, text data mining addresses “What is the text about?” but fails to answer the “Where is the text about?” Since textual data have a large amount of geographic content (estimates of about 80%), it can be safely reasoned that answering “Where is the text about?” adds significant insights about the texts. In this study, a collection of news articles from the year 2017 were analyzed using topic modelling, an unsupervised text mining technique. Topics were discovered from the text collections using Latent Dirichlet Allocation method, a popular topic modelling technique. Topics are probability distribution of words which correspond to one of the concepts covered in the text. Spatial locations were extracted from text documents by geoparsing them. Topics were geovisualized as interactive maps according to the probability of each spatial location word which contributed to the corresponding topic. This is analogous to thematic mapping in GIS. Coordinates obtained from geoparsed words provide basis for georeferencing the topics while the probability of such location words corresponding to the particular topics provide the attribute value for thematic mapping. An interactive geovisualization was constructed using leaflet library. A comparative analysis of the maps were made to see if they provided spatial context to the topics.

## KEYWORDS

Topic Modelling

Geoparsing

Natural Language Processing

Geovisualization

Spatial Context

## **ACRONYMS**

**API** - Application Programming Interface

**Kb** - Kilobytes

**NLP** - Natural Language Processing

**LDP** - Latent Dirichlet Allocation

**LSI** – Latent Semantic Indexing

## Table of Contents

ACKNOWLEDGMENTS .....	1
ABSTRACT.....	2
ACRONYMS .....	4
INDEX OF FIGURES .....	7
INDEX OF TABLES .....	8
1. INTRODUCTION .....	9
1.1 Background.....	9
1.2 Aims and objectives .....	10
2. Theoretical review.....	11
2.1 Discovering and extracting information from unstructured texts: Text Mining .....	11
2.2 Discovering concepts in texts: Topics and Topic Modeling.....	13
2.2.1 Latent Dirichlet Allocation (LDA) .....	14
2.2.2 Software Implementations for Topic Modelling.....	16
2.3 Natural Language Processing (NLP) .....	17
2.4 Extracting Location Information from Text.....	17
3. Relevant Works.....	19
3.1 Relevant works in text mining from newspaper articles.....	19
3.2 Relevant works in topic modelling .....	19
3.3 Relevant works in visualizing texts .....	19
4. Data .....	21
5. Methodology .....	23
5.1 Building Corpus of News Articles .....	24
5.2 Preprocessing Corpus of News Articles.....	24

5.3 Building Machine Readable Corpus and Dictionary .....	27
5.4 Extraction of Topic Model.....	29
5.5 Identification of location information from the collection of text .....	30
5.6 Geovisualization.....	30
6. Results and Discussion.....	32
6.1 Results of topics generated from LDA .....	32
6.2 Geovisualization results .....	36
6.3 Discussion .....	42
7. CONCLUSIONS.....	44
8. References .....	45
9. APPENDICES .....	47



## INDEX OF FIGURES

Figure 1: Overview of text mining methods (Source: <a href="http://chdoig.github.io/acm-sigkdd-topic-modeling/#/">http://chdoig.github.io/acm-sigkdd-topic-modeling/#/</a> ) .....	11
Figure 2 : Sample terms in topics(Source: (D. Blei et al., 2010)).....	13
Figure 3: Intuitive digram for Topic Modeling using LDA (D. Blei et al., 2010)....	15
Figure 4: LDA Plate Notation (D. M. Blei & Blei, 2008) .....	15
Figure 5: Chart of Overall Methodology .....	23
Figure 6 : Choropleth Map of Topic 1 .....	37
Figure 7: Choropleth Map of Topic 2 .....	37
Figure 8: Choropleth Map of Topic 3 .....	38
Figure 9: Choropleth Map of Topic 4 .....	38
Figure 10: Choropleth Map of Topic 5 .....	39
Figure 11: Choropleth Map of Topic 6 .....	39
Figure 12: Choropleth Map of Topic 7 .....	40
Figure 13: Choropleth Map of Topic 8 .....	40
Figure 14: Choropleth Map of Topic 9 .....	41
Figure 15: Choropleth Map of Topic 10 .....	41

## INDEX OF TABLES

Table 1 : Text mining functions and methods.....	12
Table 2 : Common topic modelling methods.....	14
Table 3: List of software for topic modelling .....	16
Table 4: List of popular NLP software .....	17
Table 5: List of popular geoparsing applications.....	18
Table 6: Categories and number of news.....	21
Table 7: Dictionary of Corpus built on BOW model.....	28
Table 8: Document and its representation in form of bag-of-words.....	28
Table 9: Text and visualization of geoparsed text .....	30
Table 10: Top 5 topic terms with 7 topics and 50 iterations.....	32
Table 11: : Top 5 topic terms with 7 topics and 400 iterations.....	33
Table 12: : Top 5 topic terms with 21 topics and 50 iterations.....	34
Table 13: : Top 5 topic terms with 21 topics and 400 iterations.....	35
Table 14: : Top 5 topic terms with 21 topics and 600 iterations.....	35
Table 15: Comparison between news categories, discovered topics and countries highlighted in the maps .....	42

# **1. 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 only the opportunity to 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 Turgenev, “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.

In this study, we demonstrate that spatial context can be provided to topics generated from topic modeling by mapping them based upon the probability of spatial terms in the topics. A series of topics and maps are prepared which in tandem provide not only insight into the content of collection of news but also provides spatial insight into the news.

## **1.2 Aims and objectives**

The primary aim of this research is to demonstrate that a tandem of topics discovered using topic modelling and geovisualization of individual topics can provide both information about the hidden concepts in the collection of text documents as well as provide spatial insight into the collection.

The objectives of the research are as follows:

- a. Extract topics from corpus of news collection using topic modelling.
- b. Extract probability of spatial terms in each topic and aggregate them to prepare interactive geovisualizations at the scale of country.
- c. Demonstrate that the geovisualizations provide spatial context to the topics through comparative analysis

## 2. THEORETICAL REVIEW

### 2.1 Discovering and extracting information from unstructured texts: Text Mining

The massive rate of text data generation along with already existing text data is one of the biggest impetus to the rapid development in the field of text mining. Text data constitutes useful information in various form of books, newspaper, tweets, posts, blogs, etc. 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 hidden information (Hearst, 1999). Understanding and making sense of natural language is a trivial matter to humans but making machines do the same is a formidable challenge.

Text mining was first introduced in (Feldman et al., 1998) as the technique to extract non-trivial information from text data(Allahyari et al., 2017). Several text mining methods based on supervised and unsupervised machine learning methods have been developed. The figure below gives an overview of the major methods in text mining.

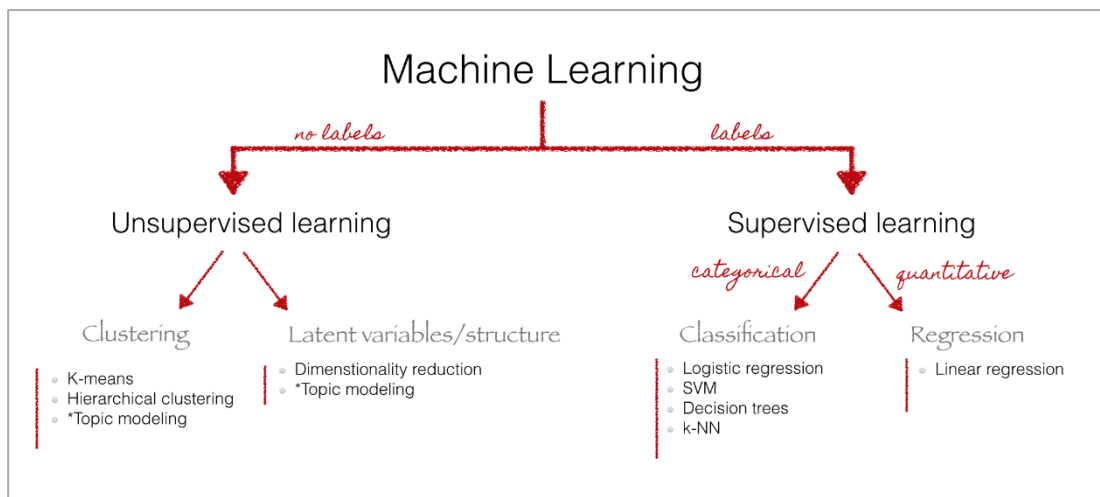


Figure 1: Overview of text mining methods (Source: <http://chdoig.github.io/acm-sigkdd-topic-modeling/#/>)

Supervised learning methods require labelled training data so as to learn from them and make predictions on the unseen data such as classifying documents into predefined categories. The major limitation in employing supervised learning is the availability of labelled training data that is specific to the purpose and domain of the work. Algorithms based on supervised learning that provide good results for a specific purpose and domain may not provide equally good results for other purposes or different domains. Unsupervised learning in text mining have edge over supervised learning as they do not require training samples. However, unsupervised learning method do not replace supervised learning as all text mining tasks cannot be done only by using unsupervised learning.

The supervised and unsupervised methods developed in text mining are oriented towards fulfilling the following major functions: text categorization; text clustering; concept mining; information retrieval and information extraction(Ghosh, Roy, & Bandyopadhyay, 2012). The major functions in text mining and the methods in accordance to (Allahyari et al., 2017) are presented in the table below.

**Table 1 : Text mining functions and methods**

<b>Sn.</b>	<b>Text mining function</b>	<b>Methods used</b>
1	Classification	<ul style="list-style-type: none"> <li>a. Naïve Bayes Classifier</li> <li>b. Nearest Neighbor Classifier</li> <li>c. Decision Tree Classifiers</li> <li>d. Support Vector Machines</li> </ul>
2	Clustering	<ul style="list-style-type: none"> <li>a. Hierarchical Clustering Algorithms</li> <li>b. K-means Clustering</li> <li>c. Probabilistic Clustering and Topic Models</li> </ul>
3	Information Extraction	<ul style="list-style-type: none"> <li>a. Named Entity Recognition (NER)</li> <li>b. Hidden Markov Models</li> <li>c. Conditional Random Fields</li> <li>d. Relation Extraction</li> </ul>

## 2.2 Discovering concepts in texts: Topics and Topic Modeling

In order to visualize a large collection of texts, it is of prime importance to know the different hidden concepts that the text contains without manually going through all of the texts. Clustering functions in text mining come very close to it. K-means clustering and Hierarchical clustering algorithms cluster the documents into different groups. However, there is no warranty that a single document contains only a single concept. This limitation is overcome by topic modelling. The outcome of topic modelling are a list of group of words that provide human with a sense of concept covered in the text. Each group of words is called a topic. The diagram presented below from the paper (D. Blei, Carin, & Dunson, 2010) gives a lucid idea of what a topic means in context of topic modelling.

<b>"Genetics"</b>	<b>"Evolution"</b>	<b>"Disease"</b>	<b>"Computers"</b>
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

**Figure 2 : Sample terms in topics**(Source: (D. Blei et al., 2010))

The top 15 words in each of the four topics are displayed in the above figure. The list of words gives insight to the contents in the collection of texts to a human reader. It is clear from the figure above that the contents of the texts from which the topics were generated covered at least four concepts, namely genetics, evolution, disease and computers.

The premise of topic modelling 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 methods have been proposed and developed for topic modelling. Some of the common methods used for topic modelling are listed in the table below.

**Table 2 : Common topic modelling methods**

<b>Sn.</b>	<b>Topic Modelling Method</b>
1	Latent Semantic Analysis (LSA)
2	Probabilistic Latent Semantic Analysis (PLSA)
3	Latent Dirichlet Allocation (LDA)
4	Correlated Topic Model (CTM)

Each of the topic modelling have their own methods, strengths and weaknesses going through each of which is beyond the scope of this study. There are several papers which discuss about these such as (Sharma & Sharma, 2017), (Alghamdi & Alfalqi, 2015)

In this study, Latent Dirichlet Allocation (LDA) was implemented for topic modelling LDA is improvement over LSA and it provides probabilistic topics which are interpretable rather than a semantic space. The probabilistic topics are key to geovisualization as the probabilities are treated as attribute values. Also, LDA could be modelled as finite number of topics which meant there would be finite number of visualizations. Given the rational above LDA was the optimal choice for topic modelling with respect to the needs of this study. LDA is discussed in further details in the section below.

### **2.2.1 Latent Dirichlet Allocation (LDA)**

LDA 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 models(D. M. Blei & Lafferty, 2009). 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). The assumption of LDA that each



document contains more than one topic is very rational as each document tends to be heterogeneous in nature covering many concepts. This is shown in the figure below which was included in the paper (D. Blei et al., 2010).

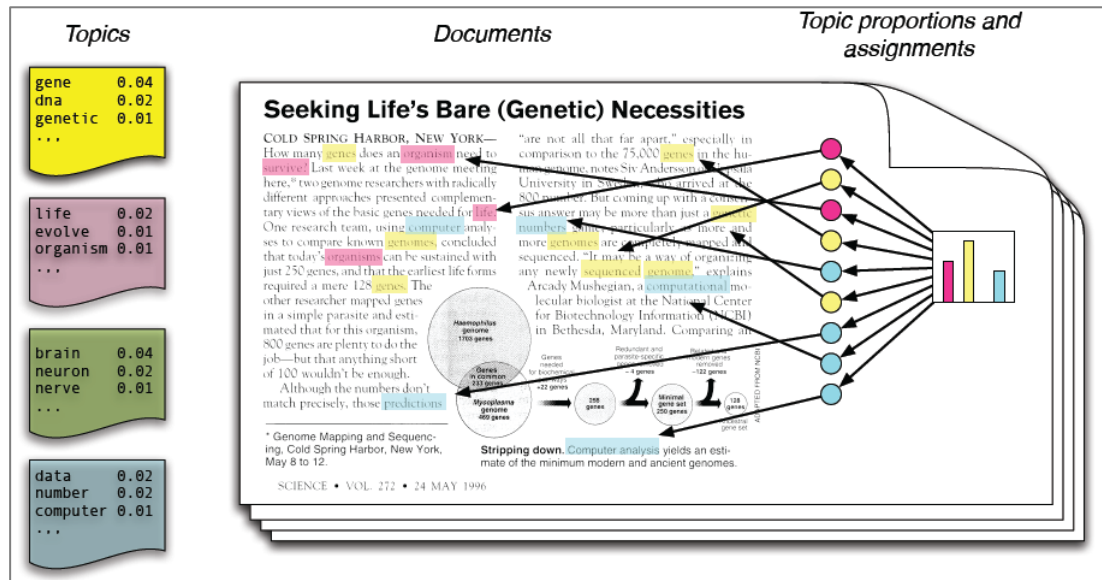


Figure 3: Intuitive diagram for Topic Modeling using LDA (D. Blei et al., 2010)

Although the figure above shows a single document, it is modelled as constituting of finite number of topics. The document is modelled into four topics as shown in the left side of the figure above. In a document, the topics have various proportions depending upon the content of the document.

The plate notation of LDA is shown in figure below.

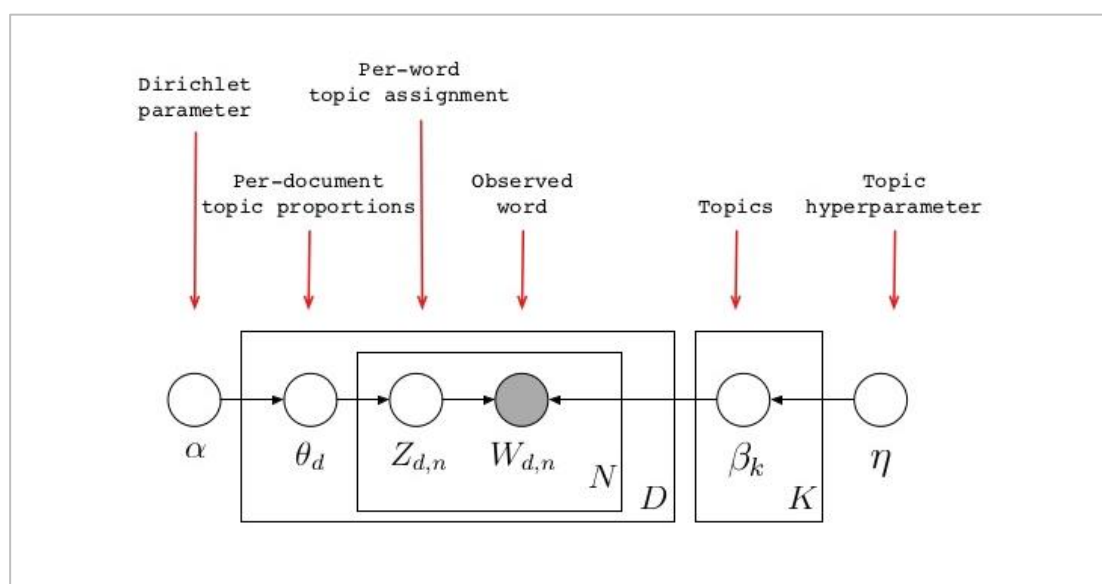


Figure 4: LDA Plate Notation (D. M. Blei & Blei, 2008)

### 2.2.2 Software Implementations for Topic Modelling

There exists a large number of tools that can perform topic modelling, particularly LDA. These are listed in the table below.

**Table 3: List of software for topic modelling**

Sn.	Tool name	Implementation language
1	Mallet	Java , Wrapper in R
2	Topic Models (Package)	R
3	LDA (Package)	R
4	Gensim	Python
5	LDA-C	C
6	GibbsLDA++	C and C++
7	Stanford Topic Modeling Toolbox	Java

In this study, we selected Gensim as our choice for implementing topic modelling. Gensim is open source python library for topic modelling. It has large user and developers' community. It supports topic modelling, document indexing and similarity retrieval. Gensim has implementations of Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP). The Gensim package implements a variation of LDA based upon the paper Online Learning for Latent Dirichlet Allocation (Hoffman, Blei, & Bach, 2010) which allows for handling of large amount of document collection including data that arrives in stream. The choice for Gensim in this study was due to the memory efficiency of the package. The package uses generators and iterators which are part of the Python for streamed data processing. The streamed data processing allows processing of large amount of text data even with lower processing capabilities.

## 2.3 Natural Language Processing (NLP)

Natural Language Processing (NLP) 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. NLP provides tools and techniques for text mining as well as topic modelling. In this study, NLP is used for preprocessing the textual data before they are feed into topic modelling algorithms.

A list of common NLP open source software are presented in the table below:

**Table 4: List of popular NLP software**

Sn.	Software/Tool/Package	Implementation Language
1	Natural Language Toolkit	Python
2	Spacy	Python
3	Stanford Core NLP	Java, Python Wrappers
4	Apache OpenNLP	Java

In this study, Spacy was selected as NLP tool which defines itself as an industrial strength NLP. It is open source and has a large user and developer's community. It comes with pre-trained statistical models and supports multiple languages. One of the reason for selecting Spacy was that it is implemented in Python which made integrating it to topic modelling much easier.

## 2.4 Extracting Location Information from Text

While topics extracted using topic modelling provides theme for visualization, location information from texts are essential for georeferencing the topics for geovisualization. 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 field is Named Entity Recognition (NER). NER identifies words that denote person, organization, location, object, etc. Different NER implementations have different classes of entities(Atdağ & Labatut, 2013). While NER

identifies several classes of entities, only entities having location information are of interest for the purpose of georeferencing the topics. 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 and unambiguous georeferencing of location information in a text which is called geoparsing (Goodchild & Hill, 2008). Some of the popular geoparsing tools and services are listed in the table below.

**Table 5: List of popular geoparsing applications**

Sn	Software/Tool/Service	Implementation
1	Clavin	Java
2	Mordecai	Python
3	Geoparse.io RESTful web API	Python API

In this study, Geoparse.io is used for geoparsing the texts. 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. Although it is not free, it allows 1000 API calls for free per month. Geoparse.io was selected for this study as it had small learning curve and had API in python.

### **3. Relevant Works**

#### **3.1 Relevant works in text mining from newspaper articles**

There has been several domain specific studies that have utilized topic modelling to discover concepts from collection of texts as topics. Text mining from newspaper is one of the most research active area. Newspaper are the source of unedited and unmodified version of history which arouses interest to researchers who want to get insight to the history(Cheney, 2013). The digitization of historical newspapers by libraries has also opened the opportunity for research in this field. The availability of such a data that spans decades if not centuries provides opportunities to study the changes in human history. (Torget, Mihalcea, Christensen, & McGhee, 2010) used sample of around 230,000 pages of historical newspapers analysing the quantity and quality of the digitized content along with measurement of language pattern. (Godbole & Srinivasaiah, 2007) analysed sentiment from news and blogs. (Akhter, 2015) extracted information related to road accidents and visualized them interactively.

#### **3.2 Relevant works in topic modelling**

Studies that have considered both topic modelling and location component have primarily intended to improve topic modelling 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. (Pölitiz, 2015) used spatial locations in newspapers and social media for explore topics in those regions. (Yin, Cao, Han, Zhai, & Huang, 2011) also used documents which were embedded with GPS coordinates for topic modelling to find topics that are coherent in a particular geographic region.

#### **3.3 Relevant works in visualizing texts**

A large number of studies exists in text visualizations. Also hundreds of text visualization techniques have been developed. (Cao & Cui, 2016) reviewed more than 200 papers based on text visualization techniques accumulated in Text Visualization Browser (<http://textvis.lnu.se/>). The paper identified five categories of text visualizations which are listed below.

- a. Visualization of document similarity.
- b. Visualization for revealing content of the document

- c. Visualization of sentiments and emotions in the text
- d. Visualization of the corpus
- e. Visualization of domain-specific rich-text corpus









These techniques cover 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. The emphasis is on visualizing “What is the text about?” The component of “Where is the text about?” remains unanswered. There is a compelling motivation for development of text visualization technique that combines both the textual and spatial components of text data and provide spatial insight from large collection of texts.

## 4. Data

In this section, a brief summary of the news data used in this study is provided. An overview of the collection and filtering process is also discussed.

News from different online news sources were downloaded that belonged to the year 2017. All of the news were downloaded based upon the following seven keywords (category) in google. The keywords and the number of news articles are presented in the table below.

Table 6: Categories and number of news

Sn.	Keyword (Category)	No. of articles	Visual representation of number of articles
1	World Cup	60	
2	Wildlife Poaching	35	
3	Tornado	84	
4	Nuclear War	46	
5	Election	98	
6	Ebola	91	
7	Deforestation	98	
	<b>Total</b>	<b>512</b>	

The articles were downloaded by automatizing the process by writing scripts in python. The scripts are included in the appendix section. Newspaper3k python package was used for downloading the articles. The package can be accessed on GitHub at

<https://github.com/codelucas/newspaper>. The articles were sampled to check if they had any anomaly. Significant number of files less than 3Kb contained artefacts or were empty. These articles were filtered. Also, files larger than 8Kb were filtered as geoparsing API header had limitation of 8Kb in size. The above table shows the numbers of articles that persisted after filtered. Samples of the downloaded news articles are included in the appendix. It is to be noted that although the news articles are downloaded based on several categories, these are feed into topic modelling algorithms without any annotations. The categories are used only for making comparisons between news collection, topics and geovisualizations.



## 5. Methodology

This section elaborates on the process that initiates from collection of the online news articles to geovisualization of the articles in the form of interactive maps. The theoretical background of the methodologies applied here are discussed in chapter 2: theoretical review. A chart of overall methodology is presented in the figure below.

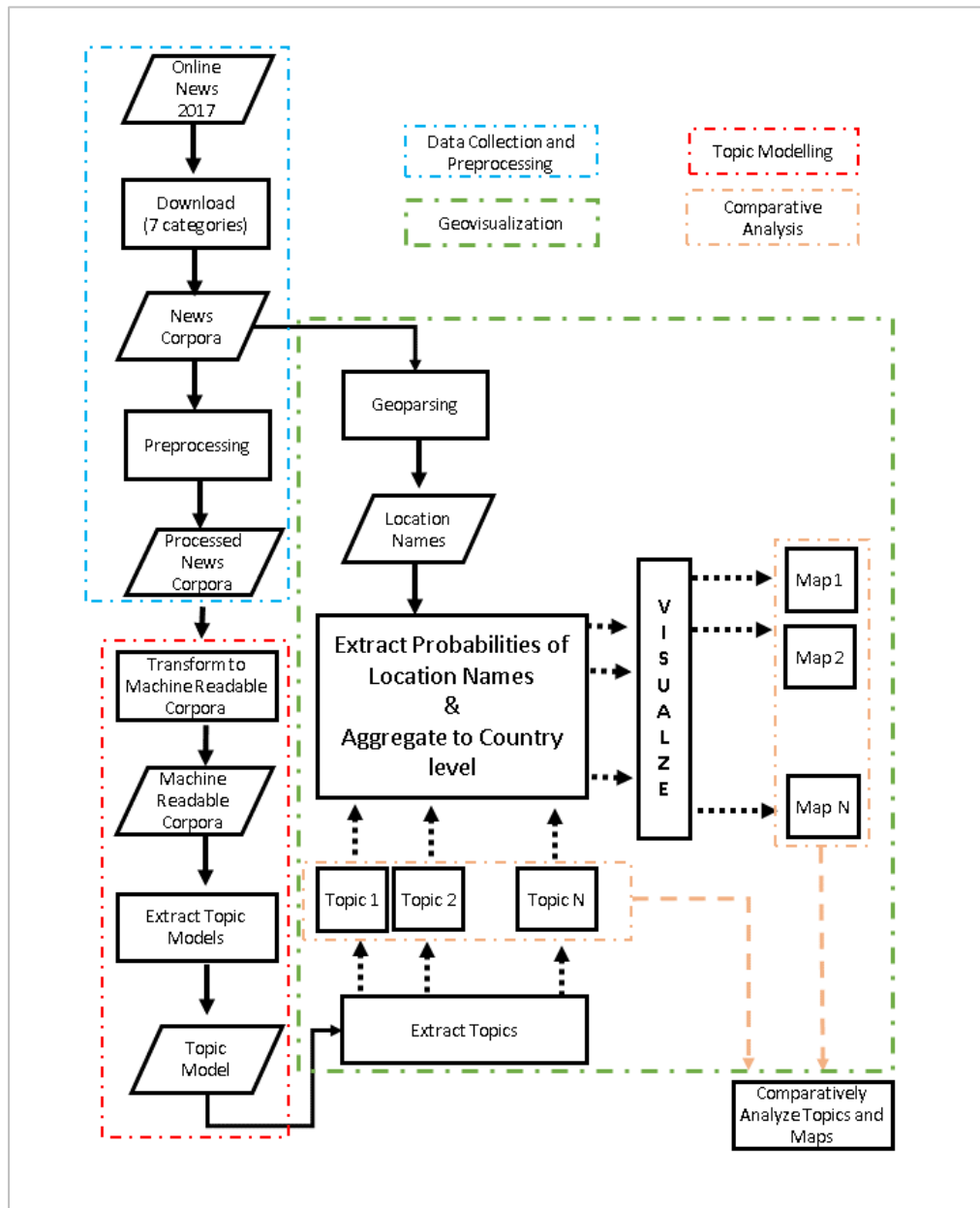


Figure 5: Chart of Overall Methodology

## 5.1 Building Corpus of News Articles

This procedure of collection and filtration of news articles is explained in Chapter 4: Data. Once done through this process, we already have a corpus suitable for our study. In our study, we perform topic modelling using Latent Dirichlet Allocation (LDA) algorithm which is an unsupervised learning algorithm. Hence, the news articles are not annotated based upon their category. The necessity of building our own corpus is well expressed in (Xiao, 2010) which states that although there exists thousands of corpus, these are created for certain research purposes and most are not available publicly. It is one of the most time consuming and costly process in text mining.

## 5.2 Preprocessing Corpus of News Articles

Before any textual data is fed into a topic modelling algorithm, it must be preprocessed in order to get rational results. This is fairly expressed in the saying, “Garbage in, Garbage out”. In fact, it is one of the most time consuming process after the data has been collected. The quality of the results is highly sensitive to the preprocessing steps. Preprocessing of the corpus involves several Natural Language Processing (NLP) tasks. In this study, Spacy was used for preprocessing. The rationale for choosing Spacy is explained in Section 2.2.2: Software Implementation. The following preprocessing were applied to the corpus of the news articles.

### a. Tokenization

Tokenization is the process of segmenting a document into individual word, punctuations, symbols, etc. Tokenization is based upon series of tokenization rules. These rules dictate when to split the words and when to preserve them.

Example document:

*“One of the measures introduced by lawmakers would remove conservation protection from 1.2 million hectares of the Amazon forest, an area larger than Jamaica. U.K. has made expressed its resentment to this decision.”*

Tokenized document:

*[One, of, the, measures, introduced, by, lawmakers, would, remove, conservation, protection, from, 1.2, million, hectares, of, the, Amazon, forest, , an, area, larger, than, Jamaica, ., U.K., has, made, expressed, its, resentment, to, his, decision, .]*

It can be observed that words such as U.K and number 1.2 are preserved based upon the tokenization rules.

b. Text Lemmatization

Lemmatization converts the words into its canonical or citation form (Bird, Klein, & Loper, 2009).

The lemmatized form of sentence mentioned in above example is presented below.

Lemmatized document:

*['one', 'of', 'the', 'measure', 'introduce', 'by', 'lawmaker', 'would', 'remove', 'conservation', 'protection', 'from', '1.2', 'million', 'hectare', 'of', 'the', 'amazon', 'forest', ',', 'an', 'area', 'large', 'than', 'jamaica', '.', 'u.k.', 'have', 'make', 'express', '-PRON-', 'resentment', 'to', 'this', 'decision', '.']*

These words are lemmatized because words such as “*measures*”, “*Measure*”, “*measured*” have same meaning in context of topic modelling and hence are transformed to its canonical form “*measure*”. Spacy lemmatizes the pronoun such as “*its*” to “*-PRON-*”.

c. Removal of stop-words

Stop-words are such words in a text which do not add any value to the purpose of the study. Which words are to be assumed stop-words depend upon the purpose of the study. Each NLP 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. In fact, 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. The following example shows the document after removal of stop-words.

Document after removal of stop-words from the example document in (a :

*['one', 'measure', 'introduce', 'lawmaker', 'remove', 'conservation', 'protection', '1.2', 'million', 'hectare', 'amazon', 'forest', ',', 'area', 'large', 'jamaica', '.', 'u.k.', 'express', 'resentment', 'decision', '.']*

The words 'of', 'the', 'by', 'would', 'from', 'an', 'than', 'have', 'make', '-PRON-', 'to' and 'this' were removed from the above document as these words add no value to topic modelling in our study.

d. Remove numbers, punctuation marks, symbols

Numbers, punctuation marks and symbols are removed from the text as these do not add any value to the topic modelling but instead act as noise.

Document after removal of numbers, punctuation marks and symbols:

*[one, measure, introduce, lawmaker, remove, conservation, protection, hectare, amazon, forest, area, large, jamaica, express, resentment, decision]*

e. Detect bigrams and trigrams

A combination of words such as nuclear weapon, 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.

Examples of bigrams and trigrams detected from the study corpus:

*[national\_team, alexis\_sanchez, semi\_final, fifa\_world\_cup, sri\_lanka, saudi\_arabia, emergency\_management\_society]*

### **5.3 Building Machine Readable Corpus and Dictionary**

Computers are inefficient in processing text data. In text mining, it is common to process thousands of documents. Hence, it is necessary to convert the text corpus into a format which facilitates faster processing. The words are converted into vectors for rapid processing since computers are much more efficient in handling numbers than strings. There are several algorithms for converting words to vectors. Some of the methods consider the grammar and the word order. In our study, bag-of-words (BOW) model was applied to convert words 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 documents is given a unique integer id. This mapping of word into unique id is called dictionary. The dictionary is used to map the unique ids back to tokens after the topic models are generated. Each word in a particular document is represented by a 2 dimensional vector. The first element of the vector is the unique integer id and the second element is the frequency of the word in that document. In this way, a document is a collection of 2 dimensional vectors. In Gensim, this collection is represented a list. Also all of the documents are also represented as lists. Hence, the machine readable corpus implementation in Gensim is a list of lists.

In our study, there are 17446 unique words after preprocessing the corpora of news articles. The dictionary is shown in the table below.

**Table 7: Dictionary of Corpus built on BOW model**

Sn	Word (token) : Unique Token Id
1	'norway' : 0
2	'reexamine' : 1
3	'financial' : 2
4	'commitment' : 3
.	.
.	.
.	.
.	.
.	.
17445	'susceptibility' : 17444
17446	'presser' : 17445

Our study data has 512 documents. So, the corpus is a list of 512 lists. A sample of a document and its representation in corpus is shown in the table below.

**Table 8: Document and its representation in form of bag-of-words**

Document	Corpus in form of word vectors
<i>Norway reexamined its financial commitment to the Amazon Fund as a result of Brazil's ever-weakening environmental protection policies. On Friday, Brazilian President Michel Temer met with Norwegian Prime Minister Erna Solberg in Oslo to promote investment in the South American country.</i>	[(0, 2.0), (1, 2.0), (2, 2.0), (3, 2.0), (4, 6.0), (5, 5.0), (6, 2.0), (7, 9.0), (8, 2.0), (9, 6.0), (10, 6.0), (11, 3.0), (12, 1.0), (13, 2.0), (14, 3.0), (15, 1.0), (16, 4.0), (17, 2.0), (18, 3.0), (19, 1.0), (20, 1.0), (21, 3.0), (22, 1.0), (23, 1.0), (24, 1.0), (73, 1.0), (74, 1.0), (75, 1.0), (76, 1.0), 1.0), (126, 1.0), (127, 1.0), (128, 1.0), (129, 1.0), (130, 1.0),
.	.
.	.
.	.
.	.
.	.

<i>Other proposed measures include relaxing the environmental licensing rules for big infrastructure projects, opening sales of farmland to foreigners and loosening rules for approving new mining projects. They are expected to be passed by Brazil's Congress in coming months.</i>	(161, 1.0), (162, 1.0), (163, 1.0), (164, 1.0), (165, 2.0), (166, 1.0), (167, 1.0), (168, 2.0), (169, 1.0), (170, 1.0), (171, 1.0), (172, 1.0), (173, 1.0), (174, 1.0), (175, 1.0), (176, 1.0), (177, 1.0), (178, 1.0), (179, 1.0), (180, 1.0)]
---	---

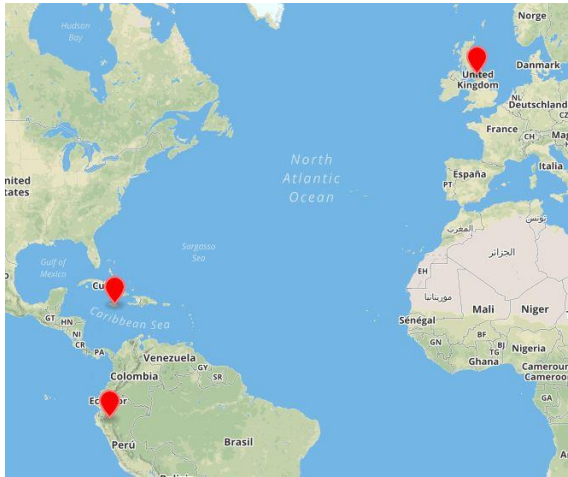
## 5.4 Extraction of Topic Model

Topic models are generated from the machine readable corpus (list of list of word vectors). In this study, the topic model was generated using Latent Dirichlet Allocation (LDA). The rationale behind selecting LDA for topic modelling in this study is discussed in section 2.2. The functioning of LDA is explained in section 2.2.1. The implementation was done using Gensim, a python package for topic modelling. The scripts written for the process is included in the appendix. The only hyperparameter in implementation of LDA is the number of topics. 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. Though the mathematical models such as perplexity and topic coherence provide some information, these measures are no sure shot indication of good topics (Chang, Gerrish, Wang, & Blei, 2009). In this study, the topics were evaluated based on human interpretation that it fits the purpose of the study. Also, pyLDAvis, a python library for visualizing LDA topic models was used to visually evaluate the topic models. This library can be accessed on GitHub at <https://github.com/bmabey/pyLDAvis>. The number of iterations was also fine-tuned based on trial and error. The topics generated with different model parameters are presented in the results and discussion section.

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

The location information is critical in Geovisualizing the topics as it is used to georeference the topics. The location information is extracted by geoparsing the texts. In this study, Geoparser.io is used for geoparsing the texts. The rationale behind using Geoparser.io is discussed in section 2.4. 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. In this study, only the name and coordinate of the location is utilized. A sample geovisualization of the geoparsed text is shown in the figure below.

Table 9: Text and visualization of geoparsed text

Text	Geovisualization of the geoparsed text
<p><i>“One of the measures introduced by lawmakers would remove conservation protection from 1.2 million hectares of the Amazon forest, an area larger than Jamaica. U.K. has made expressed its resentment to this decision.”</i></p>	

## 5.6 Geovisualization

The mapping of topics is analogous to mapping of a thematic layer in cartography. The thematic layer is first georeferenced according to the location it covers. Then it is symbolized according to one of its attribute value. For example, a population map is georeferenced based upon the administrative boundaries. Then the population of each administrative unit is used to symbolize the map. Similar approach is followed in



mapping the topics. The mapping is done at the level of country. The locations for georeferencing the topics are obtained by geoparsing all the texts in the corpus. The attribute value of each location is different for each topic. The attribute value at a location (L) for a topic (T) is the (coefficient) probability of the location L in the topic T that is obtained after as a result of LDA. As, the mapping is done at country level, the values of locations are aggregated based upon where this values lie. A web based interactive visualization was built using Leaflet visualization library. Choropleth maps were prepared for geovisualizing the topics. These maps are included in the result section. Also, qualitative analysis of the maps are done to see if they fit the concepts expressed in the collection of newspaper articles.

## 6. Results and Discussion

In this chapter, we discuss the results of the topic modelling in Section 6.1. The results of topic modelling in relation to LDA parameters are presented the section. In section 6.2, the geovisualization of topics are presented. Also a qualitative examination is done to inspect if the geovisualizations were capable of answering the “Where is the text about?” question.

### 6.1 Results of topics generated from LDA

Each topic generated using topic modelling consists of probabilistic distribution of word. If the words with the highest probabilities (top words) are coherent and indicates a common concept, only then the topics can reveal the concepts from the collection of texts. In this study, different parameters were set for number of topics and number of iterations in LDA to obtain topics. The results are presented in this section. It is to be noted that LDA is a generative model. Hence, each time the model is run, the terms in the topics slightly vary. Each terms the topic also have a probability coefficient. However, for the purpose of clarity, these are not included in the tables below. Also, topics having less than 4% of total tokens are not listed in the table.

A. Topics generated with 7 topics and 50 iterations

Table 10: Top 5 topic terms with 7 topics and 50 iterations

Topic Number	Top 5 words in topics	Percentage of tokens
1	Forest, company, year, tornado, time	25.2
2	Government, country, election, people, time	18.1
3	Election, year, people, vote, country	15.2
4	Country, world_cup, people, team, election	12.5
5	Time, people, election, report, home	11
6	Country, time, group, year, company	9.2

7	Country, election, government, work, party	8.9
---	--	-----

The topics generated with the aforementioned LDA parameters yield mixed results. Top words in topic numbers 2, 3 and 6 are close to election but other topics have either words that could be assigned to different categories. We conclude that it a poor topic model.

#### B. Topics generated with 7 topics and 400 iterations

**Table 11: : Top 5 topic terms with 7 topics and 400 iterations**

<b>Topic Number</b>	<b>Top words in topics</b>	<b>Percentage of tokens</b>
1	Forest, deforestation, company, country, government	23.6
2	Ebola, <u>north_korea</u> , virus, country, people	21
3	Tornado, damage, storm, county, home	13.6
4	Party, election, vote, win, year	11.7
5	World_cup, team, group, home, <u>tornado</u>	11.7
6	Elephant, year, wildlife, country, animal	10
7	Election, vote, result, president, people	8.4

The topics generated after increasing the number of iterations have significant improvement. Most topics have terms that indicate towards a common concept and each topic could be assigned to one of the seven categories of our data. However, topic number 2 and 5 still have mixed terms which are incoherent to the rest of the terms. These are highlighted in the table above.

Similarly topics were generated by increasing the number of iterations to 600 and then 800. However, similar pattern emerged with some topics containing mixed terms.

Instead, attempts were made to get good topics by increasing the number of topics. The results are shown in the tables presented below.

#### C. Topics generated with 21 topics and 50 iterations

**Table 12: : Top 5 topic terms with 21 topics and 50 iterations**

<b>Topic Number</b>	<b>Top words in topics</b>	<b>Percentage of tokens</b>
1	Forest, deforestation, company, palm_oil, year	10.4
2	Tornado, storm, damage, county, home	10.3
3	Ebola, outbreak, disease, people, country	9.7
4	Deforestation, forest, company, country, policy	8.1
5	Election, vote, president, kenyatta, party	7.7
6	North_korea, nuclear, missile, trump, military	6.7
7	World_cup, play, tournament, time, team	5.2
8	Group, team, world_cup, england, win	4.5
9	Party, election, merkel, vote, coalition	4.1
10	Damage, tornado, storm, air, home	4.1

Visual analysis of the topics shows that the top words in the topics are coherent and indicates a common concept.

Keeping the number of topics constant, the number of iterations were further increased to see the changes in the topics. The results are presented in the table below.

#### C. Topics generated with 21 topics and 400 iterations

Table 13: : Top 5 topic terms with 21 topics and 400 iterations

Topic Number	Top words in topics	Percentage of tokens
1	Forest, deforestation, company, land, palm_oil	15.3
2	Damage, tornado, storm, county, home	8.2
3	Virus, ebola, outbreak, disease, vaccine	7.9
4	Country, north_korea, nuclear, work, <u>ebola</u>	7.8
5	Election, vote, result, kenyatta, kenya	7.2
6	World_cup, team, group, play, win	6.7
7	Party, election, vote, poll, year	6.6
8	Ebola, sierra_leone, work, people, hospital	4.3
9	Tornado, storm, <u>ivory</u> , area, county	4

The topic modelling with 400 iterations has coherence topic terms for most topics except for topic number 4 and 9. The incoherent terms are underlined in the table above. The number of iterations were increased to examine if there are more coherent topics with higher number of iterations.

#### D. Topics generated with 21 topics and 600 iterations

Table 14: : Top 5 topic terms with 21 topics and 600 iterations

Topic Number	Top words in topics	Percentage of tokens
1	Forest, deforestation, company, land, area	14.8

2	Tornado, damage, storm, county, home	12.1
3	Ebola, virus , outbreak, disease, health	8
4	World_cup, team, group, play, win	7.5
5	Election, vote, president, kenyatta, result	7.2
6	North_korea, nuclear, trump, missile, kim	6.8
7	Elephant, ivory, wildlife, trade, ban	5.5
8	Party, election, vote, merkel, campaign	5.2
9	Ebola, work, people, time, die	4.8

All of the topic terms in the above table are coherent and indicate a common concept. Topic modelling was also done with 800 iterations. The results were very similar to what was achieved with 600 iterations. The results are included in the appendix. Since, the topics extracted at 600 iterations with 21 topics were coherent, this was used to create geovisualization which is discussed in the following section.

## 6.2 Geovisualization results

Geovisualization of coherent topics (as obtained in section 6.1) was done using Leaflet Geovisualization library as explained in section 5.6. In this section, the topics and the geovisualized choropleth maps are presented in tandem. In the following section, we discuss if the tandem of topics and maps reveal the contents of the news collection and whether the maps add any spatial insight to the text collection.

i. Choropleth Map of Topic 1

Top Topic Terms:

$(0.022 * \text{"tornado"} + 0.016 * \text{"damage"} + 0.015 * \text{"storm"} + 0.012 * \text{"county"} + 0.009 * \text{"home"} + 0.008 * \text{"area"} + 0.006 * \text{"people"} + 0.006 * \text{"report"} + 0.005 * \text{"tree"} + 0.005 * \text{"near"} + \dots)$



Figure 6 : Choropleth Map of Topic 1

ii. Choropleth Map of Topic 2

Top topic terms:

$(0.010 * \text{"elephant"} + 0.008 * \text{"ivory"} + 0.007 * \text{"wildlife"} + 0.005 * \text{"trade"} + 0.005 * \text{"ban"} + 0.005 * \text{"poacher"} + 0.005 * \text{"poach"} + 0.005 * \text{"country"} + 0.005 * \text{"south\_africa"} + 0.004 * \text{"work"} + \dots)$

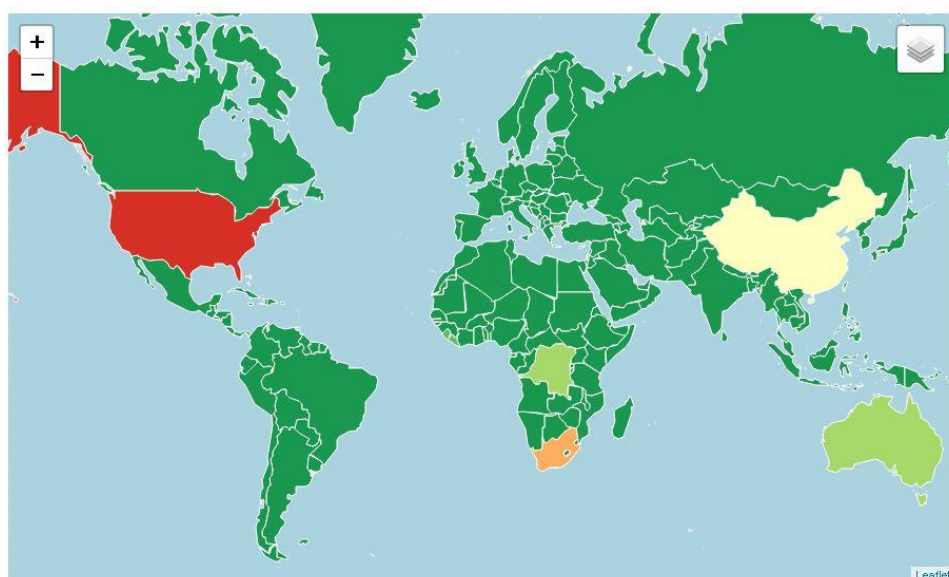


Figure 7: Choropleth Map of Topic 2

iii. Choropleth Map of Topic 3

Top topic terms:

$(0.022 \cdot \text{"forest"} + 0.020 \cdot \text{"deforestation"} + 0.014 \cdot \text{"company"} + 0.006 \cdot \text{"land"} + 0.006 \cdot \text{"area"} + 0.006 \cdot \text{"palm\_oil"} + 0.006 \cdot \text{"year"} + 0.006 \cdot \text{"policy"} + 0.005 \cdot \text{"global"} + 0.005 \cdot \text{"report"} + \dots)$



Figure 8: Choropleth Map of Topic 3

iv. Choropleth Map of Topic 4

Top Topic Terms

$(0.018 \cdot \text{"election"} + 0.010 \cdot \text{"vote"} + 0.006 \cdot \text{"president"} + 0.006 \cdot \text{"kenyatta"} + 0.006 \cdot \text{"result"} + 0.006 \cdot \text{"kenya"} + 0.005 \cdot \text{"odinga"} + 0.005 \cdot \text{"opposition"} + 0.005 \cdot \text{"country"} + 0.004 \cdot \text{"people"} + \dots)$

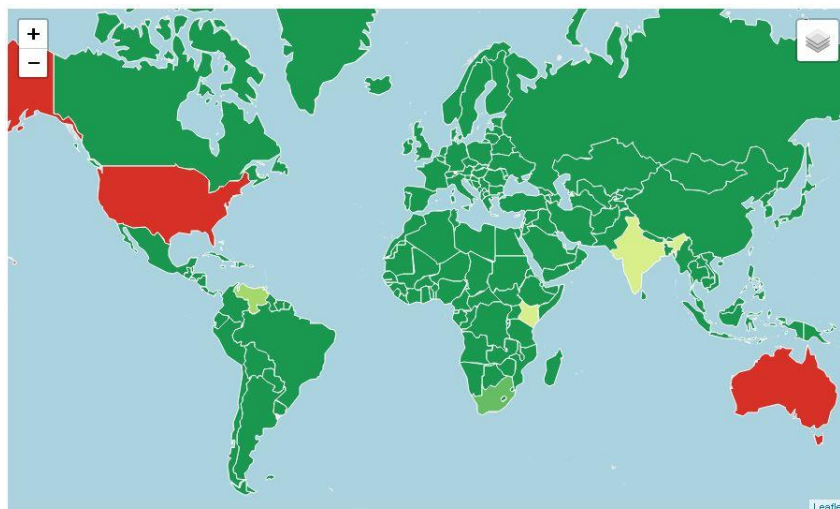


Figure 9: Choropleth Map of Topic 4



v. Choropleth Map of Topic 5

Top topic terms:

$(0.008 * \text{"charcoal"} + 0.005 * \text{"people"} + 0.004 * \text{"country"} + 0.004 * \text{"report"} + 0.004 * \text{"year"} + 0.004 * \text{"new"} + 0.003 * \text{"come"} + 0.003 * \text{"area"} + 0.003 * \text{"tree"} + 0.003 * \text{"work"} + \dots\dots\dots)$

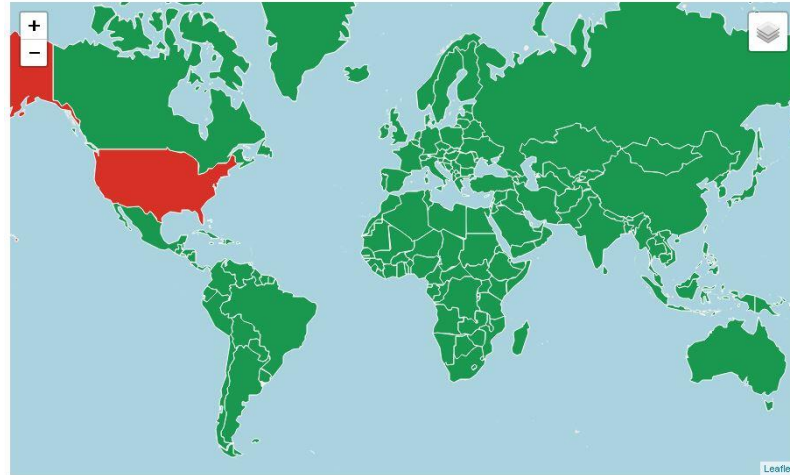


Figure 10: Choropleth Map of Topic 5

vi. Choropleth Map of Topic 6

Top topic terms:

$(0.015 * \text{"ebola"} + 0.014 * \text{"virus"} + 0.013 * \text{"outbreak"} + 0.012 * \text{"disease"} + 0.007 * \text{"health"} + 0.007 * \text{"people"} + 0.006 * \text{"country"} + 0.006 * \text{"vaccine"} + 0.005 * \text{"time"} + 0.005 * \text{"study"} + \dots\dots\dots)$

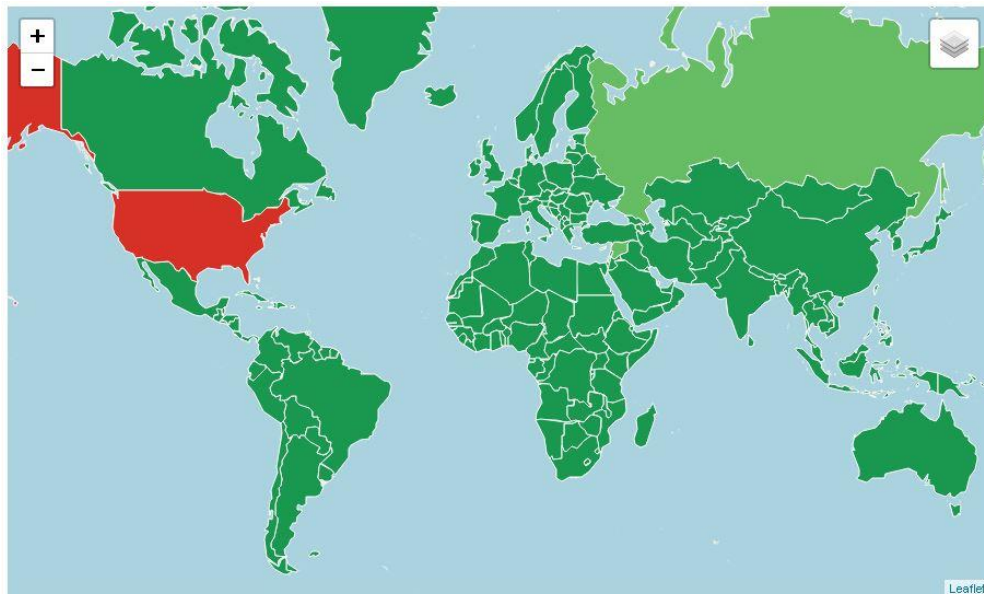


Figure 11: Choropleth Map of Topic 6

vii. Choropleth Map of Topic 7

Top topic terms:

$(0.015 * \text{"ebola"} + 0.014 * \text{"virus"} + 0.013 * \text{"outbreak"} + 0.012 * \text{"disease"} + 0.007 * \text{"health"} + 0.007 * \text{"people"} + 0.006 * \text{"country"} + 0.006 * \text{"vaccine"} + 0.005 * \text{"time"} + 0.005 * \text{"study"} + \dots)$

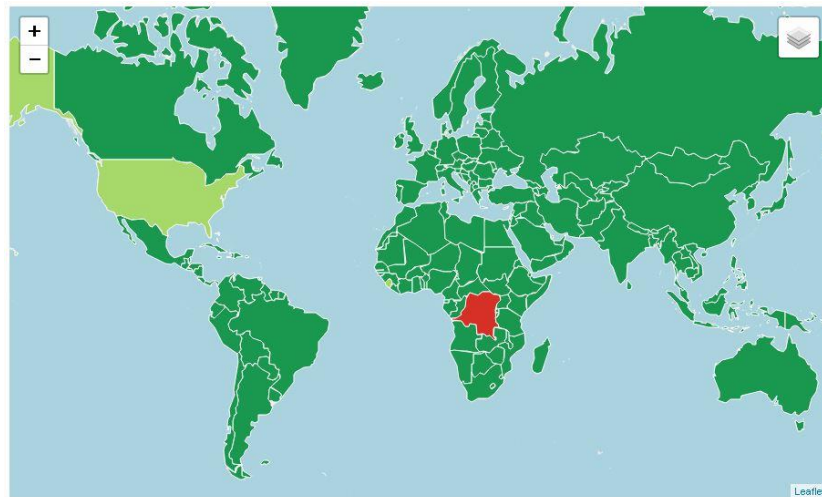


Figure 12: Choropleth Map of Topic 7

viii. Choropleth Map of Topic 8

Top topic terms:

$(0.004 * \text{"year"} + 0.004 * \text{"fema"} + 0.003 * \text{"amy"} + 0.003 * \text{"work"} + 0.003 * \text{"know"} + 0.003 * \text{"people"} + 0.003 * \text{"leave"} + 0.003 * \text{"include"} + 0.003 * \text{"time"} + 0.003 * \text{"key"} + \dots)$

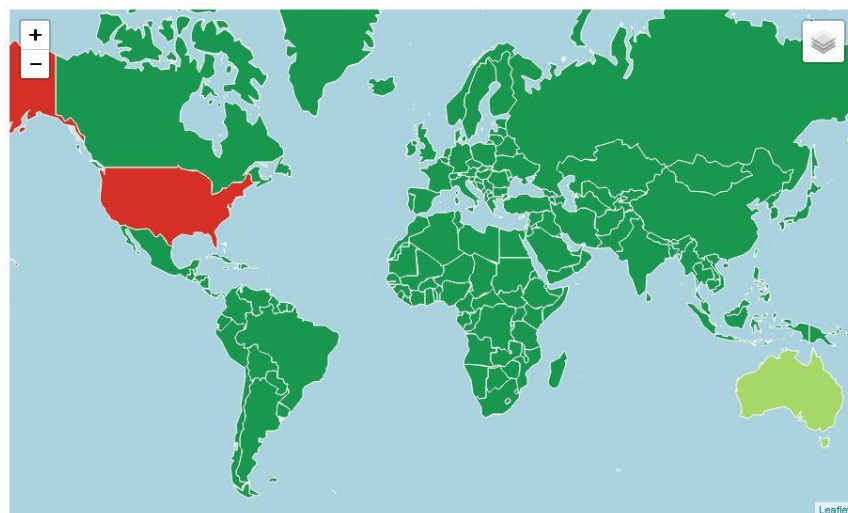


Figure 13: Choropleth Map of Topic 8

ix. Choropleth Map of Topic 9

Top topic terms:

$(0.020 * \text{"north\_korea"} + 0.011 * \text{"nuclear"} + 0.008 * \text{"trump"} + 0.008 * \text{"missile"} + 0.006 * \text{"kim"} + 0.006 * \text{"china"} + 0.006 * \text{"military"} + 0.005 * \text{"war"} + 0.005 * \text{"north\_korean"} + 0.004 * \text{"nuclear\_weapon"} + \dots)$

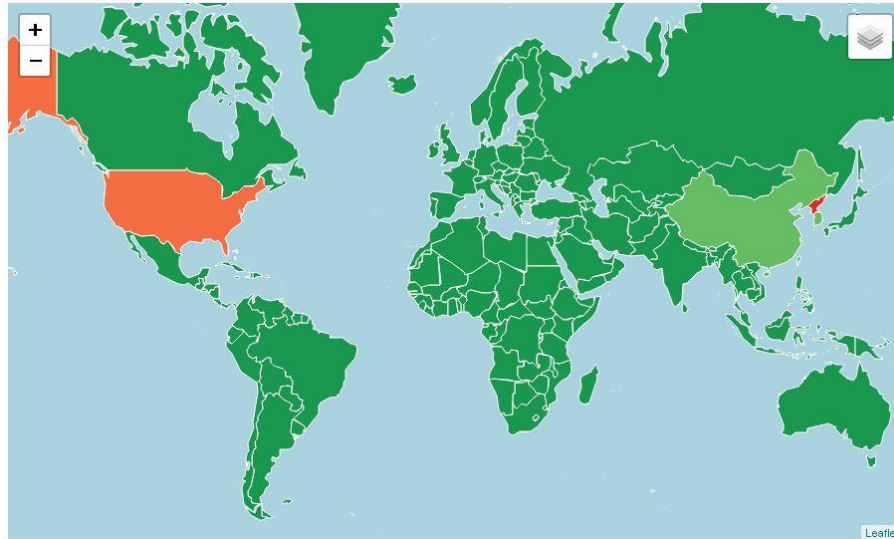


Figure 14: Choropleth Map of Topic 9

x. Choropleth Map of Topic 10

Top topic terms:

$(0.009 * \text{"kit"} + 0.008 * \text{"world\_cup"} + 0.007 * \text{"design"} + 0.006 * \text{"wear"} + 0.005 * \text{"time"} + 0.005 * \text{"year"} + 0.004 * \text{"event"} + 0.004 * \text{"russia"} + 0.004 * \text{"country"} + 0.004 * \text{"like"} + \dots)$

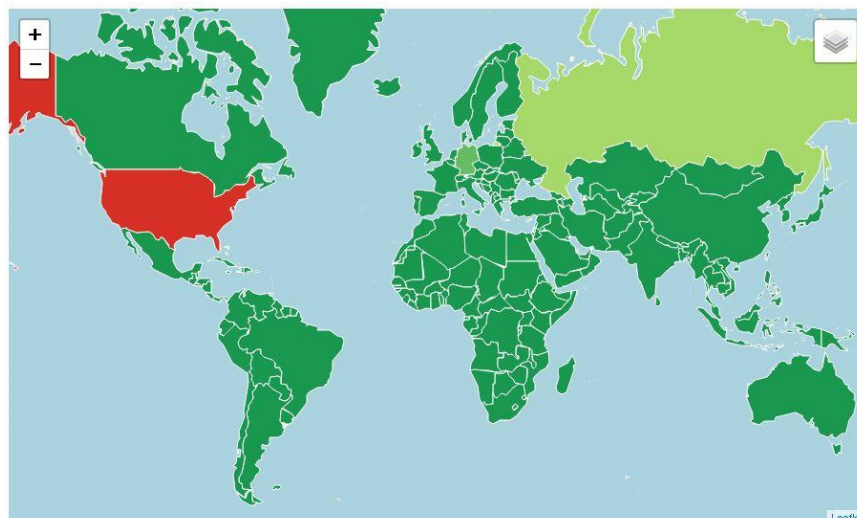


Figure 15: Choropleth Map of Topic 10

### 6.3 Discussion

In this study, news articles were collected which were related to seven different categories. It is to be noted that these articles were not annotated so as to identify its category. The corpus of news articles were preprocessed and topics were extracted using LDA. Each of the topics were geovisualized as choropleth maps. The comparison between the news collection and the tandem of topics and texts show that the tandem is able to both discover topics as well as provide spatial insight into the topics. In other words, the tandem does not only provide the answer to “What is the news collection about?” but also answers “Where in the news about?” To make the comparison more conspicuous, the news categories, topics discovered, countries highlighted in the maps and the reason for these countries to be highly covered in the news in 2017 are presented in the table below.

**Table 15: Comparison between news categories, discovered topics and countries highlighted in the maps**

<b>Categories of news articles</b>	<b>Topics matching the category with top 3 words</b>	<b>Countries highlighted in Map</b>	<b>Reason for the countries to be highly covered in the news in 2017</b>
World Cup	Topic 10 : kit , world cup, design	US, Russia	US failed to qualify for World Cup. Russia hosts World Cup in 2018.
Wildlife Poaching	Topic 2: Elephant, ivory, wildlife	US, South Africa, China	US considered removing ban on importing elephant trophies. 1028 rhinos killed in 2017 in South Africa. China is largest market of ivory.
Tornado	Topic 1 : Tornado, damage, storm	US, UK	One of the most active start to tornado season in US since 2008. UK has more tornadoes r than any other country relative to its size.
Nuclear War	Topic 9: North_korea, nuclear, trump	North Korea, US, China	North Korea, US. and China have a long history of adversary over North Korea seeking to achieve Nuclear capability.

Election	Topic 4: Election, vote, president Topic 6: Moore, republican, president	US, Australia, Kenya	Debates continue over hacks in 2016 US presidential election. State elections conducted in Western Australia. General elections was conducted in Kenya which was later annulled.
Ebola	Topic 7: Ebola, virus, outbreak	Congo, US	Ebola outbreak was contained in Congo. Organizations such as UNO, Red Cross and research organizations are based in US.
Deforestation	Topic 3: Forest, deforestation, company	US, Indonesia, Canada	Large chunks of forests cut down for growing oil palm.

The comparisons from the table strongly suggest that geovisualization of corresponding topics is able to provide spatial context to the topics. Each countries highlighted in the map are the countries have the highest influence corresponding to the topic. For example, map of topic number 7 (Ebola, virus, outbreak) highlights Congo and US. Congo had outbreak of Ebola but was successful in containing it. Organizations in US such as UNO, Red Cross, IBM, etc. are helping to contain the outbreak through direct assistance as well as research.

## 7. CONCLUSIONS

The existing massive amount of textual data along with the surge in rate of generation of textual data calls for automatic methods for information discovery. Currently available techniques in topic models are able to extract the hidden concepts from these massive text collections. However, these models do not provide spatial insight. Significant amount of textual data are geographic in nature and hence geovisualizing text helps in better understanding of the text. Also, information demonstrated in a form of map is easier and faster to grasp than text information.

In this study, topics were extracted from collections of news from 2017 using topic modelling. The news collection belonged to seven different categories. The extracted topics revealed the information from the collection of news in terms of collection of words. Choropleth maps were prepared using probability of spatial terms in LDA topics. These maps added spatial context to the topics. The comparison between the maps and the news collection showed that the maps were in line with the contents of news collection and were able to provide spatial insight to the corpora of news collection.

## 8. REFERENCES

- Akhter, H. (2015). Information Extraction and Interactive Visualization of Road Accident Related News, *I28(5)*, 37–40.
- Alghamdi, R., & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, 6(1), 147–153. <https://doi.org/10.14569/IJACSA.2015.060121>
- Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. Retrieved from <http://arxiv.org/abs/1707.02919>
- Atdağ, 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., 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](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>
- Chang, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22*, 288–296. <https://doi.org/10.1.1.100.1089>
- Cheney, D. (2013). Text mining newspapers and news content : new trends and research methodologies Abstract :, 1–5.
- Feldman, R., Fresko, M., Hirsh, H., Aumann, Y., Liphstat, O., Schler, Y., & Rajman, M. (1998). Knowledge Management: A Text Mining Approach. *Proc of the 2nd Int Conf on Practical Aspects of Knowledge Management (PAKM98, Basel, Swi*(April 2016), 1–10. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.7128&rep=rep1&type=pdf>
- Ghosh, S., Roy, S., & Bandyopadhyay, P. S. K. (2012). A tutorial review on Text Mining Algorithms, *I(4)*, 223–233.

- Godbole, N., & Srinivasaiah, M. (2007). Large-scale sentiment analysis for news and blogs. *Conference on Weblogs and Social Media (ICWSM 2007)*, 219–222. <https://doi.org/10.1177/01461079070370040501>
- 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>
- Pölitiz, 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>
- Sharma, H., & Sharma, A. K. (2017). Study and Analysis of Topic Modelling Methods and Tools – A Survey. *American Journal of Mathematical and Computer Modelling*, 2(2), 84–87. <https://doi.org/10.11648/j.ajmcm.20170202.15>
- Torget, A. J., Mihalcea, R., Christensen, J., & Mcghee, G. (2010). Mapping Texts: Combining Text-Mining and Geo-Visualization To Unlock The Research Potential of Historical Newspapers. A White Paper for the National Endowment for the Humanities, 53. Retrieved from <http://digital.library.unt.edu/ark:/67531/metadc83797/>
- Xiao, R. (2010). 7 - Corpus Creation. *The Handbook of Natural Language Processing*, 147–165. <https://doi.org/doi:10.1201/9781420085938-c7>
- Yin, Z., Cao, L., Han, J., Zhai, C., & Huang, T. (2011). Geographical topic discovery and comparison. *Proceedings of the 20th International Conference on World Wide Web - WWW '11*, 247. <https://doi.org/10.1145/1963405.1963443>
- 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>



## **9. APPENDICES**