Summarising large collection of documents through topic modelling

By

Paul Hakeem Atandoh (paul.atandoh@aims.ac.rw)

June 2017

AN ESSAY PRESENTED TO AIMS RWANDA IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN MATHEMATICAL SCIENCES

DECLARATION

This work was carried out at AIMS Rwanda in partial fulfilment of the requirements for a Master of Science Degree.

I hereby declare that except where due acknowledgement is made, this work has never been presented wholly or in part for the award of a degree at AIMS Rwanda or any other University.

Scan your signature

Student: Paul Hakeem Atandoh

Scan your signature

Supervisor: Dr. Yabebal Fantaye

ACKNOWLEDGEMENTS

My sincere thanks goes to the Almighty God for His grace and guidance throughout this program. I would like to express my heart-felt gratitude to my supervisor; Doctor Yabebal Fantaye, whose coaching and influence yielded this research project. I also thank my co supervisor; Doctor Xavier Vollenwelder for his contribution towards making this project a success. To my tutor; Doctor Jan Hazla, i say thank you very much for the assistance and helping scrutinize my work. To my project partner; Alice Irankunda i appreciate all the support and assistance. Next, my heartfelt gratitude goes to the Tutors, my Classmates, the Academic Director and the entire AIMS Family for their diverse contributions in one way or the other to making me successful in this program, i say a big thank you and may God richly bless you. Finally, I would like to express my appreciation to my family back in Ghana for their support and encouragement.

DEDICATION

This is optional.

Abstract

¡¡¡¡¡¡¡ HEAD Information retrieval is one important but challenging practice carried by organizations that deal with data. For example, we can be given thousands of documents and be interested in a single piece of information from one of them. Manually going through document by document in search of the information can be tedious and time consuming. It would be helpful to at least narrow the search by identifying which documents are relevant to the topic our information belongs.

Topic modelling provides a simple way to understand the contents and extract information from large documents without reading through them. We used the Latent Dirichlet Allocation (LDA) for this task. The LDA is a statistical model that groups words that occur together in document based on context and then assigns a probability value for each word in the group. These groups are called the "topics" and each group represents the summary content of one or more documents. As a case study we used reports from the International Federation of Red Cross and Crescent Societies (IFRC). ======= Information retrieval is one important but challenging practice carried by organizations that deals with data. [Jan: s/deals/deal] For example given thousands of documents and the only interest is just a single information from these numerous documents. [Jan: Rephrase: "For example, we can be given thousands of documents and be interested in just a single piece of information from one of them."] Manually going through document by document in search of information can be tedious and time consuming. [Jan: Add sentence: "It would be helpful to at least narrow the search by identifying which documents are relevant to the topic our information belongs to."] [Jan: Make a new paragraph here.] Topic modelling provides a simple way to understand the contents and extracting [Jan: s/extracting/extract] information from large documents without reading through them. We used the latent Dirichlet allocation (LDA) for this task. [Jan: Capitalization] The LDA is a statistical model that group [Jan: s/group/groups] words that occur together in document based on context and then assigns a probability value for each word in the group. These groups are called the "topics" and each group represents the summary content of one or more documents. As a case study we used reports from the international federation of red cross and crescent societies (IFRC). [Jan: Capitalization.] ¿¿¿¿¿¿¿ 3e4e6217f7743e21f214deeef2ba0e72e632ebcc

Contents

1. Introduction

This research is interested in how information can be accessed easily from a large chunk [Jan: s/chunk/collection] of data. As the stored data increases, [Jan: s/stored data increases/amounts of stored data increase] we are faced with the challenge and the difficulty in trying to explore and extract what we need. For the purpose of easily understanding the contents of the data or know [Jan: s/know/knowin] what it talks about, topic models serves as a tool for handling this task. [Jan: Rephrase: "Topic models serve as tools for the purpose..."]

Topic models are statistical tools with a collection of algorithms that reveals [Jan: s/reveals/reveal] the key components to understanding a document. Huge magnitude of unlabelled text can be analysed with a topic model.

Topic model algorithms provides the environment that allows [Jan: s/provides/provide] users to explore the text in details and summarize it no matter how large their size. Topic model is very useful in identifying the patterns of words in a document and in the event that more than one document is involved in the ML, documents with similar patterns can be related. [Jan: Change "ML" to "machine learning".]

Topic models are unsupervised method of ML, [Jan: Full stop here. Later: "Through various algorithms, they are able to..."] through various algorithms is able to produce cluster of words that represent somewhat a summary of a document. They are applied in search engines to recommend to users what they are interested in. A typical summary for a collection of documents is for the analysis of web search, producing results for users in further search (Turpin et al., 2007). [Jan: I didn't get last sentence.]

This research focuses on summarizing reports from the international federation of red cross and crescent societies (IFRC). [Jan: Capitalize: International Federation of Red Cross and Crescens Societies.] [Jan: Do not jump between topics. Put information about IFRC after the formula with L and S.] The summary provides a representative topic for each document or in other words best cluster of words that summarize the document. Mathematically, it can be perceived as a function that takes a large text and converts to small one, in a way that thematic structure of the large original document is preserved.This can be represented as :

$$f : L \longrightarrow S, \quad \text{Such that} \quad |S| << |L|.$$

[Jan: Also use "\ll" instead of <<.] $L =$ Large text or document

$S =$ Summarized document or small document.

Intuitively the size of S is smaller than L. Manually going through the reports and trying to understand what each is talking about can be time consuming and challenging. The IFRC is a non-governmental organization that provides humanitarian assistance to victims who suffers [Jan: s/suffers/suffered] a disaster event. Through this aid the IFRC generates data of the occurrences of disaster [Jan: s/disaster/disasters] worldwide. Large volumes of complex information are locked in the reports and it is hard extracting [Jan: s/extracting to extract] them going by the manual approach.

This research will employ the Latent Dirichlet Algorithm (LDA). The LDA model extract the contextual meaning from a given large text.

This research is divided into five chapters, the second chapter elaborates some key concepts of topic modelling, IFRC and similar work. The third chapter discuss [Jan: s/discuss explicitly LDA/discusses LDA explicitly] explicitly LDA, and the tools that were also used to arrive at the final results. The fourth chapter covers results and discussions. Chapter five presents conclusion and recommendation.

2. Literature Review

2.1 Unsupervised learning (SL)

This is a machine learning task of inferring a function to describe hidden structure from unlabelled data. This type of ML does not require any prior manual categorization of observations in the data.

The distinction between supervised learning and unsupervised learning (UL) is that in unsupervised learning there is no evaluation of accuracy of the algorithm used, because data fed to the learner is unlabelled . Also one advantage of UL over SL is that time and cost is saved in labelling as required in SL.

2.2 Natural language processing (NLP)

It is a multidisciplinary area that deals with the automatic processing of human language. This automation allows communication between humans and computers. The computer accept input in the form of text or speech and then produces structured representations showing the meaning of those strings as their output.

2.3 Word embeddings

Word embeddings is a dense representation of words in a low dimensional vector space. Bengio et al. (2003) introduced the concept of word embedding and then train them in neural language jointly with model parameters. (Mikolov et al., 2013) came out with the popular word embedding model known as the Word2vec. Faruqui et al. (2014) released Glove. The Glove and the Wor2vec are both aimed at producing word embeddings that ecode the general semantic relationship.

2.4 Topic model

A topic model is a statistical tool that produce a short description of an original document. Topic models can be applied on a single document or a collection of documents. (Blei, 2012b) described topic models as algorithms that discovers the main themes existing in a large text or document and otherwise the combination of two or more documents. He further reveal that the development of probabilistic topic modelling by ML researchers as a set of algorithms that is geared towards revealing and describing large archives of documents with thematic information. Topic models analyses words in the large text document to discover the themes that pervades them, the connection that exist between the words and their occurrence with time. In topic

modelling the stress of having to label the documents prior to annotations is saved as it is been done in supervised learning. From the analysis of the original document the topics are obtained. Given a very large volume of electronic archives that is impossible for human annotations, topic modelling can help to summarize and organize it.

2.5 Topic model methods

Topic modelling techniques have been developed to automatically summarize document or large text. Some popular models are, Latent Semantic Indexing/Allocation (LSI/LSA), the Latent Dirichlet Allocation (LDA) and the Probabilistic Latent Semantic Analysis (PLSA) . This research will be restricted to LDA.

2.6 Vector space model (VSM)

In Natural Language Processing (NLP) especially in semantics similarity, documents can be represented as a vector of words in in a vector space model Salton et al. (1975). The frequency of the word in the document determines its importance. Given two documents with words "desk" and "shirt". From the matrix table below "desk" appears 6 and 4 times in document 1 and document 2 respectively, and the word "shirt" appears 3 times in document 1 and 5 times in document 2. Geometrically this can be represented as shown in figure 2.1.

Table 2.1: Document of words

	desk	shirt
Doc_1	6	3
Doc_2	4	5

Computing the distance between them tells the extent of their similarity. The dot product shows how close the two vectors are to each other. The dot product of two vectors is given as

$$X \cdot Y = |X||Y| \cos(\theta).$$

Although the VSM is a simple model to use, one main limitation is that, it cannot handle polysemy and synonyms issues. Polysemy is a term that describes words with multiple meaning and synonyms are words that have similar meaning. For example, a polysemy word such as "light" can be used in the context of weight of an object or to describe a type of electromagnetic radiation. Words like "detect", "find", "uncover" and "reveal" are synonymous, they can be used to describe a particular event.

Considering a query search in Google, by the SVM, the documents relating to the query will not be revealed in the search results. However a research conducted by Erk and Pado pointed out the reason leading to the limitations in SVM. The paper indicated that the existing model does not take syntatic structure into account. Their research resulted in a model called "structured vector space (SVS)". This work incorporates the context in which words are used.

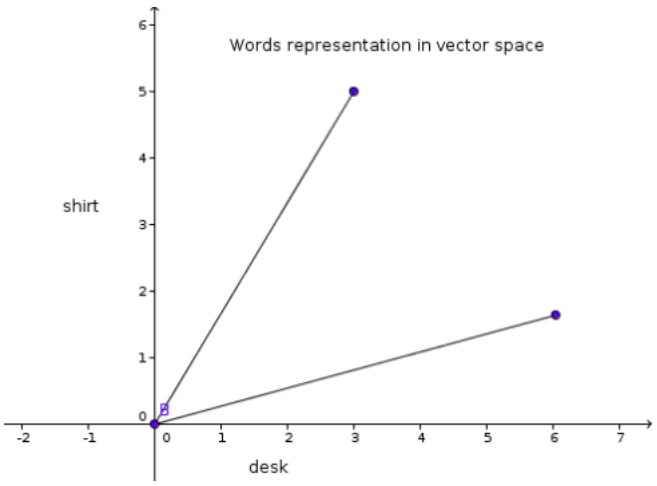


Figure 2.1: Documents represented in vector space, the cosine angle between the two vectors measures the similarity between the two vectors.

2.7 Latent semantic analysis (LSA)

Also known as the latent semantic indexing (LSI) is a topic model method that transforms documents of high dimension to low dimension of words. One useful role played by LSI in topic modelling is its ability to deal with polysemy and synonyms Deerwester et al. (1990).

Preliminarily it constructs a matrix $M \in \mathbb{R}^{n*k}$, from the documents $d_1, d_2, ..., d_k$ of words $w_1, w_2, ...w_n$. The rows represents the different words and the columns can be viewed as different documents. For example from table 2.2, m_{ij} shows the position and the frequency of the word w_i in document d_j. To achieve reduction in the dimension of the matrix M the truncated Singular Value Decomposition is applied, given as:

$$M \approx A_t \sum B_t^T.$$

A_t and B^T are orthogonal matrices, whilst $\sum$ is a diagonal matrix. Reducing the dimension leads to reduction of noise Deerwester et al. (1990).

Table 2.2: Corpus of documents

	d_1	d_2	d_3
food	0	0	2
school	2	5	0
cash	0	1	0
automobile	1	0	4

2.8 Latent dirithchet allocation (LDA)

Blei (2012b) referred to LDA as the simplest topic model . The ideas underlying this model is every document has several topics existing in it. He defined topic to be a distribution over a

fixed a vocabulary. Each topic is made up of words that are very related to the topic. Considering an article with a title "Seeking Life's Bare (Genetic) Necessities," for which data analysis was used to determine the number of genes an organism needs to survive. By hand, words pertaining to three different vocabularies were highlighted with different colours. Words such as "computer", "prediction" linked to the topic "data analysis" highlighted blue, "life" and "evolve" about "evolutionary biology" highlighted pink and words like "gene", "DNA" describing the topic "genetics" is highlighted yellow. Stop words that occur frequently in the article are removed.

The LDA as a statistical tool uses this idea based on the assumption that topics are generated prior to words assignment. The LDA also assumes a model of generating documents. All words in each vocabulary has a probability value and depending on the topic each word finds itself would be high or low. For example the word "gene" will have a low probability value if it is in the domain of the vocabulary "data analysis" compared to when it belongs to the topic "genetics". The idea describing the process of generating documents using words is:

1. From the documents, a random selection of some topics deemed to describe the documents.
2. for each word in the documents:

2a. Randomly choose a topic from the selected topics in step 1.

2b. Randomly choose a word from the selected topic. The topic has a collection of words of which randomly one is chosen at a time.

The LDA model reflects the idea of multiple topics exhibited by documents.
Figure2.2 gives a picture of the whole intuition of this generative probabilistic process:

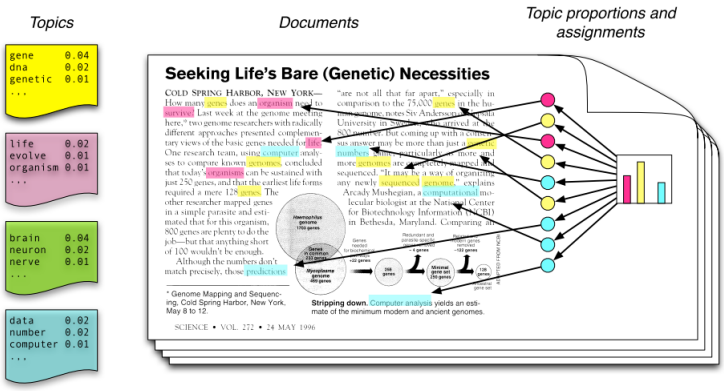
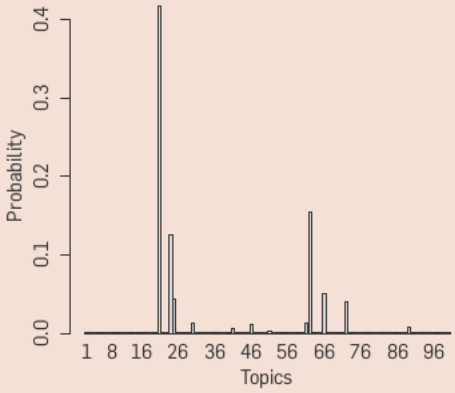


Figure 2.2: Generating documents by reverse process of the LDA model, By hand, words pertaining to three different vocabularies were highlighted with different colours. Words such as "computer", "prediction" linked to the topic "data analysis" highlighted blue, "life" and "evolve" about "evolutionary biology" highlighted pink and words like "gene", "DNA" describing the topic "genetics" is highlighted yellow.

In short description of Figure 2.2 the idea underlying LDA is that, first of all some number of topics that are distribution over words is assumed (far left). In generating for each document,

firstly choose a distribution over topics (far right ie. histogram), then the circles of different colours are topic assignment for which words drawn from the document corresponds to.
Figure 2.4 shows real inference with LDA, using 17000 articles from the journal of science. "Genetics", "Evolution", "Disease" and "Computers" represents the topics from one article and the words below each are top 15 most frequent words. The graph on the left shows the probability values for each topic. The probability values for this article for a given set of topics may be different from another article. In effect, even though some documents or articles may share the same topics, each article exhibits the topics in different proportions.



"Genetics"	"Evolution"	"Disease"	"Computers"
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 2.3: Four Inferred topics consisting of the 15 top words for each topic, the bar graph on the left is the topic probability distribution in the document.

2.9 Graphical model of LDA

Figure 2.4provides a graphical representation, showing both the observed and latent variables involved in the generative process. Latent variables are variables that are not directly but inferred from the the observed. The only observed part is the shaded circle $W_{d,n}$, α and η are parameters from the Dirichlet distribution. What the notations stands for:

- D:the number of documents
- N: number of words in each document
- K: number of topics
- θ_d: the topic proportion for each document d.
- $Z_{d,n}$: is the topic assignment for word n in document d.

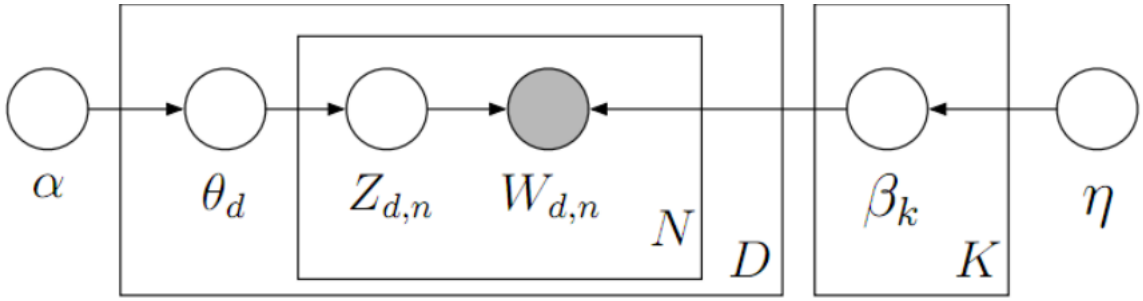


Figure 2.4: Graphical representation of LDA model

- $W_{d,n}$: is the observed worD n in document d,
- β_K:topics

From 2.4 the joint distribution or the total probability of both latent and observed variables is given by:

$$P(\beta, \theta, Z, W, \alpha, \beta) = \prod_{k=1}^{K} P(\beta_i) \prod_{d=1}^{D} P(\theta_d) \prod_{n=1}^{N} P(Z_{d,n}|\theta_d) P(W_{d,n}|\beta_k, Z_{d,n}). \tag{2.9.1}$$

2.10 Posterior distribution

This is a type of Bayesian statistic that describes how latent variables are obtained given the observed data. From the LDA documents generative model a joint probability distribution of both the hidden structures and observed variables. To compute the conditional distribution given the observed variables (words), the posterior is used, given by:

$$P(\theta_{1:D}, \beta_{1:K}, Z_{1:D}|W_d) = \frac{P(\theta_d, \beta_k, Z_d, W_{1:D})}{P(W_{1:D})}.$$

$P(\theta_d, \beta_k, Z_d, W_{1:D})$ can be computed easily for any setting of the hidden variables. $P(W_{1:D})$ is the marginal probability of the observed word variable. This is computed by considering all possible instances of the hidden topic structure by summing the their joint distribution. Because the possible topic structure is large, it is very hard to compute using this posterior relation.
There exist a number of algorithms categorized as "sampling based algorithms" and "variational based algorithms". These algorithms approximate the posterior distribution based on the joint probability distribution between the latent variables and the observed in the posterior relation.

2.11 Variational bayes (VB)

This is a method for approximating the posterior distribution. Recalling from the difficulty encountered by the posterior in computing the marginal probability of the observed variable $P(W_{1:D})$. This form of approximation operates when we have both observed and hidden variables, but also latent parameters.
The VB serves two purposes, of which are:

- VB gives an analytical approximation to the posterior probability given the observed and latent variables, so as to use the approximate values to make some inference.

- To determine from it a lower bound for a marginal probability of the observed data, also known as the evidence. That is, it is the marginal probability of the data given the model. This idea is used to select a best model, on the premise that the model with the highest marginal likelihood shows how best is the data used to fit the model. This also justifies that there is a high probability that the data used was generated by the model.

2.12 Diritchlet distribution

It is from the exponential family of continuous multivariate probability with the parameter α of positive real. It is denoted by $D(\alpha)$.
Suppose $S = [S_1, S_2, ..., S_d]$ represent the probability mass function (pmf), for each $S_i \geq 0$, then $\sum_{i=1}^{d} S_i = 1$. Also suppose $\alpha = [\alpha_1, \alpha_2, ..., \alpha_d]$ with $\alpha_i > 0$ for each i, and let $\alpha_0 = \sum_{i=1}^{d} \alpha_i$. Then S is said to have a Dirithchlet distribution with parameter α, which is denoted by $S \backsim \text{Dir}(\alpha)$, if s is a pmf then,

$$f(s, \alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{d} \Gamma(\alpha_i)} \prod_{i=1}^{d} s^{\alpha_i - 1}. \tag{2.12.1}$$

However if s is not a pmf it has $f(s, \alpha) = 0$.
Where $\Gamma()$ is the Gamma distribution.

2.13 Advantages and disadvantages of LDA

LDA can be used as a built in module in other models to perform more difficult tasks. The LDA model is used in other applications, not only in text. Blei and Jordan (2003) carried out a work that used pairs of LDA modules to model relationships between images and their corresponding descriptive captions . But also include problems involving collections of data, including data from domains such as collaborative filtering, content-based image retrieval and bioinformatics.

The bag of words assumption (the order words in a document does not matter) of LDA makes it unrealistic, however it is reasonable if only our task is to uncover the course thematic structure of the texts Blei (2012b).

2.14 Extentions of LDA

Wallach (2006) developed a model that does not ignores the assumption that the order of words does not matter (bag of words). This means that the each word generated by the topic depends on the previous word. Griffiths et al. (2007) combined the idea of syntactic and semantics to produce a generative model. This model is capable of simultaneously finding syntactic classes and semantic topics despite having no knowledge of syntax or semantics beyond statistical dependency.

3. Methodology

This chapter presents detailed description of the concepts underlying the techniques and tools that are employed to achieve the goal of the research. Detailed description of how the LDA algorithm works in Gensim with respect to the how the results were obtained. All other concepts and prior tools adopted for data preprocessing are also explained.

3.1 Data

The source of the data for this research is from the website of IFRC. Practically the data was obtained by algorithms implemented in the R studio, automatically downloaded the over one thousand pdf reports from the website. Each report is named a name of a country depicting that the reports describes disaster that occurred in a particular country. Each report has an appeal id, several documents might refer to the same appeal id. The appeal id is the unique code given a particular report. Reports describing the same event have the share the same appeal id.

The total number of reports obtained from the IFRC website is 1259 pdf files. The reports were converted to txt format before preprocessing and the model training were carried out. It is easy and convenient to perform such tasks when files are in the txt format.

3.2 Packages and modules

Before training the model on the data some modules that are important to enable smooth running of the algorithms were imported. Apart from Gensim and NLTK that are deeply explained in subsequent sections of this chapter, the role of the remaining are briefly highlighted below.

- os: This module was used to locate path of the directory containing files needed to do the machine learning. And also to combine the various paths to these files.
- codecs: Specifically codec.open from the codecs module opens the encoded text files. It takes 5 positional arguments but 4 were necessary to obtain the required output. They are "filename", "mode=r", "encoding=utf-8" and "errors=ignore", the "bufffering" argument was ignored.

3.2.1 Gensim.

Gensim is an open source toolkit implemented in python to execute task involving vector space models and topic modelling Rehurek and Sojka (2010). Some features of Gensim employed in this research are "term frequency inverse document frequency (TFIDF)" and "LDA". Before executing the above features the "corpora" and "doc2bow" modules are used to represent large collection of texts as vectors and to convert the vector of texts into numbers respectively.

3.2.2 Natural language toolkit (NLTK).

This is also an open library with set of modules that enhances the processing of human language. It is originally developed by Steven Bird and Edward Loper both in the Department of Computer and Information Science at the University of Pennsylvania. This provided a landmark for researchers to contribute to making it more robust and an efficient library. The "corpus" and the "tokenize" are some modules relevant in topic modelling.

3.3 Data preprocessing

This section highlights the preliminary techniques applied on the data sets before training the model. Preprocessing helps to prepare the data to be compatible to the training algorithms, but also removes noise (irrelevant items) from the data, since their presence do not contribute towards obtaining desired results. This practice is very important in topic modelling because it gives an idea to understanding the various terms in the corpus.

3.3.1 Tokenization.

This describes the process of splitting a text or a collection of texts into each single term constituting the text. Each term is known as the token. It can be a "word", "symbols", "punctuations", "numbers". For example, given the text "she won a prize worth 30 million dollars", after tokenizing we have "she", "won", "a", "prize", "worth", "30", "million", "dollars". Prior to creating a vector representation of terms in the document tokenizing is done. The NLTK library is employed to implement this process.

3.3.2 Stop Words, punctuations and irrelevant Terms.

Prior to training the model on the data, words commonly known as stop words, such as "is", "at", "the", "or" are removed. This words are removed because mostly they do not represent the major significant terms to understanding a given article or document. Stop words are also removed to reduce the number of unique tokens in the corpus. The nltk has a model that automatically removes stop words. Punctuations and numbers are removed, it is reasonable to mention that the former cannot contribute to uncovering the thematic meaning of documents, they are just symbols. Because we are dealing with a number of documents and documents have pages, terms related to page numbers are removed. As well as terms with length of one, for example "x", "y" and empty spaces are deleted.

3.3.3 Dictionary and corpus.

To be able to train a topic model, the data sets have to transformed to specific form to make sure the model is able to work with it. The dictionary is a list of unique words in all the documents. It is created with the "corpora.Dictionary" module from gensim tool kit. Having a dictionary created allows us to know the total number of unique terms and all words positions in all the documents collections.

A corpus represents the converted words to bag of words, each word is in a tuple form , the first and second component stands for the "word id" and its "frequency" respectively in a particular document. For example $(1452, 100)$ in document 1 means the $1452th$ word appears 100 times in document 1.

3.4 Term frequency inverse document frequency (TFID)

This measures the extent to which words are important in a document. In topic modelling we want to find a group of words that describes a vocabulary. For example topic modelling a document that talks about a university, words such as "classrooms", "library", "lecturs", "courses", "grades" would tend to be the most important words that describe the topic. It is worth noting that important words are not necessarily the most frequent words, possible to be judged by our intuitive notions.

The TFIDF transforms a vector of integer values into a vector of real values, maintaining the dimension of the original vector.After transformation features which are not frequent in the corpus may have their values increased. That does not mean that all rare words important, some may not be significant at all in the description of the topic. For instance dealing with our "university" document, a word like "congregation" may be rare, but then it is significant towards describing the vocabulary. On the other hand a word such as "consequently" may appear very frequent which in this case does not really say anything about the topic. The most frequent words are mostly words such as "the" or "and," which helps to construct a sentence, thereby making it readable and understandable. These words do not carry any importance to help topic model a document.They are stop words and they are removed before the modelling irrespective of their number.

Given a collection of document with each document d containing words, where each word in the document is denoted i. The frequency of occurrence of a word i in document d is denoted f_{id}. The term frequency TF_{id} computed as:

$$TF_{id} = \frac{f_{id}}{\max_t f_{td}}$$

. Which means that the frequency of the word i in document d is f_{id} normalized by dividing it by the term with the highest frequency in the same document of occurrence with stop words exclusive. Intuitively the word which occurs most frequently would have would have a TF of 1, and other words get fractions as their term frequency for this document. The IDF for a word i which occurs in d of the D documents is computed as:

$$IDF_i = log_2 \frac{D}{d_1}.$$

The TFIDF for a term i in document d is will now be:

$$\text{TFIDF} = TF_{id} \cdot IDF_i.$$

3.5 LDA implementation

This model is a type of topic model that is based on the imaginary assumption that 'words are used to generate documents' or simply "it is a document generative process". For example given N words for a particular document. To generate the documents given the words, firstly choose words deemed to be abstracts topics of the entire corpus, where each topic has a probability value

θ_d, known as the per document topic proportion. For each word we choose a topic $Z_{d,n}$, from the list of topics already chosen. And then choose a word $W_{d,n}$ from that topic . Each word chosen is placed in the document and the process is repeated until all the N words are exhausted to fill the document. The only variable that is observed is the $W_{d,n}$, the remaining are hidden variables. In practical terms the LDA model works in a reversed manner. What it actually does is that it finds the latent variables given the $W_{d,n}$, using the posterior distribution. Because the posterior is difficult to compute the variational inference approximates the posterior. This approximation are embedded in the LDA algorithm. The LDA model in gensim takes several positional arguments but only seven is used, their functions in the model are briefly described below.

- Corpus: This is the vectorized form of all the documents. Each list of tuples represents a document.
- Id2word: Instead of the model using the word numeric id's in the corpus, it is assigned a dictionary containing the words. This makes it easier to read the model output because words constituting a topic will be displayed. Without assigning the id2word the dictionary the words which are distribution over a topic will be shown as numeric id's.
- Num_topics: This takes the number of topics that the model should output. The number of topics chosen is subjective and could be subjected to the number of documents in the corpus.
- Chunk_size: Depending on the number of documents that the model should process at a time is assigned to this parameter.The choice of documents to be processed is dependent on the total number of documents in the corpus. It our model the 1259 is processed at once.
- Update_size: This tells the model how many times to update per the given chunk_size. The default size in the model is 1 which was maintained.
- Passes: Is the number of times that the corpus should be passed over in the modelling process. High value of passes is recommended to obtain better results.
- Alpha: The alpha is a hyper parameter value responsible for controlling the per document topic distribution from the Dirichlet distribution. The "alpha"parameter can be assigned to any arbitrary value, but in our case it assigned to "auto" to allow the model to learn the data and properly assigned an alpha parameter value. Higher alpha values produce more topics, and thus makes documents looks more similar to each other, and for low alpha values, results in documents having few topics.

  Griffiths and Steyvers (2004) Suggested an alpha (α) value of $\frac{50}{T}$, where T is the number of topics and 0.1 for eta (η).

3.6 Visualizing the topic Model

Visualizing topic models is very important because it gives a a clear graphical representation of the topics and their associated terms. One tool that is used to visualize topic models is the

LDAvis. After the LDA model is fitted, the LDAvis is used to interpret the model and understand it better. It takes the "corpus", "Dictionary" and the "LDA model" to produce a visual results. The following can be determined with the LDAvis tool.

- Proportions of each word in a topic and in the entire corpus.
- How often Topics that occur in the topic model.
- The relationship between topics in the topic model.

  The stacked and heat graphs were also used to visualize the content of document topics.

3.7 Python and R

The research employs the Python and R programming languages to implement all the task necessary to achieve the desired goal associated with this research. Preliminary the R studio provided an environment to execute algorithms that scrapped the pdf reports from the IFRC website. Converting these reports from pdf to txt formats were also handled by the R studio. All other algorithms pertaining to the machine learning process were done in Python.

4. Results and discussions

This chapter presents the results obtained from the series of algorithms implemented in Python and detailed discussions of the results are also presented.

4.1 Removal of red cross-specific stop words

Prior to plotting figure (4.1) the top ten occurring words in the dictionary of the 1259 text documents are "red", "cross", "support", "ifrc", "national", "activities","operation", "volunteers", "people" and "emergency". These words were removed because they are perceived to be often used in almost all the text documents, considering the nature of the reports and the source. In other words they are key words that are always used in reporting events concerning disaster and related issue. Even though there is no clear motive for removing these words, one reason thought to be convincing is that these words repeats more frequently in almost all the topics when trained on the LDA model. And also it was difficult to identify unique topics as result of this repetitions. However these repetitions may be viewed to make sense because the reports talks about similar issues. These repetition effect continued even after the initial top words in the documents were removed, but the repetitions decreased with the next set of frequently occurring words.

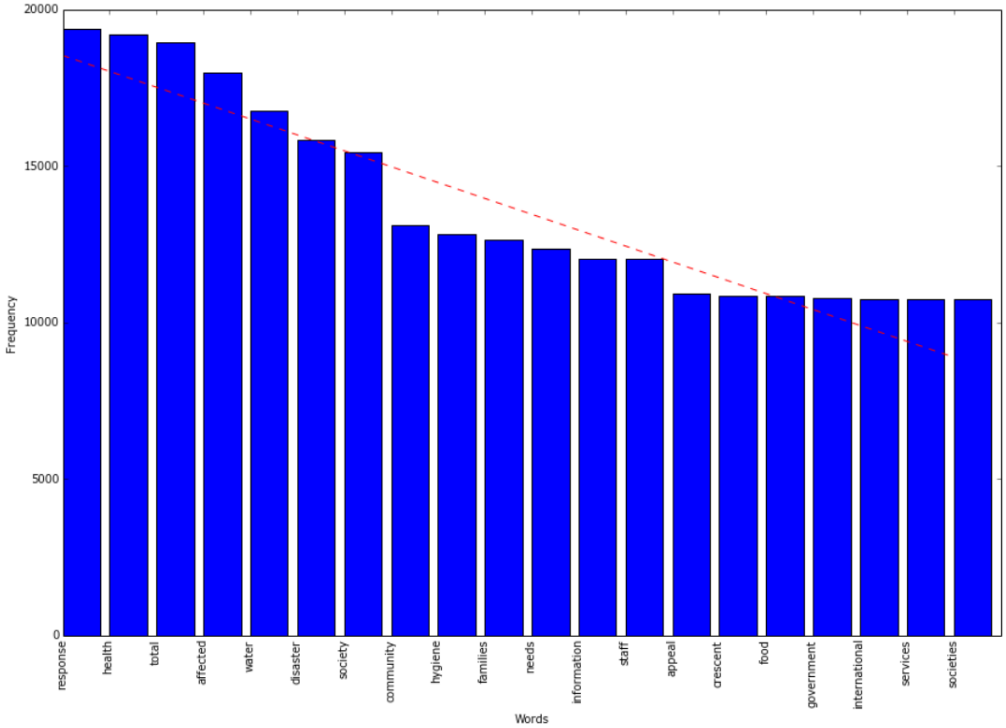


Figure 4.1: Top 10 words in the text collections, the red dashed lines shows a steady decrease in the trend of words in the corpus.

4.2 Topics Distribution over words

Table 4.2 shows 11 topics selected out of the 30 topics that were generated by the LDA model. These topics were selected on the basis of uniqueness of their constituent words distribution.

Topic number 7, 10, 13, 14, and 18 can be considered to describe cash transfer programs (CTP). Food, cash and shelter are some forms of CTP's given to victims who have suffered a disaster event. Some reports out of the 1259 text files are CTP reports. It is very likely that these topics are associated with these particular set of reports. Topic 18 particularly relates to documents describing the source of the CTP's. The "sarc" word in this topic is an abbreviation that stands for Syrian Arab Red Crescent, a humanitarian NGO (private) that provides some form of CTP to victims.

The term "mena" in topic 27 is an acronym for countries in Middle East and Northern Africa. The other words occurring together with "mena" in this topic such as "syrian", "lebanon",and "iraq" are terms related to countries that falls in the domain of MENA. For easy query of documents or reports in relation to countries of MENA, topic 27 can be used to trace and easily find such documents.

Topic 28 has 3 very common disease outbreaks that affected some countries across the globe. The "zika", "cholera", and "chikungunya". The chikungunya and zika diseases spread are associated with mosquitoes, and the Dominican republic has been hit with the zika and chikungunya diseases.

	Topics	Words
0	7	[food, cash, appeal, security, households, children, nutrition, monitoring, beneficiaries, health]
1	9	[societies, health, response, disaster, relief, items, needs, families, floods, dref]
2	10	[rc, total, report, timeframe, budget, expenditure, income, funding, rcrc, appeal]
3	13	[appeal, financial, statements, chf, societies, expenditure, contributions, international, costs, period]
4	14	[nrcs, districts, bdrcs, families, cash, affected, shelter, water, district, distribution]
5	16	[health, cases, total, dref, social, response, outbreak, community, mobilization, vaccination]
6	17	[ebola, guinea, government, total, community, sierrra, leone, liberia, response, virus]
7	18	[from, private, donors, line, donations, total, sarc, chf, society, government]
8	19	[krcs, county, kenya, counties, total, staff, affected, health, nairobi, garissa]
9	27	[ircs, lrc, food, rc, iraq, crescent, syrian, lebanon, mena, displaced]
10	28	[health, cholera, cases, community, zika, prevention, information, dominican, chikungunya, outbreak]

Figure 4.2: Top 11 topics from the trained LDA model.

4.3 Documents distribution over topics

The trained LDA model revealed that several topics are related to the various documents used to train the model. Figure 4.3 shows the extent to which documents can consist of multiple topics. Each of the vertical bars corresponds to a document and each division is a topic. All the ten documents in 4.3 exhibits more than one topic. Document 7 is distributed over 10 topics with documents 2 and 5 having the least number of topics distribution, they have two topics each. The sum of the probabilities for each bar of its constituent rectangle values adds up to 1.

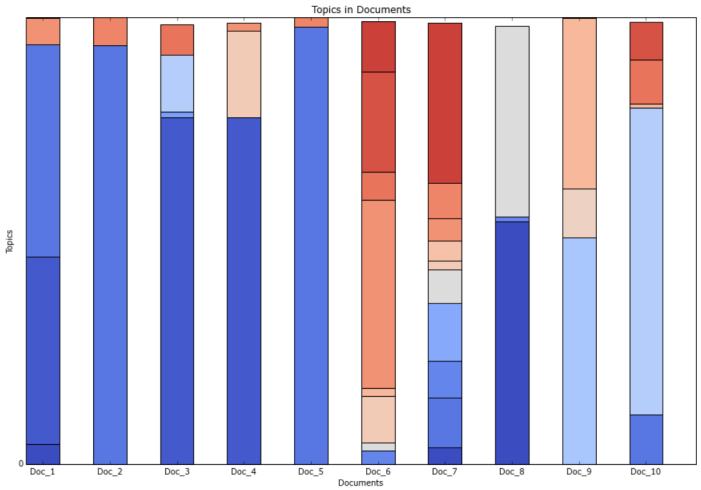


Figure 4.3: Showing the extent to which documents consist of multiple topics.

4.4 Topics Concentration in Documents

The LDA topic model informs us about the fact that documents constitutes multiple topics, but also tells us the proportion and how concentrated the topics are in each of the documents. From 4.4 the blue coloured rectangles varies with intensity. As shown in the the probability scale on the right of 4.4, deeply coloured means high probability $(0.9 - 1)$, in that order. It is observed that topic 3 is highly concentrated in documents 2 and 5. The proportion of topics 3, 4, 7, 15, 18, 19, 23 and 24 occur in small proportions in document 7.

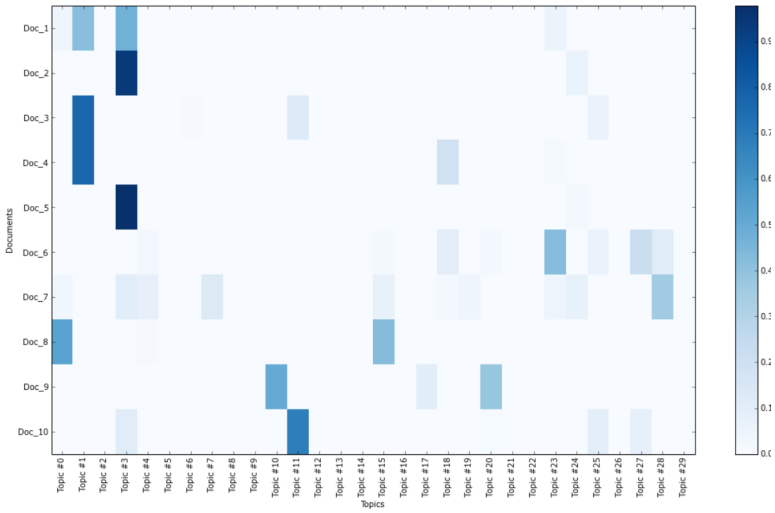


Figure 4.4: The deep blue coloured area corresponds to high proportion of topics in a documents and lightly coloured refers to low probalities in corresponding documents.

4.5 Topics and Documents relationship

The nature of the distribution of the terms in each of the topics can predict how the documents are related. That is to say if some of the terms are found in many topics, then the topics can be said to be related and hence the documents.

From figure (4.5) circles (0-29) are the topics and below it are the terms constituents of each topic. This implies that red circle (topic 26) is composed of the words "appeal" down to "allocations", these words are the top most 30 terms connected to topic 26.

The size of the circle tells how prevalent it is in the corpus. That is, the bigger the circle the more prevalent that topic is in the documents. Figure (4.5) shows how frequent topic 1 occurs in the documents more than the other 29 topics.

The map also reveals the semantic distances between the topics. Figure (4.5) shows that nineteen of the topics are crowded in one region, the extreme right end of the horizontal axes. This means that more than 50% of the topics are similar and hence gives an indication that most of documents in the corpus share similar topics.

502 ¡¡¡¡¡¡¡ HEAD

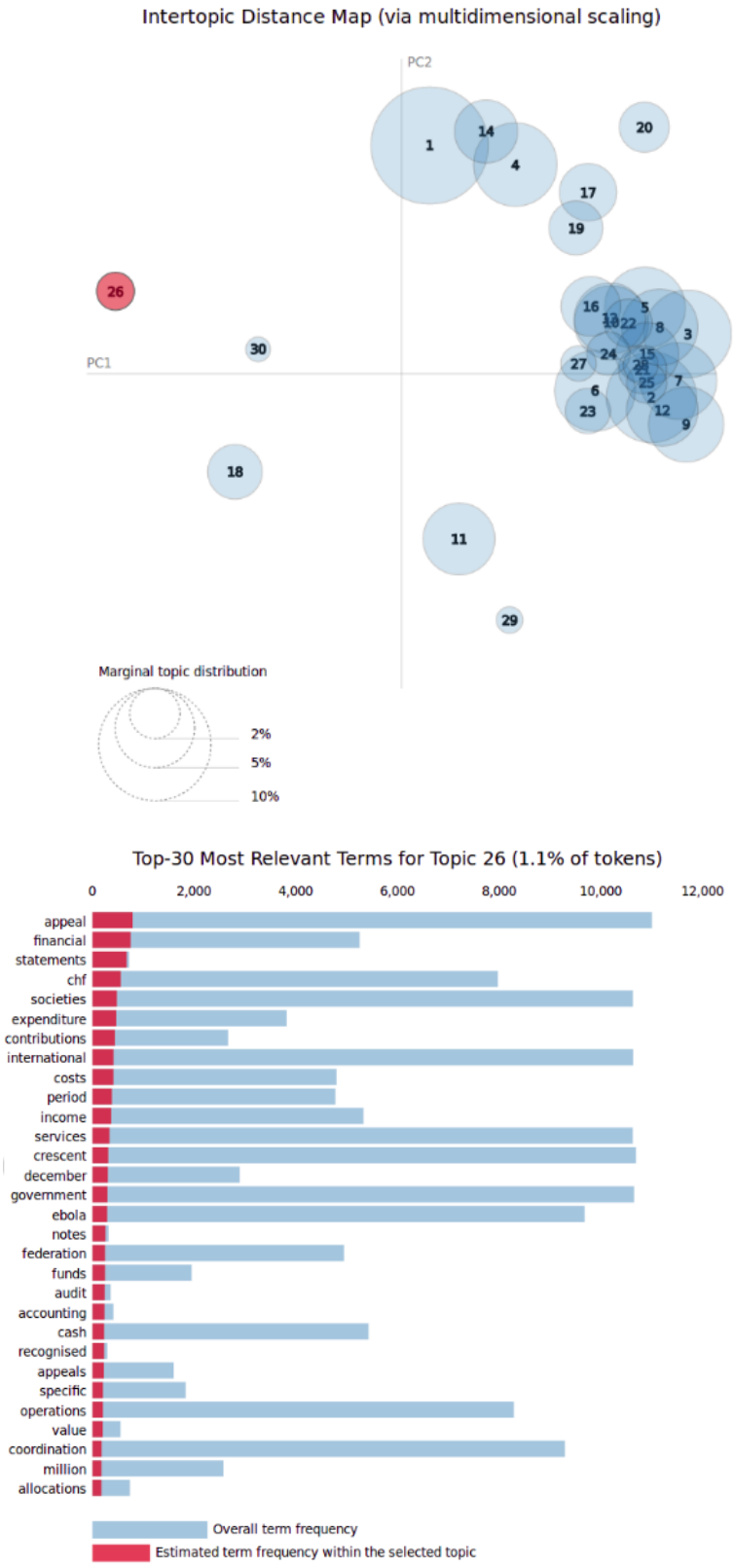


Figure 4.5: The circles (1-30) are the topics and the prevalence of words in topics and in the corpus is shown in below, the red bar and blue bar represent the word frequency in topics and corpus respectively, topic 26 is composed of the words "appeal" down to "allocations", these words are the top most 30 terms connected to topic 26.

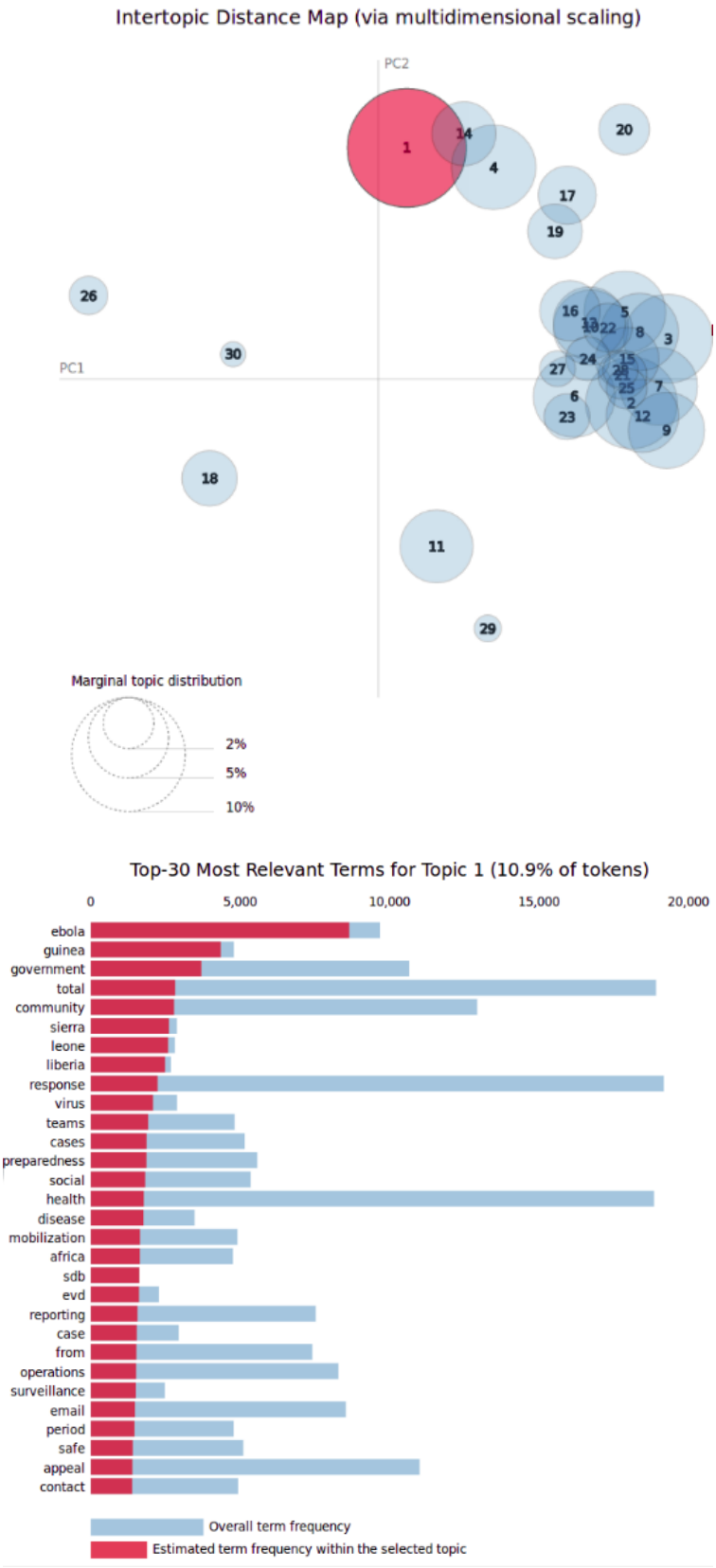


Figure 4.6: The circles (1-30) are the topics and the prevalence of words in topics and in the corpus is shown in below, the red bar and blue bar represent the word frequency in topics and corpus respectively, the first top 5 words of topic 1 are "ebola", "guinea", "government", "total" and "community', these are the top most 30 terms that constitutes topic 1, it is the most occuring topic in all the documents.

5. Conclusion and way forward

The era of topic modelling gave rise to some number of topic modelling approaches, the Vector Space Model (VSM), Latent Semantic Analysis (LSA), Latent Dirichlet Allocaton (LDA) etc. This research employed the LDA to summarise reports from the International Federation of Red Cross and Crescent Societies. The LDA model takes a number of parameters, notably the number of topics to be generated The quality of the topics that the LDA generates also depends on the other parameters. Our work assigned the model to generate 30 topics. We observed that the model exhibits some randomness in generating the topics. In the sense that when the input parameters are maintained and the model command is executed several times, different but similar set of topics are obtained. The observed difference is that, some number of new topics replaces some topics that occurred in previous results. This means that the model can generate many topics, depending on the "topic number" parameter value assigned in the model. Hence there exist more than 30 topics in the corpus.

In spite of the fact the LDA gave some good results, the results could have been better if some techniques had been applied during the data preprocessing stage. Techniques such as lemmatization and stemming aim at grouping words together that come from one root and assigning base word to them. For example, topic 18 has the words "donors" and "donations". With lematization the two words can reduced to their base form as "donor", so that everywhere in the corpus these words are replaced by "donor". Future work will incorporate these techniques to improve the quality of results. =======

6. Conclusion

[Jan: Do not use flushleft, here or elsewhere!]

The era of topic modelling gave rise to some number of topic modelling approaches, the vector space model (VSM), latent semantic analysis (LSA), latent dirichlet allocaton (LDA) etc. This research employed the LDA to summarise reports from the international federation of red cross and crescent society. [Jan: Capitalization for algorithm names and IFRC.] The LDA model takes a number of parameters, notably the number of topics. The model outputs the exact number of topic number inputs. [Jan: Merge those sentences into one: "The LDA model takes a number of parameters, notably the number of topics to be generated."] The quality of the topics that the LDA generates also depends on the other parameters. Our work assigned the model to generate 30 topics. We observed that the model is random in generating the topics. [Jan: s/is random/exhibits some randomness] In the sense that when the input parameters are maintained and the model command is executed several times, different but similar set of topics are obtained. The observed difference is that, some number of new topics replaces some topics that occurred in previous results. This means that the model can generate many topics, depending on the "topic number" parameter value assigned in the model. Hence there exist more than 30 topics in the corpus.

In spite of the fact the LDA gave some good results, the results could have been better if some techniques had been applied during the data preprocessing stage. Techniques such as lematization and stemming aimed [Jan: s/aimed/aim] at grouping words [Jan: s/together/that come] together from one root and assigning base word to them. For example, topic 18 has the words "donors" and "donations". With lematization [Jan: s/lematization/lemmatization] the two words can reduced to their base form as "donor", so that wherever [Jan: s/wherever/everywhere] in the corpus these words are replaced by "donor". Future work will incorporate these techniques to improve the quality of results.

¿¿¿¿¿¿ 3e4e6217f7743e21f214deeef2ba0e72e632ebcc

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

David M. Blei. Surveying a suite of algorithms that offer a solution to mannaging large documents archives. *Commuinication of the ACM*, 55:1–7, 2012a.

David M Blei. Surveying a suite of algorithms that offer a solution to managing large document archives probabilistic field models. *review articles april*, 2012b.

David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134. ACM, 2003.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

Katrin Erk and Sebastian Padó. A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 897–906. Association for Computational Linguistics, 2008.

Nicholas Evangelopoulos. *Comparing latent dirichlet allocation and latent semantic analysis as classifiers*. PhD thesis, University of North Texas, 2011.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*, 2014.

T Griffiths and M Steyvers. Latent semantic analysis: A road to meaning. *chapter Probabilistic topic models, Laurence Erlbaum*, 2006.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. Topics in semantic representation. *Psychological review*, 114(2):211, 2007.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.

Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

583 Claude Shannon. The best detection of pulses. In N. J. A. Sloane and A. D. Wyner, editors,
584 *Collected Papers of Claude Shannon*, pages 148–150. IEEE Press, New York, 1993.

585 Andrew Turpin, Yohannes Tsegay, David Hawking, and Hugh E Williams. Fast generation of
586 result snippets in web search. In *Proceedings of the 30th annual international ACM SIGIR*
587 *conference on Research and development in information retrieval*, pages 127–134. ACM, 2007.

588 Jeffrey D Ullman, Jure Leskovec, and Anand Rajaraman. Mining of massive datasets, 2011.

589 Hanna M Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international*
590 *conference on Machine learning*, pages 977–984. ACM, 2006.