DOI 10.29042/2018-3764-3768

Helix Vol. 8(5): 3764- 3768

Overview and Application of Text Data Pre-Processing Techniques for Text Mining on Health News Tweets

[*1]Gauri Chaudhary, [2]Dr. Manali Kshirsagar

[1]Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, India

[2]Professor, Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, India

**Email:chaudhary_gauri@yahoo.com*

Received: 09th July 2018, Accepted: 14th August 2018, Published: 31st August 2018

Abstract

With the huge abundance of textual data in electronic form on the internet, there is a growing need to identify and extract meaningful information and patterns from this underlying text data. While in a broader context this is Data Mining, since applied to text data, this gives rise to a specialization within data mining, called text mining. Various useful applications of text mining are clustering, classification, trends analysis etc. This paper focuses on various techniques of pre-processing the text data so that it is transformed into a form to which text mining can be applied. Various data pre-processing techniques have been applied to health news from more than 15 major health news agencies collected using Twitter API, available on UCI Machine Learning Repository and output at the end of each pre-processing step is specified.

Keywords: Text Pre-processing, Tokenization, Stemming, Stop Words Removal, TF-IDF, Text Mining.

Introduction

In last few decades there has been vast increase in textual data in electronic form such as collection of web pages, research papers, emails etc. Previously data mining was meant for structured data such as that stored in relational databases or data warehouses. Today, all information available with various organizations is stored electronically in text databases (or document databases). Text data is semi-structured because it is neither structured nor unstructured. Free flowing text stored on the web is an example of unstructured data. In order to extract useful information i.e. interesting patterns from unstructured textual data, we use text mining techniques.

Text mining is analysis of text data and extraction of knowledge from this data. In order to analyze and perform any kind of operation on text data, the text data needs to be transformed to numbers. Only then can any specific text mining technique be applied to it. [05]. There is a need to extract data from large number of documents, compare them, find similar documents, select relevant information from them, and find patterns across multiple documents. Hence text mining is applied on data stored in large document collections. Text mining involves various tasks like Information Retrieval, Classification, Clustering, Summarization, Trends Analysis, Association Analysis, Visualization etc.

The objective of this research is preparing or pre-processing data for text mining. Fig. 1.0 below describes the complete system flow for text mining. It starts with collection of textual data (also called the document corpus) from one or more sources (example: web pages or emails or digital libraries).

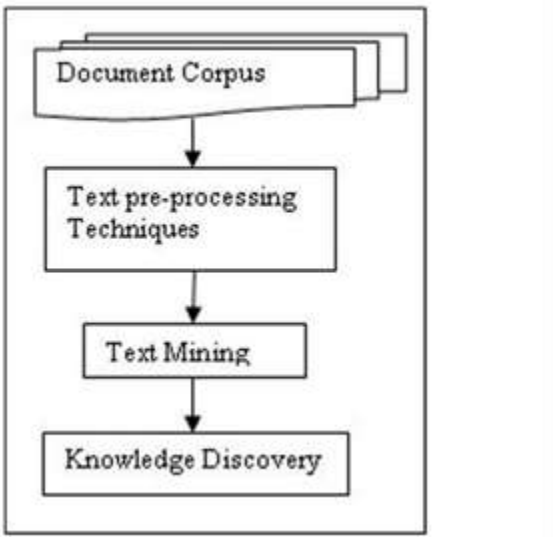

Fig 1 Text Mining Process

The text data may be noisy, so it needs to be cleaned and converted to a numerical form so that text mining can be applied to it. This is the function performed by pre-processing techniques. In this paper we describe in detail the steps carried out for pre-processing text data.

Text mining can then be applied on the pre-processed data. An example application of text mining is document clustering which can be performed. Document clustering aims at identification of natural groupings of similar texts in large document collections and partitioning them in such a way that data points in a cluster are more similar to each other than the points in other clusters. Typical application could be to group similar news articles in clusters so that they can be retrieved more efficiently from

clusters with similar key words rather than look up in the entire document corpus. This is nothing but the knowledge discovery as specified in the Fig. 1

Pre-Processing Techniques

The objective of text pre-processing is to transform and represent the raw text data in an efficient format to which text mining techniques (example: document clustering) can be applied. The various stages of text data pre-processing include the following (also depicted in Fig. 2):

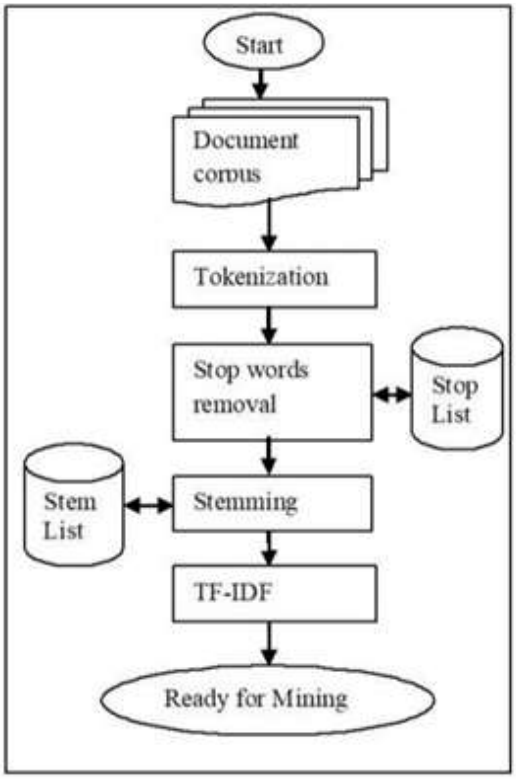

Fig 2 Text Pre-Processing Steps

Tokenization

Parsing and dividing any text into individual meaningful elements like words or phrases or symbols is called tokenization. [02]. In order to identify individual words in a text, following steps are generally followed:

Parse the text and identify the tokens using separators such as spaces, line break or punctuation mark is encountered.

Whitespaces and punctuation marks may or may not be included based on requirement.

All characters within a contiguous string are part of one token. The characters can be alphabets, alphanumeric or numeric.

Stop Words Removal

Stop words removal is a process of cleaning the documents by removing stop words (i.e. noisy words). [04]. Stop words are terms that do not contribute to the meaning of the document. They are typically articles, pronouns, prepositions etc. which just add up to the feature space. Hence they are removed from the documents. Examples of stop words are the, a, in, an, with, but etc.

Stemming

Words in any language are modified to exhibit various grammatical functions such as tense, gender, number, feeling, person, case etc. So same word is expressed in different forms according to the grammatical purpose it has to perform. [03]. Stemming is the process of reducing a term to its root. For example the stem of all the terms Push, Pushed, Pushing, Pushes would be "Push" as depicted in Fig. 3

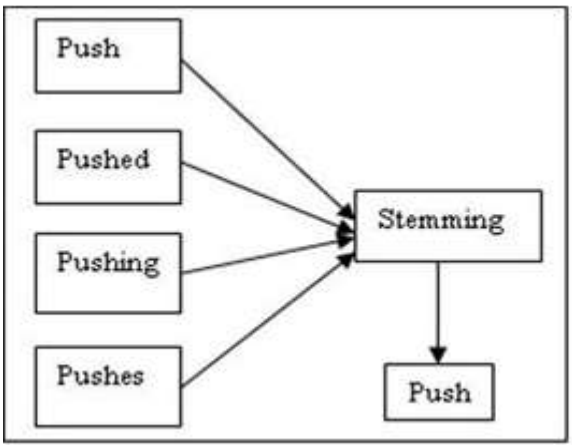

Fig 3 Stemming Process

The objective of stemming is to eliminate the suffixes from words and bring different forms of the same word to the same root. Thus, the number of words is reduced and comparison of matching words across documents becomes easy. Saving and processing only word roots also saves memory space and time. Various algorithms are available for stemming like Porters stemmer, Paice/Husk stemmer, Lovins stemmer, Dawson stemmer, Snowball Stemmer etc. [07].

One of the widely used algorithms for stemming is Porters stemming algorithm. It removes suffixes from words. In English language, generally the suffixes are built by grouping smaller and simpler suffixes. [01].

Transformation to Vector Space Model Using TF-IDF

In order to apply text mining, it is essential to convert the text data to numerical values. Vector space model is a popular method used to represent documents [08]. In the context of text mining, every distinct term in a document is considered as a feature. Thus a document is a collection of features. Any document in the corpus is represented as a vector of features and each

individual item in this vector is a term weighted score value in the specific document. [06]. One of the popular approaches for calculating the term weight is term frequency - inverse document frequency (TF-IDF).
TF-IDF is generally used to represent the document in Vector space (an array of numbers, each number representing the score TF*IDF for a term). TF*IDF is a value that indicates the significance of a term within a document compared to the complete document collection (corpus).

TF*IDF calculation:
TF of a term is the count of number of times the term occurs within a document. It is calculated as:

Number of times term appears in a document divided by the total number of terms in the document

IDF of a term is the count of number of documents in the complete document corpus that contain the term. It is calculated as:
$\log_e$ (Total number of documents divided by the number of documents containing the term)

TF-IDF is then calculated for each term as:
TF(term)*IDF(term).

The documents are thus converted into vectors, each document being represented as a vector of values, each value represents TF(term)*IDF(term) as described above. With this representation, we are ready for text mining!

Experimental Results
The Dataset
The dataset that is used is Health News in Twitter Data Set which is downloaded from UCI Machine Learning Repository. The dataset contains health news from more than 15 major health news agencies such as BBC, CNN, NYT etc.

The dataset consists of individual text file for each news agency's Twitter Account. Each line within the file contains data in the following format:

Tweet Id | Date and Time | Tweet

Application of Pre-Processing on the Dataset:
Development environment used is ASP.net using C#. Fig 4 below shows the user interface designed in ASP.net.

Fig 4 User Interface in ASP.net for Text Pre-Processing

Listed below are the steps for pre-processing the health news tweets data and the output at the end of each step:

1. Load documents
All the documents are parsed to only load the tweets first (discarding tweet ids and timestamp using separator i.e. pipe (|) and new line characters). Only the tweets are thus loaded and written to data files (one for news from each news agency). Fig. 5.0 gives a snapshot of just the tweets loaded in a text file.

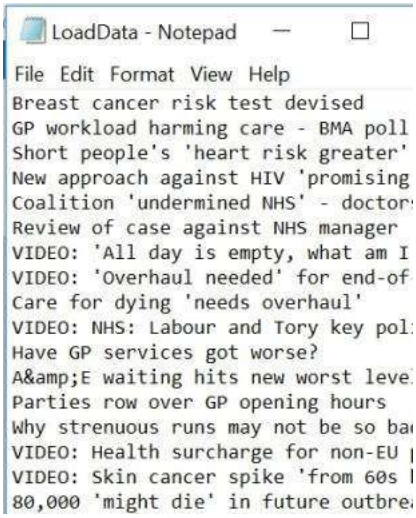

Fig 5 Loading Health News Tweets

2. Tokenization:
Each news file is then parsed for delimiters such as spaces and new line characters to separate words/ tokens.

3. Stop Words Removal
Next, a list of stop words is defined which in the context of text mining carry no meaning like articles, pronouns, prepositions etc. and those are discarded from further processing.

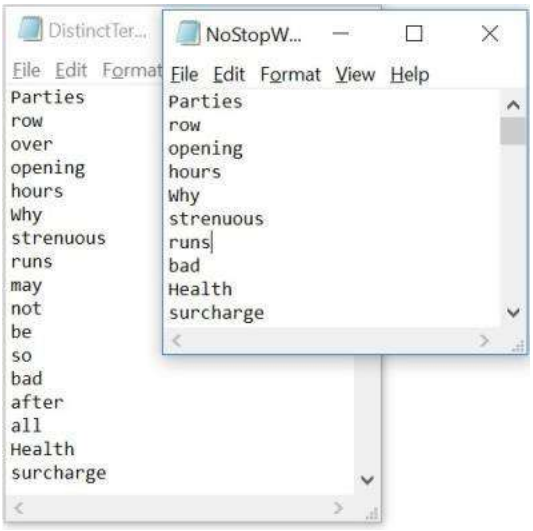

Fig 6 Tokens and Stop Words Removed

Fig 6 shows the list of tokens in one text file(labelled Distinct Terms in the Fig) and the stop words removed in another text file (labelled No Stop Words in the Fig). As can be seen in the Fig 6, the stop words such as why, may, so, be etc. are all removed.

4. Stemming

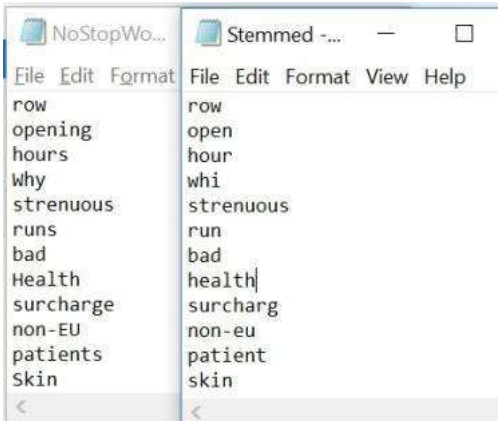

Fig. 7 Stemming Output

The next step is to reduce the terms to their root. Porter's stemming algorithm is applied for stemming. The stemming output of small subset of the data is shown in Fig 7.0. As can be seen, the word "opening" is replaced by "open, "runs" is replaced by "run" and so on.

5. Transformation to Vector Space Model

The next step is to transform the document space to vector space using TF-IDF. Following are the steps:

i) Each of the documents is represented as string of tokens removing stop words and using stemmed words as described in previous steps.
ii) Read each document string and construct a dictionary of n distinct terms in the complete document corpus.
iii) Consider each distinct term in the dictionary and calculate the term frequency against each document.
iv) Calculate Inverse document frequency for each of the dictionary term against the total number of documents.
v) Calculate the score TF*IDF for each term. The documents can thus be represented as n-dimensional vectors with TF-IDF score for each of the n dictionary terms.

Fig 8 below shows a snapshot of TF-IDF values for each of the documents. First the name of the document is displayed (e.g. Health-Tweets\bbchealth.txt). Then for each of the dictionary term, TF and IDF values are displayed separated by tabs (e.g. The TF of the term "cancer" is 191 and IDF is 3.332205 in the document bbchealth.txt. For simplicity only top 10 most frequently appearing terms in each document are displayed.

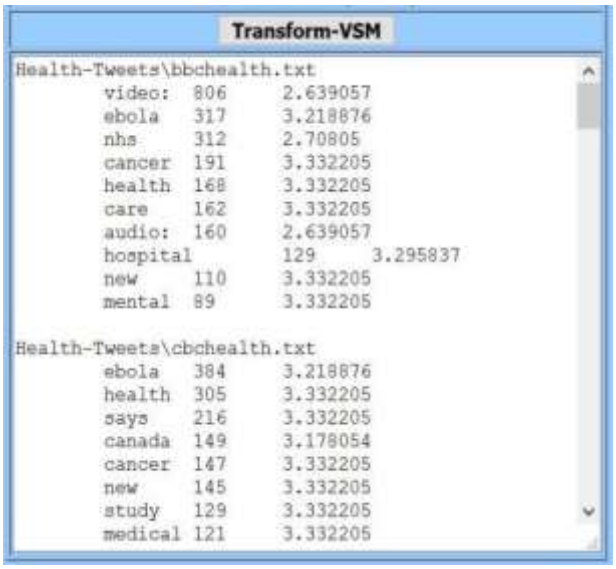

Fig. 8 Term, TF, IDF Listing for the Documents

Conclusion and Future Scope

Text mining has been increasingly important in recent times due to the huge creation and usage of electronic documents. The objective of text mining is to extract useful information such as patterns, trends or knowledge from textual data. It is essential to convert the text data to numbers to perform text mining. Various pre-processing methods are described in this

paper to convert the text documents to a form suitable for further knowledge mining. The text pre-processing techniques such as tokenization, stop words removal, stemming and document representation as TF-IDF vectors have been described and their application demonstrated on Health news tweets dataset. Thus the researchers should be able to understand and apply the pre-processing techniques described in this paper for any text mining applications.

References

1. Dr.S.Vijayarani et al. 2015. Preprocessing Techniques for Text Mining - An Overview. International Journal of Computer Science & Communication Networks, Vol 5(1),7-16. ISSN:2249-5789

2. D Sailaja, M.V.Kishore, B.Jyothi, N.R.G.K.Prasad. 2015. An Overview of Pre-Processing Text Clustering Methods. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 3119-3124.

3. R. GeethaRamani, M. Naveen Kumar, Lakshmi Balasubramanian. 2016. Identification of Emotions in Text Articles through Data Pre-Processing and Data Mining Techniques. International Conference on Advanced Communication Control and Computing Technologies (ICACCCT).

4. Chalitha Kulathunga, D.D. Karunaratne. 2017. An Ontology-based and Domain Specific Clustering Methodology for Financial Documents. International Conference on Advances in ICT for Emerging Regions (ICTer): 209 – 216.

5. Arian Dhini, I.B.N. Sanditya Hardaya, Isti Surjandari, Hardono. 2017. Clustering and Visualization of Community Complaints and Proposals using Text Mining and Geographic Information System. 3rd International Conference on Science in Information Technology (ICSITech)

6. Abdullah Alsaeedi, Khalid Aloufi, Mohamed Abdel Fattah. 2017. A hybrid feature selection model for text clustering. 7th IEEE International Conference on System Engineering and Technology (ICSET 2017), 2 - 3 October 2017, Shah Alam, Malaysia

7. Mayuri Mhatre, Dakshata Phondekar, Pranali Kadam, Anushka Chawathe, Kranti Ghag. 2017. Dimensionality Reduction for Sentiment Analysis using Pre-processing Techniques. Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication

8. Hrishikesh Bhaumik, Biswanath Chakraborty, Anirban Mukherjee, Siddhartha Bhattacharyya, Manojit Chattopadhyay. 2014 IEEE. Towards Reliable Clustering of English Text Documents using Correlation Coefficient. Sixth International Conference on Computational Intelligence and Communication Networks

9. Deepak Agnihotri, Kesari Verma, Priyanka Tripathi. 2014 IEEE. Pattern and Cluster Mining on Text Data. Fourth International Conference on Communication Systems and Network Technologies.

