Two Important Methods of Text Mining and Simple Visualization Techniques in Suggestions of Single Sided Defines

Anu Babu[*], Josemon Achenkunju [**]

[*] B. N Gupta Research Centre, Dr. S. R. Chandrasekhar Institute of speech and Hearing, Bengaluru
[**] Department of Biostatistics, St. Thomas College, Palai

DOI: 10.29322/IJSRP.9.09.2019.p9396
http://dx.doi.org/10.29322/IJSRP.9.09.2019.p9396

Abstract- Recently, the services based on internet has grown rapidly raising an imminent question about its analysis of the information about suggestions of services. The organizations can be collecting the suggestions about the services from the consumer. Many suggestions are often available only through the information such as good or bad and another contain more non useful information by local slang. It is an essential problem when different codes are used to analyze these suggestions or when specific categorization in certain content is missing. This leads to unnecessary complications. In order to improve and automate the process of analyzing suggestions or recommendations we propose an approach such as text mining. The text mining is the process of extract the accurate and overall information from the type of unstructured data. In this article mention the two important methods of text mining and the simple visualization techniques in the suggestions of the online course Single Sided Defines. The normalization and tokenization are the most important methods of text mining. The aim of this study is to develop a solution for the suggestions of the one-hour course was produced to offer insight into single sided unilateral hearing loss, using text mining methods.

Index Terms- Text mining, Single Sided Defines, Normalization, Tokenization.

I. INTRODUCTION

The purpose of this paper is to highlight text mining methods as a support to identify the relevant literature from a data for recommendations. The aim of this paper work is that examining large collections of written resources to generate new information and to transform the unstructured text into structured data using the different methods of text mining. The text mining is a common process of extracting relevant information from the using set of documents [4]. Text mining provides basic pre-processing methods, such as identification, extraction representative characteristics, and advanced operations as identifying complex patterns. Document classification is a task that consists of assigning a text to one or more categories. Text mining will ever be able to replace the accuracy that manual duration can archive in machine learning extraction [1]. Text mining comprises the discovery and extraction of knowledge from free text and can extend to the generation of new hypothesis by joining the extracted information from several publications [2]. Text mining is a knowledge-intensive task. This is gaining a wider attention in several Original Equipment Manufacturing industries, for example aerospace, automotive, power plants, medical-biomedicine manufacturing, sales and marketing divisions [3]. Text mining can be broadly defined as knowledge intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. In a manner analogue to data mining, text mining seeks to extract useful informationfromknowledgesourcesthroughtheidentificationandexplorationof interesting patterns. In case of text mining, however, the data sources are document collections and interesting patterns are found among formalized database records but in the unrestricted textual data in the documents in these collections [5, 6]. This paper only addresses the summarization of two text mining methods such as normalization and tokenization. This paper is organized as follows: in the next section is described a method about text classification. The experiments performed using the python code and the results obtained with the sets of scientific suggestions considered in the automatic text summary and text classification, are discussed in result section, which is followed by the concluding remarks in summary section.

II. RESEARCH ELABORATIONS

The section talked about the important text mining methods such as normalization and tokenization and other methods.

Normalization

The methods are needed for transferring text from human language to machine readable format for further processing. The normalization is a procedure that making the text into standardized form. There are different type of methods involved in the

normalization procedure of text mining. Firstly, the text can be converted into lower case or upper case. The next method is removed numbers if they are not relevant. To remove the leading and ending space called the white space removal. The next method is to remove stop word using Natural language toolkit (NLTK). The text normalization includes,

1. **Convert text to lowercase or uppercase**

The python code for converting text to lowercase or uppercase as follows,
variable = variable.lower()
variable = variable.upper()

2. **Remove numbers**

Remove numbers if they are not relevant to the analysis. Usually, regular expressions are used to remove numbers
variable2 = re.sub(variable1)
print(variable2)

3. **Remove punctuation**
   It can be used to remove the set of some symbols.
   variable2=variable1.translate(string.maketrans("","),string.punctuation) print(variable2)

4. **Remove whitespace**
   To remove leading and ending space, use the function.
   strip()

5. **Remove stop word**
   The stop words are the most common words in a language like the, a, all, is, all. These words do not carry important meaning and are usually removed from texts. It is possible to remove stop word using Natural Language Toolkit or NLTK, a suite of libraries and programs for symbolic and statistical natural language processing. The python code for stop word removal is,
   variable = set(variable.words(language))

6. **Remove sparse terms and particular words**
   In some cases, it is necessary to remove sparse terms or particular words from texts. This task can be done using stop words removal techniques considering that any group of words can be chosen as the stop words.

Tokenization

The tokenization is the process of splitting the given text into smaller pieces called tokens. Words, numbers, punctuation mark and others can be considered as tokens. The python code for tokenization is the function.
tokenize()

Stemming

Stemming is a process of reducing words to their word stem, base or root form. The main two algorithms are Poster stemming algorithm and Lancaster stemming algorithm. The python code for stemming are
first import PorterStemmer
variable1 = "change word"
variable1 = word_tokenize(variable1)
print(stemmer.stem(word)

Lemmatization

The aim of lemmatization, like stemming is to reduce inflectional forms to a common base form. As opposed to stemming, lemmatization does not simply chop off inflections. Instead it uses lexical knowledge bases to get the correct base forms of words.
first import WordNetLemmatizer
variable1 = "change word"
variable1 = word_tokenize(variable1)
print(lemmatizer.lemmatize(word))

Part of speech tagging

Part of speech tagging aims to assign parts of speech to each word of a given text such as nouns, verbs, adjective and others based on its definition and its context.
from textblob import TextBlob
variable1 = example variable2 = TextBlob(variable1)

print(variable2.tags

Chunking

Chunking is a natural language process that identifies constituent parts of sentences and links them to higher order units that have discrete grammatical meanings.

first import TextBlob
variable1 = example
variable2 = TextBlob(variable1)
print(variable2.tags)

Named entity recognition

Named entity recognition aims to find named entities in text and classify them into predefined categories such as names of persons, locations, organisations, times.

first import ne_chunk
variable1 = example
print ne_chunk(post_tag(word_tokenize(variable1)))

Coreference resolution

Pronouns and other referring expressions should be connected to the right individuals. Coreference resolution finds the mentions in a text that refer to the same real-world entity. In python the function is,

StanfordParser()

Collocation extraction

Collocations are word combinations occurring together more often than would be expected by chance. Collocation examples are break the rules, free time, draw a conclusion, keep in mind and so on. The python code is,

from ICE import CollocationExtractor
extractor = CollocationExtractor.with_collocation_pipeline(pos_cheak = False)
print(extractor.get_collocations_of_length(input, length))

Relationship extraction

Relationship extraction allows obtaining structured information from unstructured sources such as raw text. Strictly stated, it is identifying relations among named entities. For example, from the sentence-Mark and Emily married yesterday, we can extract the information that Mark is Emily's husband [7]. In python we use the function

relextract()

Visualization Techniques

Visualization is the simple and most accurate method for understanding the data. There is different type of visualization techniques are available in the case of text mining. In this paper also includes the simple visualization techniques using python codes. There are 3 types of visualization techniques.

1. Network model
2. Graphical representations
3. Heat map

Moreover, text mining is a one of the simple procedures to analyses the invariable recommendations or suggestions. The normalization and tokenizations are the core methods of the text mining techniques.

III. RESULT

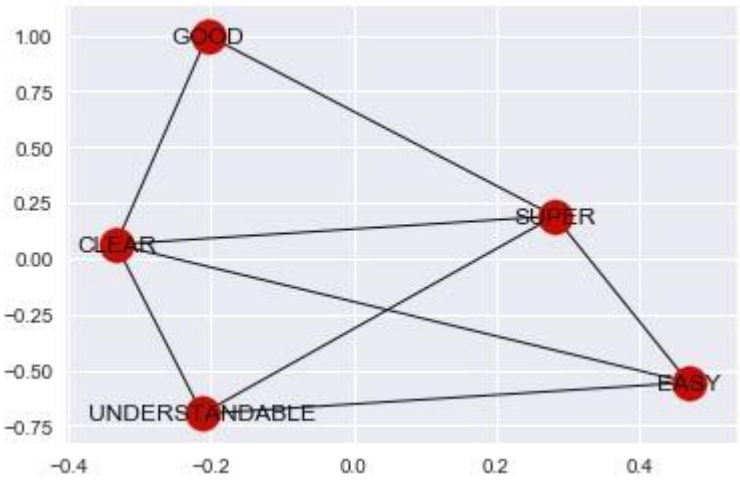
This section includes the overall information about methods of text mining and visualization techniques. In this study a secondary dataset is used. The dataset contains 50 participant's suggestions of professional class, Single sided deafness (SSD). Some of the suggestions are in a single sentence and others are in paragraph. The normalization is a procedure that making the text into standardized form. There are different type of methods involved in the normalization procedure of text mining. Firstly, the text can be converted into lower case or upper case. The next method is removed numbers if they are not relevant. To remove the leading and ending space called the white space removal. The next method is to remove stop word using Natural language toolkit (NLTK).

After applying tokenization method suggestions are splits into tokens. All type of characters except space in the suggestions are considered as tokens. Figure 1 shows the result of tokenization method.

```
[['Description', '\\', '0', 'It', 'covered', 'a', 'variety', 'of', 'options', 'to', 'amplify', 'pat', '...', '1', 'Product', 'd
riven', '2', 'content', '3', 'Very', 'informative', '.', '4', 'I', 'understand', 'the', 'SSD', 'struggle', 'with', '.', 'I', 'h
ad', 'a', 'st', '...', '5', 'A', 'very', 'comprehensive', 'description', 'of', 'the', 'availa', '...', '6', 'Basic', 'but', 'go
od', '!', '7', 'I', 'wanted', 'to', 'hear', 'more', 'about', 'counseling', 'the', 'pt', '...', '.', '8', 'Very', 'informative',
'review', 'of', 'SSD', 'as', 'well', 'as', 'impl', '...', '9', 'I', 'currently', 'serve', 'three', 'families', 'with', 'SSD-ver
y', '...', '10', 'content', '11', 'comparisions', 'made', 'between', 'devices', '12', 'Well', 'written', '&', 'organized', '1
3', 'Clear', 'overview', 'of', 'available', 'treatment', 'for', 'SSD', '14', 'EXAMPLES', '15', 'Very', 'informative', '.', '1
6', 'Well', 'organized', 'and', 'lots', 'of', 'details', 'given', 'to', 'al', '...', '17', 'great', 'flow', 'and', 'appropriat
e', 'amount', 'of', 'info', '18', 'Very', 'well', 'presented', '.', 'Organized', 'information', 'but', '...', '19', 'A', 'goo
d', 'overview', 'of', 'SSD', '.', '20', 'clarified', 'some', 'areas', 'of', 'doubt', '21', 'clearly', 'laid', 'out', ',', 'to',
'the', 'point', '.', '22', 'Interesting', 'study', 'and', 'facts', '23', 'I', 'found', 'it', 'presented', 'in', 'a', 'way', 'th
at', 'was', 'easy', 'to', '...', '24', 'Easy', 'follow', '.', '25', 'Clear', ',', 'concise', '26', 'Covered', 'current', 'fitti
ng', 'solutions', 'for', 'SSD', ',', 'wit', '...', '27', 'A', 'good', 'overview', 'about', 'the', 'challenges', 'of', 'having',
'...', '28', 'Content', '29', 'The', 'information', 'shared', '.', '30', 'A', 'lot', 'of', 'good', 'information', 'presented',
'in', 'a', 'conci', '...', '31', 'Enjoyed', 'the', 'brief', 'historical', 'perspectives', 'offe', '...', '32', 'I', 'really',
'enjoyed', 'reviewing', 'over', 'the', 'main', 'issue', '...', '33', 'It', 'had', 'so', 'many', 'new', 'breakthroughs', 'to',
'share', 'in', 't', '...', '34', 'Presentation', 'of', 'current', 'implants', 'and', 'qualifica', '...', '35', 'Very', 'clear',
'information', '36', 'all', 'meat', 'and', 'potatoes', '37', 'Easy', 'to', 'understand', 'and', 'well', 'written', '38', 'singl
e', 'sided-', 'deafness', '39', 'Well', 'written', ',', 'adressed', 'a', 'variety', 'of', 'treatments', '...', '40', 'This', 'c
ourse', 'was', 'very', 'extensive', 'into', 'the', 'world', '...', '41', 'Information', 'was', 'clear', 'and', 'concise', '.',
'42', 'I', 'have', 'started', 'working', 'with', 'a', 'child', 'with', 'this', '...', '43', 'great', 'information', '44', 'goo
d', 'overview', 'of', 'SSD', '45', 'Spanned', 'the', 'technology', 'for', 'early', 'wired', 'CROS', 'to', '...', '46', 'I', 'ha
```

Figure 1: After tokenization

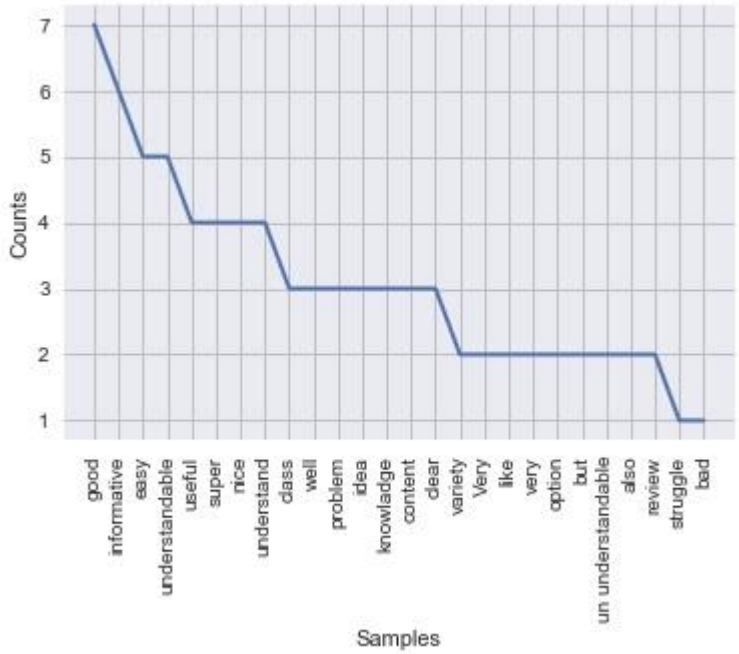
The visualization is the one of the important parts of text mining. The visualization method can useful in easily understand the result. The text data can represent in disparate method of visualization. The network model is the one of the simplest forms of visualization for text mining. The network model can easy to understand for common man. So that in text mining the network model is more preferable. The network model can be used to find the token relationship. The result of network model of connection of tokens are demonstrated in Figure 2.



IV.

V. Figure 2: Network model of connection of tokens

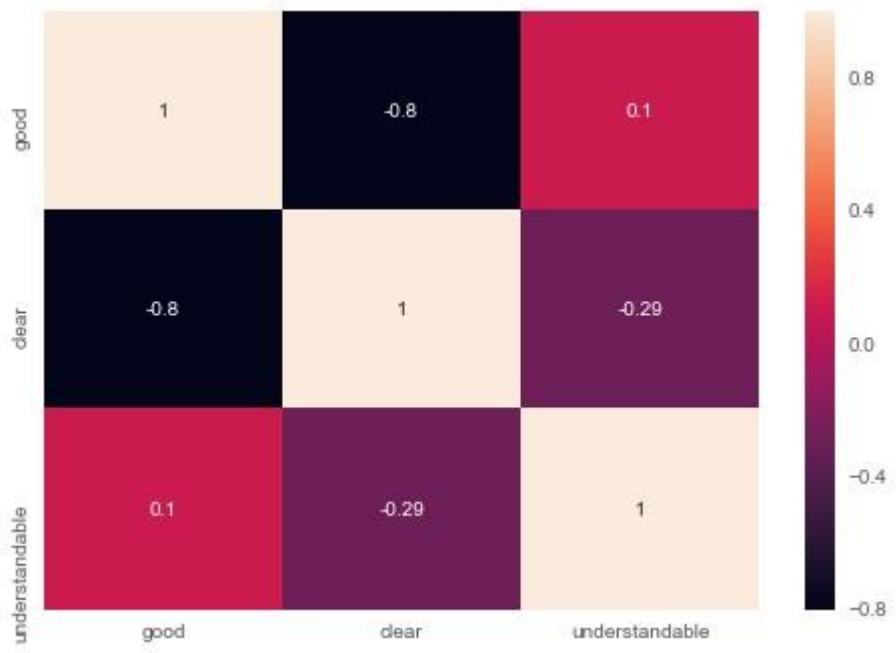
To draw the graphical representation of the text data, it represents the information about the count of tokens. To select some tokens from the list of tokens and evaluate their count.



VI. Figure 3: Count of tokens

From Figure, we conclude that, the word good is repeated more than the other words and the words struggle and bad are use only once. The range of occurrence can be proved by take some tokens from the few suggestions and count their number of occurrences.

Another method of visualization technology is the heat map. The heat map is a two-dimensional representation of data by which values are represented by colors.



VII. VIII. Figure 4: Heat map

The heat map shows that the token good are more occurring token. but the token clear is less occurring than other of the most occurring tokens. The token understandable is the token that second most token of most occurring. The result shows that better understandable of the text mining methods and visualization techniques.

IV. CONCLUSION

This project work is aimed at studying general idea of a text mining and their methods. In this study the suggestion data of SSD are used. The text mining methods are used to analyses suggestions of SSD. From the result, get the idea about the generalization of the

suggestions. That is some words are more occurring in the suggestions. To find the more occurring words using the methods of text mining and analyze it. After the analysis conclude that the overall idea of the suggestions leads to good idea about SSD. This project work is a structure of structures. The project work includes extracting information from suggestions of SSD. It is a learn by doing excursion into the art and science of text mining. To divided the project into three chapters to facilitate the learning approach. The text mining teaches software developers how to mine the vast amounts of information available on the web, internal networks and desktop files, and it turn it into usable data [8]. The text mining and analysis is a new area of research that is less than 15 years old. Much of text mining is observational. Text mining is one of those disciplines that are now emerging as cutting-edge technologies. Only a few years ago, standard desktop computers did not have the physical capability to handle huge amount of textual material or to engage in the complex analysis of page length material. Storage capacity and in particular megahertz and random-access memory were insufficient to appropriately analyze large or complex textual situations. Recently, however a type of personal computer has become powerful and fast enough to engage in sophisticated text mining analysis, which itself has the capability of assisting researchers to better understand a host of social science and biological issues [9]. So, this work includes the information extraction from the unstructured data like suggestions or recommendations.

ACKNOWLEDGMENT

I express my sincere gratitude to all faculty of B N Gupta research center providing necessary advices and supervision.

REFERENCES

[1] Daniel G. Jamieson, Martin Gerner, Farzanesh Sarafraz, Goran Nenadic and David L. Robertson. (2015). Towards semi-automated curation using text mining to recreate the HIV-1, human protein interaction database. *Database*. 2012,1-12.
[2] Dietrich Rebholz-Schuhmann, Anika Oellrich and Robert Hoehndorf. (2012). Text mining solutions for biomedical research, enabling integrative biology. *Nature reviews Genetics.* 12, 829-839.
[3] Dnyanesh G. Rajpathak. (2013). An ontology-based text mining system for knowledge discovery from the diagnosis data in the automotive domain. *Computers in industry*. 64, 565-580
[4] Gabe Ignatow and Rada Mihalcea. (2018). *An introduction to text mining.* SAGE Publications
[5] Gary Miner, Dursun Delen, John Elder, Andrew Fast, Thomas Hill, Robert A. Nisbet. (2012). *Practical text mining and statistical analysis for non-structured text data applications.* Elsevier Academic Press.
[6] Hercules Dalianis. (2018). *Clinical text mining*. Sringer International Publishing.
[7] Text mining methods, https://medium.com/@datamonsters/text-preprocessingin-python-steps-tools-and-examples.
[8] Manu Konchady. (2006). Text mining application programming. Charles River Media.
[9] Gary Miner, Dursun Delen, John Elder, Andrew Fast, Thomas Hill, Robert A. Nisbet. (2012). Practical text mining and statistical analysis for non-structured text data applications. Elsevier Academic Press

AUTHORS

First Author – Anu Babu, MSc Biostatistics, Dr. S. R. Chandrasekhar Institute of Speech and Hearing, Bengaluru anubabu2495@gmail.com
Second Author – Josemon Achenkunju, MSc Biostatistics (Student), St. Thomas College, Palai Jose1996achenkunju@gmail.com.

.