ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 1, January 2020

Identifying Major Focus of a Person Speech using Word Cloud

Dhaval Bhoi[1], Divyesh Patel[2]

Assistant Professor, U & P U. Patel Department of Computer Engineering, CSPIT, CHARUSAT University, Gujarat, India[1,2]

ABSTRACT: Rapid growth due to digital data acquisition led to huge volume of structured, unstructured and semi-structured data. Main focus of this paper is to demonstrate the use of R Programming language to highlight the main focus of the person while giving his speech or writing on a topic of different length. In this paper speech given by different people for the different purposes are used for experiment purpose. R Programming helps us to generate or extract main focus of the speaker or writer during his speech or writing. This will help us to find out the major focus areas of a person & also to identify his attitude.

KEYWORDS: Unstructured Data, R Programming, Word Cloud

I. INTRODUCTION

We are living in the world surrounded by lots of information in the form of structured and unstructured data. Handling structure data is found easy compare to unstructured data [1]. Textual data especially unstructured data mining is found challenging task. Various people gives their speech on different occasion. Teacher delivers his talk or minister delivers his speech on various occasion. People express their views, ideas & share messages in their speech or in writing.

Proper analysis of the people's speech will help us to find out main attention point of a particular speaker. In the education field student will find the main area of concern expressed by a teacher during his lecture. Students also will be able to identify important topics, critical issues and this will help them to prepare for examination in a better way.

In this paper we have taken topics of different length and tried to generate most frequent or emerging pattern from the given unstructured data. To get the best of the result we have used R Programming language [RStudio] as a programming language for implementation.

The remaining paper is organized as follows: Section II will cover Background and related work. Section III describes various process, methods and speech analysis process. Section IV and V contains results discussion and Conclusion & Future work respectively.

II. BACKGROUND AND RELATED WORK

R Programming is a platform independent programming language and freeware software environment for statistical computing and graphics. It is supported by the R Foundation for Statistical Computing. R programming language is also a strong option for accomplishing various data mining task. R Programming language is having various distinct features like, it is open source, graphical, statistical, can handle large data, better visualization and many more. R programming is developed in FORTRAN, C++ and R Programming itself.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 1, January 2020

R Programming helps in performing various data mining tasks like learning hierarchical clustering, association rule mining, for evaluating methods and metrics data visualization and in the field of geoscience [2]. R Programming can also handle big data [4]. As R programming is open source its development is worldwide and it's licensed under free software, GNU GPL 2+. Experiment we performed on the various topics, are done in RStudio.

R Programming languages are widely used in various industries like Facebook, Google, Ford Motor etc. [2]. Industries are using R Programming for effective business planning and its execution. R Programming helps in achieving high paid jobs in the industries. To showcase that uses of R Programming various industries are shown in Table 1. The purpose of uses is also shown.

Table 1. Usage of R Programming in Industries

Name of the company	Uses of R
Facebook	Big Data Visualization, Status update Analysis
Google	To find ROI of advertising (to make online advertising more effective), Statistical Analysis, manipulation and visualization
Microsoft	To understand the user behaviour
Twitter	Data Visualization and Semantic Clustering
Ford Motor	Statistical Analysis and Data driven decision support
John Deere	Time Series Modelling and Geospatial Analysis
Lloyds of London	Development of motion chart to help investor for risk analysis
New York Times	Data Journalism

Using R Programming language, we can integrate with other languages like Python, C, C++ & Java also. It also enables to communicate with many data sources like Access, database and SPSS, Minitab, SAS, Stata like statistical packages. R Programming language and it comparison with other programming language for choosing this specialized and advanced data mining task is also done [3]. Where R Programming is found best suitable to work with big data, link, graph mining spatial data analysis, time series analysis, semi- supervised learning, data streams, text mining, parallelization and deep learning. R Programing is indeed a powerful tool for performing computational text analysis [6]. Hence, we have decided to take R Programming as implementation medium due to its uses in various domains and industries like Facebook, Google, Microsoft, twitter etc. Our suggested approach and process steps are described below.

III. PROCESS, METHODS FOR SPEECH ANALYSIS

To understand and analyse the speech of a person, we have taken very famous speech of Swami Vivekananda that he delivered at Chicago in year 1893. 'tm' package available in R Programming provides a framework to work with text data structure [7].

Process Steps:

1. Installing packages like "tm", "Snowballoc", "wordcloud" and "RColorBrewer" are required.
2. Speech of person is read in a text format [Unstructured Data]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 1, January 2020

3. Conversion from text to document format
4. Inspection of document and its pre-processing
    a. Pre-processing steps include converting text to lower case
    b. Removing numbers
    c. Removing common English common stop words
    d. Removing our own stop-words if any
    e. Text stemming is performed
5. Text document matrix is prepared
6. Creation of word cloud
7. Finding frequent items & Plotting a bar-plot

Once we perform the above steps we get two things as major result or outcome.

1. Word Cloud
2. Most Frequent Words

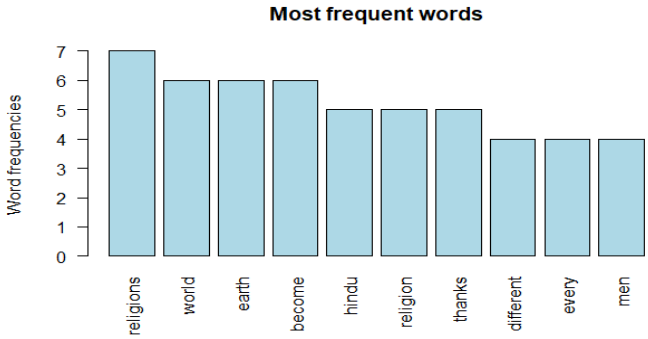

Fig 1. Most Frequent Words for a speech of Swami Vivekananda

From the above Fig 1 and word cloud generated in Fig 2 on the next page clearly shows that religions, world, earth, become and Hindu are the most important 5 words from the speech of Swami Vivekand. From this experiment we can clearly understand the attitude, thought and orientation of a person.

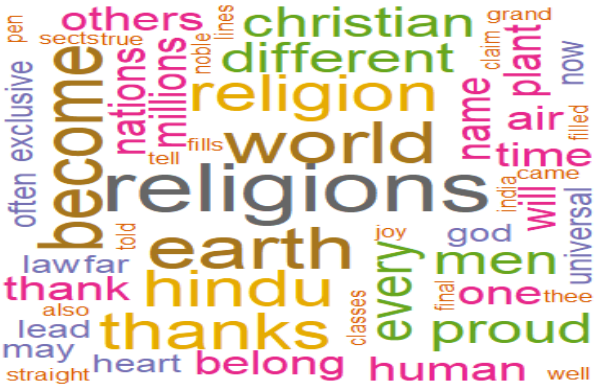

Fig 2. Word Cloud created from the speech of Swami Vivekananda

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 1, January 2020

The above analysis has been performed in RStudio 1.0.136. To extend & analyse the above experiment in more details in the education field we have also compared the essay written by different students on the same topic: "Education" of different length n. We also compared & analysed their results. The way people look and feel towards the same topic is different. The thinking of different age people also found different.

Table 2. Emerging words for different length topic

Length of topic (In words)	Most Emerging Words	Word with highest frequency Word, Frequency
100	Education, life, helps, every, throughout, development	Education, 5
250	Education, life, country, get, people	Education, 10
400	Education, people, areas, country, development, everyone	Education, 15
1223	Education , system, sector, development, country	Education, 30

IV. RESULT AND DISCUSSION

Generated cloud of words as a result of different length same topic is shown on the next page [See Fig. 3, 4 & 5]. Emerging patterns of are shown in Fig. 3 and Fig. 4 for different length of topic e.g. n=100 words and n=1223 words respectively. We can observe that as the size of topic in length increases the focus on the people on the word also increased and frequency of word education also increased linearly. Even the focus of the people for topic of any length, is around the major words like education, life, people, country and development etc.

Fig 3. Word cloud for n=100 words

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 1, January 2020

Fig 4. Word Cloud for n= 1223 words

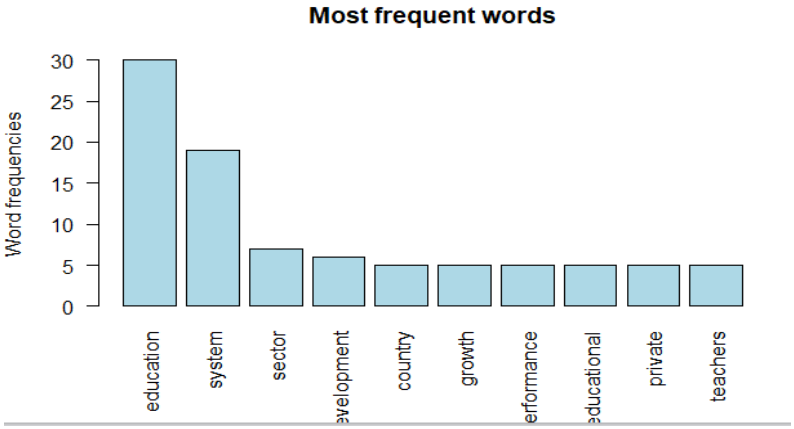

Fig 5. Frequency count for topic of the length 1223 words

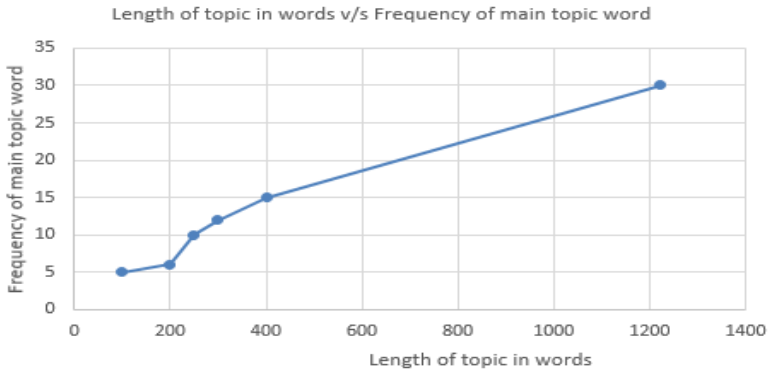

Fig 6. Freq. of Main topic word Vs. Length of topic

Graph of Length of topic document Vs. Frequency of the main topics word is shown in figure 6.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 1, January 2020

V. CONCLUSION AND FUTURE WORK

From the above result and its discussion, we can conclude that the word cloud created from the human speech is not the only end result of their expertise but also their nature, subject expertise, habits etc. As we extended our experiment with different length topic written on the same issue is also as per our expectation. In the future work, person's thought and attitude before & after an event can be analyzed.

VI. ACKNOWLEDGEMENT

The authors would like to thank Head of Department, Principal and Dean of Faculty of Technology and Engineering, and CHARUSAT Management, Charotar University of Science and Technology, Changa for their suggestions, encouragement and support in undertaking this work.

REFERENCES

1. H. Déjean, "Extracting structured data from unstructured document with incomplete resources," in Extracting structured data from unstructured document with incomplete resources, Tunis, 2015.
2. R. M. Bishwal, "Potential use of R-statistical programming in the field of geoscience," in 2nd International Conference for Convergence in Technology (I2CT), Mumbai, 2017.
3. S. Kumar, P. Singh and S. Rani, "Sentimental analysis of social media using R language and Hadoop: Rhadoop," in 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida.
4. Malaviya, A. Udhani and S. Soni, "R-tool Data analytic framwork for big data," in Symposium on Colossal Data Analysis and Networking (CDAN), 2016.
5. Jovic, A. a. Brkic, K. a. Bogunovic and Nikola, "An overview of free software tools for general data mining," in Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on, Pratia, Crotia, MIPRO 2014.
6. K. Welberes, W. V. Atteveldt and K. Benoit, "Text Analysis in R," Communication Methods and Measures, vol. 11, no. 4, pp. 245-265, 2017.
7. Feinerer, K. Hornik and D. Meyer, "Text Mining Infrastructure in R," Journal of Statistical Software, vol. 25, no. 5, 2008.
8. M. C. J., The R Book, John Wiley & Sons Ltd., 2007.