CONTENT MINING AND VISUALIZATION OF PAPERS REVIEWS USING R LANGUAGE

RAMESHKUMAR VALLABHDAS MODHA,RESEARCH SCHOLAR,SAURASHTRA UNIVERSITY,RAJKOT
DR. GIRISH C. BHIMANI, RESEARCH GUIDE, SAURASHTRA UNIVERSITY,RAJKOT

ABSTRACT:

These days, individuals share and examine logical papers via web-based networking media, for example, the Web 2.0, large information, online gatherings, websites, Twitter, Facebook and researcher network, and so forth. Notwithstanding an assortment of measurements, for example, quantities of reference, download, suggestion, and so on., paper audit content is likewise one of the compelling assets for the investigation of logical effect. The web-based social networking apparatuses improve the examination procedure: recording an arrangement online insightful practices. This paper expects to investigate the immense measure of paper audits which have created in the online life stages to investigate the understood data about research papers. We actualized and demonstrated the consequence of content mining on audit writings utilizing R language. Also, we found that Zika virus was the examination hotspot and affiliation look into strategies were broadly utilized in 2016. We likewise mined the news survey around one paper and inferred the general assessment.

KEYWORDS: R language, Text mining, Visualization, Word cloud

INTRODUCTION:

With the approach of the Web 2.0 and the huge information, online gatherings, web journals, Twitter, Facebook and other internet based life administrations have grown quickly. Specialists start to direct their work stream via web-based networking media apparatuses. Academic writing is shared and examined on Twitter and Facebook, composed in social reference administrators like Mendeley and ReadCube, remarked in online journals and small scale websites, announced in news, peer-assessed after production in Faculty of 1000. While the internet based life instruments improve the exploration procedure and researcher correspondence proficiently, they have another amazing bit of leeway: recording a progression of online insightful practices. The arrangement of online academic practices are sorts of advanced follows [1]. In "altmetrics: a declaration", Priem et al. [2] characterize altmetrics as pursues: This assorted gathering of exercises (that reflect and transmit insightful effect via web-based networking media) frames a composite hint of effect far more extravagant than any accessible previously. We call the components of this follow altmetrics (http://altmetrics.org/proclamation/). As indicated by altmetric.com, altmetrics are measurements and subjective information that are corresponding to conventional, reference based measurements. They can incorporate (however are not restricted to) peer surveys on Faculty of 1,000, references on Wikipedia and in open

approach records, dialogs on explore web journals, prevailing press inclusion, bookmarks on reference supervisors like Mendeley, and notices on informal organizations, for example, Twitter. Contrasted and customary bibliometrics and webmetrics, altmetrics are prevalent in that they give quick, ongoing, open and straightforward reports on logical effect, and spread a broad non-scholastic group of spectators and expanded research discoveries and sources [3]. Internet based life stages contain a great deal of remark messages about logical articles. We should better break down them through measurable examination, slant investigation, content order and grouping, and AI to acquire understood, obscure helpful data from them, and subsequently better help logical research and revelation. In this paper, we led content mining on the audits of articles via web-based networking media, trying to follow the focal point of survey and the heading of general supposition reflected in news reports.

II. RELATIVE WORKS AND DATASETS

Content mining includes an immense field of hypothetical methodologies and techniques with one thing in like manner: message as information data. This permits different definitions, going from an augmentation of old style information mining to writings to increasingly complex plans like "the utilization of huge online content assortments to find new actualities and patterns about the world itself" [4]. When all is said in done, content mining is an interdisciplinary field of action among information mining, semantics, computational measurements, and software engineering. Standard procedures are content arrangement, content bunching, philosophy and scientific classification creation, archive rundown and dormant corpus investigation. Moreover a great deal of methods from related fields like data recovery are generally utilized.

The advantage of content mining accompanies the huge measure of important data inactive in writings which isn't accessible in old style organized information groups for different reasons: content has consistently been the default method for putting away data for many years, and for the most part time, individual and cost requirement deny us from bringing writings into well-organized arrangements (like information edges or tables).

The issue of content mining is of significance to distributers who hold enormous databases of data requiring ordering for recovery. This is particularly valid in logical controls, in which exceptionally explicit data is regularly contained inside composed content. In this way, activities have been accepted, for example, Nature's proposition for an Open Text Mining Interface (OTMI) and the National Institutes of Health's basic Journal Publishing Document Type Definition (DTD) that would give semantic signs to machines to answer explicit questions contained

inside content without expelling distributer obstructions to free.

The programmed examination of immense literary corpora has made the likelihood for researchers to investigation a huge number of archives in different dialects with extremely constrained manual mediation. Key empowering advancements have been parsing, machine interpretation, theme arrangement, and AI.

The programmed parsing of literary corpora has empowered the extraction of entertainers and their social systems on a tremendous scale, transforming literary information into arrange information. The subsequent systems, which can contain a huge number of hubs, are then broke down by utilizing devices from arrange hypothesis to distinguish the key on-screen characters, the key networks or gatherings, and general properties, for example, heartiness or basic security of the general system, or centrality of specific hubs [5]. This robotizes the methodology presented by quantitative account examination [6], whereby subject-action word object triplets are related to sets of on-screen characters connected by an activity, or sets framed by on-screen character object [7].

Content investigation has been a customary piece of sociologies and media reads for quite a while. The mechanization of substance investigation has permitted a "major information" upheaval to happen in that field, with contemplates in online networking and paper content that incorporate a huge number of news things. Sex inclination, comprehensibility, content closeness, peruser inclinations, and even state of mind have been broke down dependent on content mining strategies more than a large number of archives [8-11]. The examination of intelligibility, sexual orientation inclination and theme predisposition was exhibited in Flaounas et al. [12] indicating how various points have distinctive sexual orientation inclinations and levels of intelligibility; the likelihood to recognize temperament moves in a huge populace by breaking down Twitter content was exhibited too [13].

In this paper, we picked the 100 most noteworthy - score articles in 2016 on Altmetrics.com, downloaded the datasets (December 7, 2016) through the connection (https://figshare.com/collections/Altmetric_Top_100_2016/3590951).

III. Techniques

First we delivered a plain book document "Top100.txt" which incorporates the rundowns of all the 100 articles. At that point we chose the most elevated score article "US Health Care Reform: Progress to Date and Next Steps" in 2016 and delivered a content document dependent on prevailing press remarks on it gave by Altmertics.com. As needs be, we arranged two

plain content documents (one for the entire, and one for parts) for later content mining.

We utilized the R Studio form 3.3.3, including its factual condition and the accompanying bundles: tm, dplyr, wordcloud2, and so on we executed literary investigation of remark messages by contemplating the entire first and afterward narrowing the examination degree to concentrate on some of them to acquire imagined word mists and determined remarks.

IV. RESULTS AND ANALYSIS

In ceaseless dispersal via web-based networking media, logical articles leave advanced records as well as pull in a large group of remark messages on news outlets, blog and Twitter, and so forth.

These writings are significant, uncommon wellspring of solid help for assessing the effect of logical articles. We directed a printed investigation dependent on the rundown document of the 100 articles contained in the datasets and the news report record of one specific article among them. To start with, we entered the writings and the outline record of the 100 articles into the framework. Second, we pre-prepared the writings, for example, erasing spaces, changing over them into lowercase, erasing accentuation stamps and words that are never again being used. Third, we determined the word recurrence. At long last, we sent out the pictured word mists as indicated by the word recurrence. We utilized R language to program and the R content as pursues:

```
1 library(wordcloud2)
2 library(dplyr)#data getting and cleaning
3 library(tm)
4 ##data cleaning, delete the blanks and punctuations
5 filePath<- "D:/R/top100wordcloud.txt"
6 text = readLines(filePath)
7 txt = text[text!=""]
8 txt = tolower(txt)
9 txt <- removeWords(txt,stopwords('english'))
10 txtList = lapply(txt, strsplit," ")
11 txtChar = unlist(txtList)
12 txtChar = gsub("\\.|,|\\!|:|;|\\?","",txtChar)
13 txtChar = txtChar[txtChar!=""]
14 data = as.data.frame(table(txtChar))
15 colnames(data) = c("Word","freq")
16 ordFreq = data[order(data$freq,decreasing=T),]
17 wordcloud2(ordFreq, size = 0.5,shape = 'star')
```

Consequently, from the datasets we separated 1,447 words and the seven most much of the time utilized words are recorded in Table. 1.

Table 1: high Frequency words

Words	Frequency (%)
Human	17
Cancer	13
Virus	12
Zika	12
Association	10
New	10
Life	9

Fig. 1. Visualized word cloud of comments on Top 100 articles.

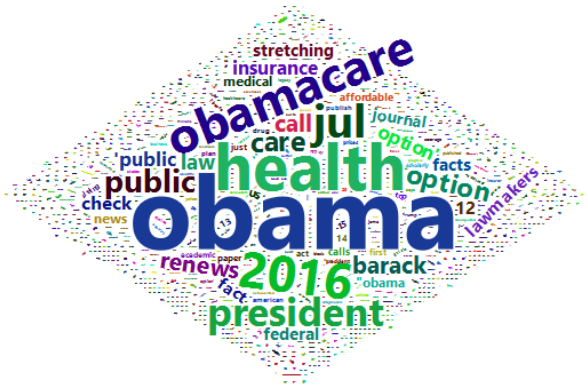

Fig. 2. Visualized word cloud of news review bout one paper.

The words in the informational collection were shown as word cloud as indicated by word recurrence. From Fig. 1 we can see that in 2016, individuals were increasingly keen on the investigations of people, specifically in the investigations of malignant growths and the Zika infection that cleared crosswise over Africa. From the every now and again utilized word "affiliation", we found that the greater part of the exploration was interdisciplinary, demonstrating the covering and combination of logical research. Plus, the examination is "New", which means

that analysts embrace new techniques, new viewpoints and new approaches for spearheading research.

What's more, one paper in the datasets "US Health Care Reform: Progress to Date and Next Steps" has gotten ceaseless media consideration since its distribution. We slithered a sum of 31 titles of news provides details regarding it and built up the pictured word cloud by utilizing a similar technique. Fig. 2 gives that the normal subject of these news reports shows that "previous US president Obama turned out Obama care in July 2016".

V. CONCLUSION

Bormmann [14] thought about that future research should concentrate more on the estimation of the broad effect of the exploration, not on the examination of altmetrics and conventional measurements. As per Davis et al. [15], content mining innovation ought to be applied to follow roundabout references of printed substance of research discoveries, especially in web journals, news reports and government records. We directed content mining on the article outline document of the

datasets and found the focal point of consideration in logical research from the open viewpoint and another way to deal with the general collaboration in logical research in 2016. Content mining was likewise performed on titles of news provides details regarding one specific article. Media remarks about the article were envisioned by word cloud. Misleadingly straightforward, content mining discloses to us what the numbers recorded by altmetrics can't tell. The pictured word cloud likewise makes the outcome increasingly direct and straightforward.

Altmetrics give us a remarkable social point of view to investigate the effect of scholastic research discoveries and follow scholarly correspondence among perusers. There is a large group of datasets to help the investigations in scholastic long range informal communication practices and even in the cooperation between various measurements [16]. What's more, representation of scholarly trade and network found at the online life level is another significant research subject [17].

Web based life stages contain a ton of remark messages about logical articles. We should better break down them through measurable investigation, assessment examination, content characterization and bunching, and AI to acquire understood, obscure helpful data from them, and in this way better help logical research and disclosure.

REFERENCES

[1] K. Weller, "Social media and altmetrics: an overview of current alternative approaches to measuring scholarly impact," in Incentives and Performance. Cham: Springer International Publishing, 2015.

[2] J. Priem, T. Taraaborelli, P. Groth, and Neylon, "Altmetrics: a manifesto," 2010 [Internet], Available: http://altmetrics.org/manifesto/.

[3] P. Wouters and R. Costas, "Users, narcissism and control: tracking the impact of scholarly publications in the 21st century," 2012 [Internet], Available: http://apo.org.au/node/28603.

[4] M. A. Hearst, "Untangling text data mining," in Proceeding of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL), College Park, MD, pp. 3–10, 1999.

[5] S. Sudhahar, G. De Fazio, R. Franzosi, N. Cristianini, "Network analysis of narrative content in large corpora," Natural Language Engineering, vol. 21, no. 1, pp. 81-112, 2015.

[6] R. Franzosi, "Quantitative narrative analysis," Journal of Bacteriology, vol. 191, no. 7, pp. 2388-2391, 2016.

[7] S. Sudhahar, GA. Veltri, and N. Cristianini, "Automated analysis of the US presidential elections using big data and network analysis," Big Data & Society, vol. 2, no. 1, pp. 1-28, 2015.

[8] I. Flaounas, M. Turchi, O. Ali, N. Fyson, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini, "The

structure of EU Mediasphere," PLoS ONE, vol. 5, no. 12, pp. e14243, 2010.

[9] V. Lampos and N. Cristianini, "Nowcasting events from the social web with statistical learning," ACM Transactions on Intelligent Systems and Technology, vol. 3, no. 4, pp. 1-22, 2012.

[10] I. Flaounas, O. Ali, M. Turchi, T. Snowsill, F. Nicart, and T. De Bie, "NOAM: news outlets analysis and monitoring system," in Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, Athens, Greece, pp. 1275-1277, 2011.

[11] N. Cristianini, "Automatic discovery of patterns in media content," in Combinatorial Pattern Matching. Cham: Springer International Publishing, pp. 2-13, 2011.

[12] I. Flaounas, O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini, "Research methods in the age of digital journalism," Digital Journalism, vol. 1, no. 1, pp. 102-116, 2013.

[13] T. Lansdall-Welfare, V. Lampos, and N. Cristianini, "Effects of the recession on public mood in the UK," in Proceedings of International Conference on World Wide Web, Lyon, France, pp. 1221-1226, 2012.

[14] L. Bornmann, "Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics," Journal of Informetrics, vol. 8, no. 4, pp. 895-903, 2014.

[15] B. Davis, I. Hulpuş, M. Taylor, and C. Hayes, "Challenges and opportunities for detecting and measuring diffusion of scientific impact across heterogeneous altmetric sources," 2015 [Internet], Available: http://altmetrics.org/wp-content/uploads/2015/09/altmetrics 15_ paper_21.pdf.

[16] M. Taylor, "Exploring the boundaries: how altmetrics can expand our vision of scholarly communication and social impact," Information Standards Quarterly, vol. 25, no. 2, pp. 27-32, 2013.

[17] C. P. Hoffmann, C. Lutz, and M. Meckel, "A relational altmetric? Network centrality on ResearchGate as an indicator of scientific impact," Journal of the Association for Information Science and Technology, vol. 67, no. 4, pp. 765-775, 2015.