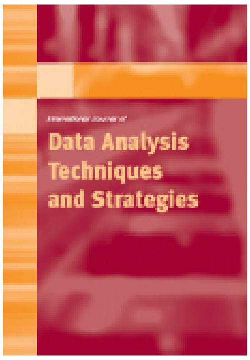


International Journal of Data Analysis Techniques and Strategies

ISSN online: 1755-8069 - ISSN print: 1755-8050
https://www.inderscience.com/ijdats

Using free open-source tools for text visualisation over unstructured corpus effectively

Gowri R. Choudhary, Iti Sharma

DOI: 10.1504/IJDATS.2023.10054976

Article History:

Received:	09 June 2022
Last revised:	06 February 2023
Accepted:	11 February 2023
Published online:	28 July 2023

Using free open-source tools for text visualisation over unstructured corpus effectively

Gowri R. Choudhary*

Computer Science Department,
Career Point University,
National Highway 52, Opp Alaniya Mata Ji Mandir,
Alaniya, Kota, Rajasthan 325003, India
Email: rgk.choudhary@gmail.com
*Corresponding author

Iti Sharma

Computer Science Department,
Government Polytechnic College,
Kota, Rajasthan, India
Email: itisharma.uce@gmail.com

Abstract: Text visualisation is an essential analytical task in many applications. Researchers from such fields can utilise the graphical output but need tools for its generation due to lack of expertise. Though several open-source tools are available, there is a challenge in choosing a suitable tool and preparing a compatible input because the corpora are often self-collected and unstructured. This paper describes ten popular open-source text visualisation tools for word-clouds. It is observed here that these tools take only a single text file as input while unstructured corpora have multiple-file format. So further in this paper a priority window technique is proposed to convert corpus into a small single text file that retains the characteristics similar to a standard model found in structured corpora. Experiments are performed over a self-collected corpus to demonstrate a real-life scenario from journalism application. The output word clouds show the effectiveness of proposed frequency-based technique. This technique is aimed at ease of use by users that lack expertise of data mining.

Keywords: text visualisation; summarisation; text mining; word-cloud; open-source visualisation tools; term weighting; feature weighting.

Reference to this paper should be made as follows: Choudhary, G.R. and Sharma, I. (2023) 'Using free open-source tools for text visualisation over unstructured corpus effectively', *Int. J. Data Analysis Techniques and Strategies*, Vol. 15, Nos. 1/2, pp.129–146.

Biographical notes: Gowri R. Choudhary is a student pursuing her PhD in Computer Science Engineering at the Career Point University, Kota. Her research interest areas are text visualisation and clustering which comes under the area of text mining. She had completed her post-graduation from the Gujarat Technical University, Gujarat in Computer Science Engineering field.

Iti Sharma is working as an Assistant Professor in the Computer Science Department at Government Polytechnic College, Kota, India. She has completed her PhD in Computer Science in 2018. She has several research publications. Her areas of interest are WSNs, security, data analytics and text mining.

1 Introduction

Text is the most comprehensible form for human beings. Humans developed the script for communication because people communicate majorly in the text form not in the numerical form. Thus, text being the major part of human communication, comprehension and interaction, a copious amount of data is available in textual form like emails, webpages, newsfeeds, articles, stories, etc. All these texts have different patterns inside them and when a professional derives those patterns through automated process, it is termed as text analytics (Sarkar, 2019). According to Bhoslay and Bali (2021), text analytics enable the user to convert the text into information so that the major five tasks can be achieved to benefit the society, science and business as follows:

1 information extraction (Rai, 2019)

2 information retrieval

3 clustering/categorisation

4 summarisation.

Here, we focus on summarisation tasks that make a large body of text understandable in less time. Summarisation (Syed et al., 2021; Ying et al., 2021) is a task of condensing large amount of text data into its most important parts such that information loss is minimal. Bhargav et al. (2021) categorise summarisation approaches into two major forms: text-to-text and text-to-graphical. Text-to-text form (Mandal, 2021; Iskender et al., 2021) is also called text summarisation implies shortening up long text by recognising and collecting the important points without altering the meaning of the text. This requires expertise and reader is referred to Awasthi et al. (2021) and Widyassari et al. (2020) for description of various techniques. Text-to-graphical form is extracting the important words from a large amount of text and arranging into a graphical form (Yadav et al., 2021). This is also called text visualisation. Text visualisation is a collective term for methods used to convert results of text analysis in a visual form. Gan et al. (2014) and Elmqvist et al. (2014) have described it as transforming the text into a visual by considering the words, sentences and their relationships to make the user understand better and reduce mental workload of facing massive text. Due to involvement of graphics, text visualisation is achieved automatically through visualisation tools. Kucher and Kerren (2014, 2019) have presented a survey on text visualisation techniques through an online visualisation browser. Similarly, SoSVis (Alharbi and Laramee, 2018) where a user can compare visualisation tools by experimenting, though the tools are not available for free use. Majority of users need visual analysis demand freely available tools for their one-time needs. Other aspects that concern the users are the form of input to be given to tool and output desired by their application at hand.

Text visualisation output can be like tagcloud (Hearst and Rosner, 2008; Jänicke et al., 2018), word-cloud (Kulahcioglu and Melo, 2019; Lee et al., 2010; Cui et al., 2010), graph (Havre et al., 2002; Viegas et al., 2009), graph-of-words (Antoine et al., 2016), chart (Koh et al., 2010), map (Kucher et al., 2018), text data stream (Wanner et al., 2014), social networks (Preim et al., 2013) and others like timeline, tree-map (Wattenberg, 2006), head-map, and spark-line. As compared to outputs like histograms that need mathematical interpretations, visuals that are capturing and attractive to the human eye like word-cloud are preferred. Word-clouds depict words of input text arranged in space varied in size, colour, and position based on word frequency, categorisation, or significance (Heimerl et al., 2014; Vilaplana and Montoro, 2014). The words are highlighted through font and colour as per their significance.

Hong and Park (2019) identify that form of input affects visualisation approach as:

1 single-text approach

2 collection-of-text approach.

Moreover, the visualisation tools take input as either:

1 direct-text

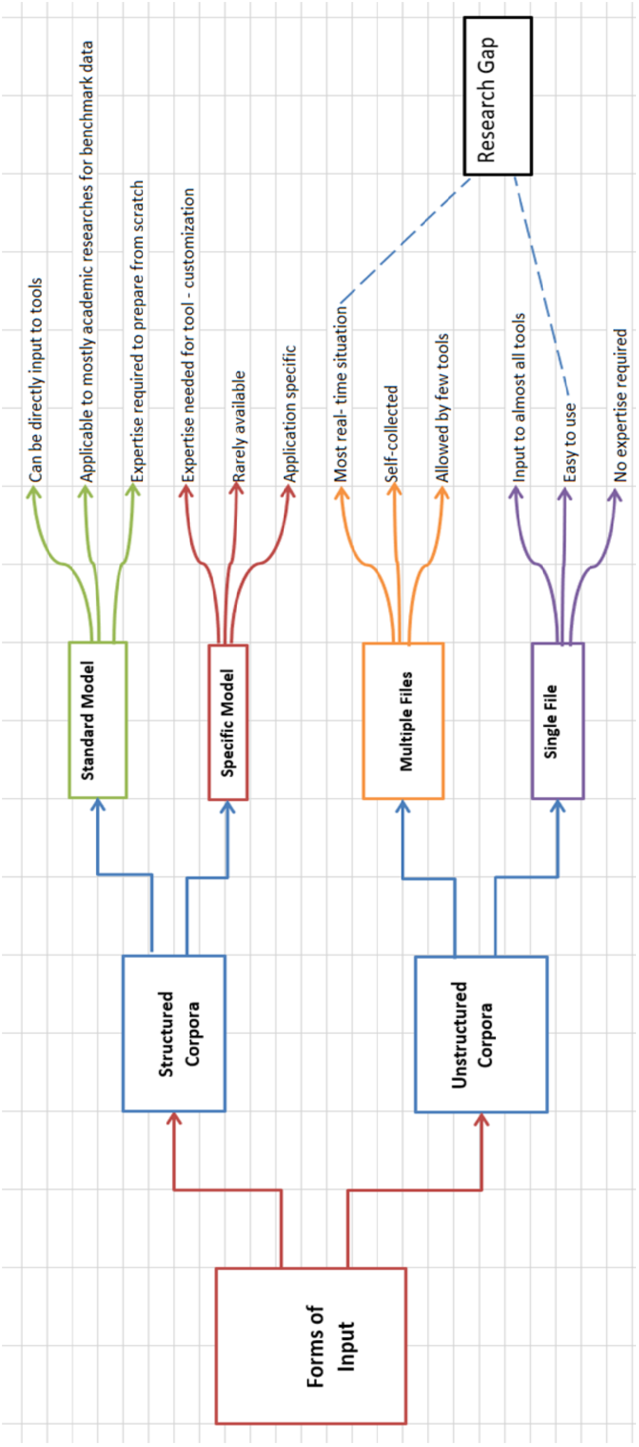
2 representation-of-text.

The direct input is like copying a text file into the tool interface. Or a tool is designed to take input only a standard representation of the corpus based on a language model like frequency table. While direct text input is easy for many users as it saves time, it is often word-limited, that is it has a single-text approach. Very few tools have a collection-of-text approach. Tools that need a representation prefer standard models and may not be suitable for specifically designed corpus. Here, we digress to explain that corpora are either structured or unstructured. When extra information like part-of-speech tags, semantics, and pragmatics is involved into the corpus, it is called a structured corpus. Such corpora are available for benchmarking academic researches. Real-life applications have self-collected corpora that do not have any annotations and are called unstructured. Figure 1 outlines what different inputs are available for visualisation tools. Our research is aimed at the broad gap between the easy-to-use single-text tools and available unstructured corpora. The challenge is of converting an unstructured corpus into a single text file that retains its text data characteristics.

The applications that may benefit from our research are those where users have their own text data but lack expertise to build a language model out of it. Like material science (Kononova et al., 2021), healthcare (Elbatta et al., 2021), social media (Chen et al., 2017; Kabir et al., 2018; Wu et al., 2016), services management (Kumar et al., 2021), consumer reviews (Alper et al., 2011), theme-crowds (Archambault et al., 2011), and business performance analysis (Hong and Park, 2019). These users rely on freely available tools to suffice their irregular needs of text visualisation.

In this paper, we propose a scheme to convert a corpus into small single text file without affecting the effectiveness of its visualisation. The rest of paper first describes few popular and easy-to-use visualisation tools, and then discusses proposed ‘priority-window’ techniques. Word-clouds are shown to visually compare the output.

Figure 1 Various forms of input for text visualisation tools and their characteristics (see online version for colours)


Forms of Input
Structured Corpora
Unstructured Corpora
Standard Model
Specific Model
Multiple Files
Single File
Can be directly input to tools
Applicable to mostly academic researches for benchmark data
Expertise required to prepare from scratch
Expertise needed for tool - customization
Rarely available
Application specific
Most real- time situation
Self-collected
Allowed by few tools
Input to almost all tools
Easy to use
No expertise required
Research Gap

2 Tools description

This section briefly introduces some popular open-source free text visualisation tools, especially for word-clouds, with their salient features and major drawbacks.

2.1 WordArt/Tagul

WordArt is a good tool which provides several settings such as text colour, shape of the word-cloud, size, and density of the word, fonts and more. Input is to be provided as a list of keywords. It has a much fancier look than most of the other tools. The main drawback of this tool is it neither allows single text file upload nor a corpus.

2.2 Tagcrowd

An easy-to-use tool that allows only pasting of text or single file upload. There are no options for changing shape, colour, etc. The words are shown in alphabetical order with variation in size and thickness for emphasis.

2.3 Word-it-out

A simple tool with preset design options for word-clouds. The major drawbacks of this tool are:

1 no option of file upload

2 limited design options

3 style and fonts are not up to the mark.

2.4 Voyant

It is versatile tool that generates:

1 Cirrus: it is a 'cloud' generated from the most frequent words.

2 Bubble lines: a word frequency graph throughout the text.

3 Text arc: for word distribution and their interconnection.

Besides pasting single text, user can upload single text file as URL, pdf, or MSWord format or entire corpus.

2.5 WordSift

WordSift is collection of text analysis tools. Major features are word-cloud and visual Thesaurus. It allows only pasting text, with recommended limit up to 10,000 words for analysis. The main advantage is that we can control the scale of the words, density, and orientation.

2.6 Wordclouds

Wordclouds is simple and much similar to the WordArt tool. The set of shapes for word-clouds are as per fields like traffic, love, pets, etc. It has many settings for font, style, colour palettes, size of cloud, invert, and masking options. The main drawback is:

1 slow speed

2 single text input.

2.7 Jasondavies word-cloud

This tool is one of the best online word-cloud generators because it provides funny and exciting shapes. It allows only pasting text input. Another drawback is that we cannot change the settings like shapes and colours.

2.8 Daniel Soper's word-cloud generator

It is a free tool and easy to use because of its simple interface. The output cloud can be customised using the options panel. The drawbacks are:

1 no option of uploading file

2 not many options for shape, colour, and design.

2.9 Lexos

Lexos is a web-based platform tool which has great resources for visualising large text. This site allows uploading the entire corpus, i.e., multiple files. Then prepare the data to visualise and analyse it. The output can be word-clouds, multi-cloud, bubbleviz, rolling-window graph and analysis like statistical analysis, clustering, similarity query, and top word. The drawback is that we have to login every time before using the tool.

2.10 Vizzlo

This tool has more potential than any other word-cloud generator because it has many paywalls. It allows the user to upgrade, change shapes, front and can fix the maximum numbers of words. One of the main features of this tool is used for removing the watermark from the word-cloud. If you can pay upfront, this tool allows you to change more settings and download files in PNG form with a transparent background.

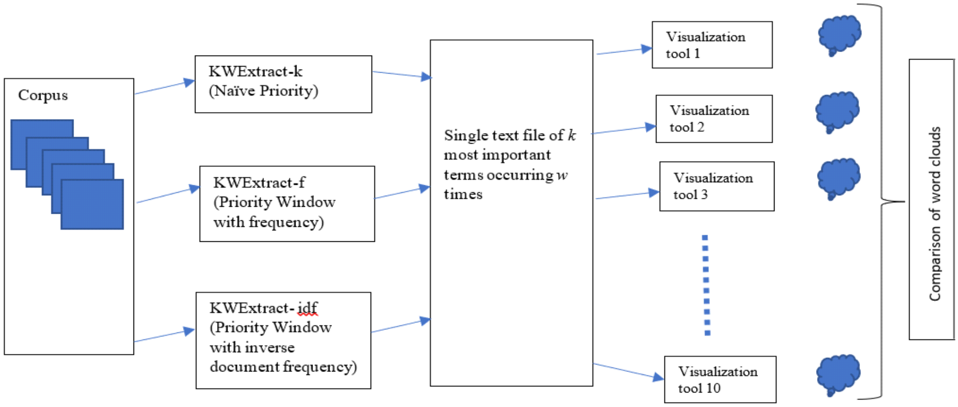
3 Problem statement

The main challenge is that most of the tools restrict the input to be a single text file of limited words, while the user is interested in visualising entire corpus. Their corpus is often unstructured and their computing expertise limited. We present how simple techniques can be used to convert corpus into a single text file of desired word length.

4 Proposed schematic

A Naïve approach to convert corpus into single-text is concatenate all the files of corpus. The obvious drawback is the size of file, which already is a restriction in most of open-source tools. Also, the number 'n' itself may be too large to make a simple concatenation a memory-intensive operation. Hence, we propose a schema as illustrated in Figure 2.

Figure 2 Flow of the proposed schema (see online version for colours)


Corpus
KWExtract-k (Naïve Priority)
KWExtract-f (Priority Window with frequency)
KWExtract- idf (Priority Window with inverse document frequency)
Single text file of k most important terms occurring w times
Visualization tool 1
Visualization tool 2
Visualization tool 3
Visualization tool 10
Comparison of word clouds

The process consists of processing a corpus through priority window technique to prepare a single text file of certain length and then in putting it to a tool for visualisation. Here we discuss in detail the proposed priority window approaches.

There are three basic approaches already exists in text mining where a word is taken has its face value, a term is taken proportional to the frequency, and a term is taken proportional to the tf-idf. So, these three approaches have been listed here with one concept that we have introduce as priority window. The approaches are:

1 selecting something from the bag which is priority window

2 selecting something from the bag proportional to the frequency

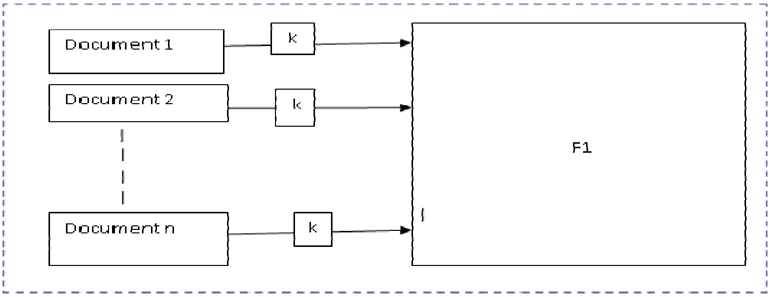
3 selecting something from the bag which is proportional to its tf-idf.

4.1 Priority window (KWExtract-k)

The first approach is the priority window (pw) approach that aims at collecting most important words from each file and the number of words (say k) is kept fixed and same for each file. Thus, for a corpus of n documents, extract k most important keywords from every file and concatenate into a single file of size kn. The process is shown in Figure 3, the single output file referred as F_1. Here the glitch is that suppose a particular term occurs in first document of the priority window may not be in the priority window of other document but occurring in the document then that term will get eliminated. The terms which are common to all documents and most occurring in all documents will get higher priority and words which rarely occur will not get priority in the entire corpus. Thus, this technique eliminates document bias for those terms that have a very high frequency in very few or only one long document. Therefore, only those words which are common to all documents and mostly occur in all documents will gain the priority, hence named as priority window. A possible drawback is that frequency of term is considered only within the document context and not in the corpus context. The formula is

$$F_1 = concentrate(pw) \tag{1}$$

Figure 3 Schematic representation of priority window (KWExtract-k) approach (see online version for colours)



4.2 Priority window with frequency (KWExtract-f)

The above approach has a drawback that every priority word appears in file F_1 only once per document. To prepare a file that reflects the frequency characteristics of entire corpus, we now propose to repeat every term as many times as their frequency in that particular document. Each document contributes towards the single text file k different words written many times. In Figure 4, this process is shown where f indicates frequency of term in that document. In the final single text file F_2, the frequency of each word is going to be different. We can observe in this approach that the frequency of the term in the corpus is not reflecting but the frequency of the term in the particular document is visible. So, the drawback cannot be clearly stated but a loophole can be identified in this approach where a term may be just out of the priority window of certain article and that frequency may not be contribute in the final single text file frequency of that particular term. Thus, it is possible in certain cases that a document having a particular term of very high frequency is able to affect the frequency of that term in final text file. The formula is:

$$F_2 = concat[pw * frequency] \tag{2}$$

4.3 Priority-window with inverse document frequency (KWExtract-idf)

This approach is based on the inferences that have been already done in text mining in so many years. Several researchers agree that using inverse-document-frequency (idf) eliminates the document bias and also gives weightage to the terms which hold the importance in all over the corpus instead of in a single article. Consider an approach where k-sized priority window is constructed and each term is repeated as per its *idf* value to obtain single text file F3 (shown in Figure 5). The *idf* is computed using

$$idf = \log\left[\frac{N}{1+df}\right] \tag{3}$$

where *df* is the document frequency. Therefore, the priority window with *idf* equation (3) can be defined as:

$$F_3 = concat[pw * idf] \tag{4}$$

Figure 4 Schematic representation of priority window with frequency (KWExtract-f) approach (see online version for colours)

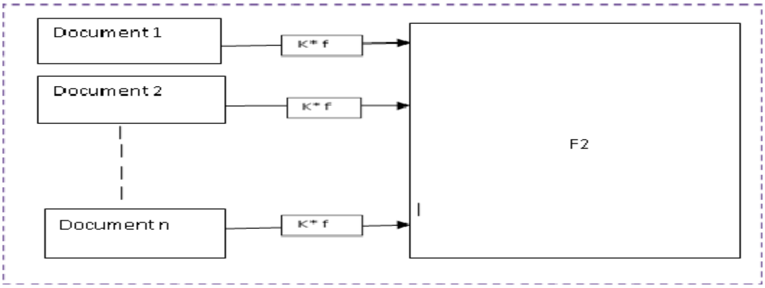
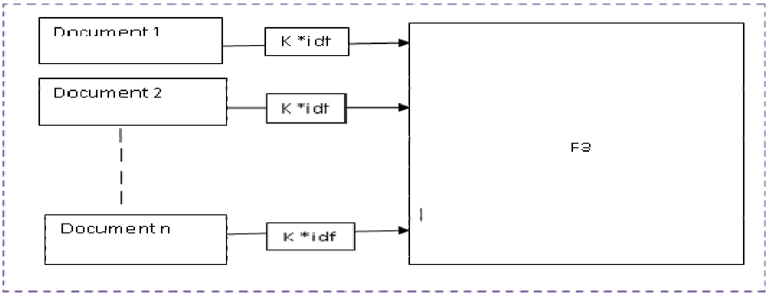


Figure 5 Schematic representation of priority window with inverse document frequency (KWExtract-idf) approach (see online version for colours)



4.3.1 Note on choosing value of k

Value of *k* needs to be set for tools having restriction on input-size. While it is direct for KWExtract-k approach by setting $kn \leq Limit$, for KWExtract-f approach we can only have bound as $knf_{avg} \approx Limit$ where is f_{avg} average term-frequency in the document. In our experiments we have observed that accepting at most 8 or 10 words from each document in the corpus is sufficient.

5 Experimental evaluation

5.1 Preparation of corpus

In order to see how the proposed techniques will perform in a real-life situation like analysing news stories related to same topic occurring in different sources, we use a self-collected corpus for experiments. News articles related to Ayodhya Ram Mandir (ARM) issue available online on websites of newspapers like Hindustan Time, The Times of India, etc. are procured. The date of started preparing our corpus was from 21st September 2019 to 30th November 2019. We have extracted keywords from each text document individually keeping size of priority window as 10. A text file is prepared by concatenating all these 1,055 keywords of ARM corpus and input to the visualisation tools to represent the naïve approach. Similarly, the document frequency, inverse document frequency and term frequency of terms occurring in the priority window were recorded and a single text file was constructed using above mentioned approaches.

Table 1 Wordcloud obtained by using KWExtract-k, KWExtract-f and KWExtract-idf approaches from different open-source text visualisation tools, (A) WordArt, (B) Tagcrowd, (C) WordItOut, (D) Voyant, (E) Wordsift, (F) WordCloud, (G) Jasondavies, (H) Daniel Soper's Word cloud, (I) Lexos, (J) Vizzol (see online version for colours)

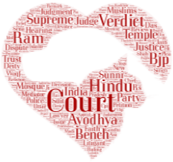
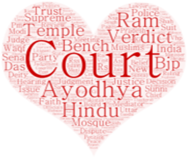
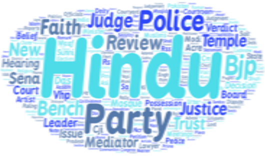
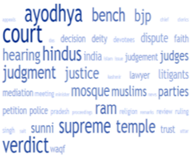
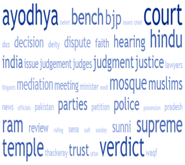
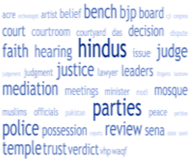
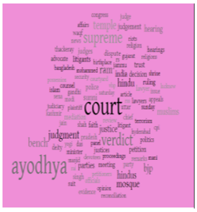
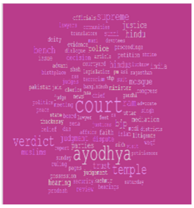
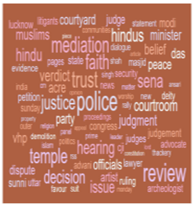
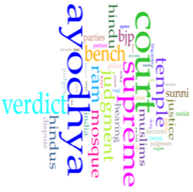
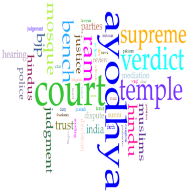
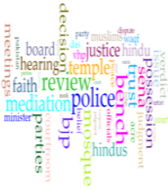
Tools	*KWExtract-k*	*KWExtract-f*	*KWExtract-idf*
A			
B			
C			
D			

Table 1 Wordcloud obtained by using KWExtract-k, KWExtract-f and KWExtract-idf approaches from different open-source text visualisation tools, (a) WordArt, (b) Tagcrowd, (c) WordItOut, (d) Voyant, (e) Wordsift, (f) WordCloud, (g) Jasondavies, (h) Daniel Soper's Word cloud, (i) Lexos, (j) Vizzol (continued) (see online version for colours)

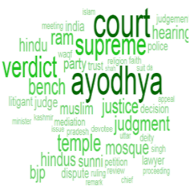
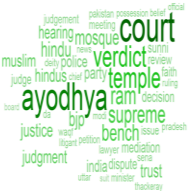
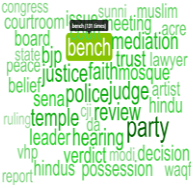
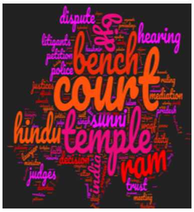
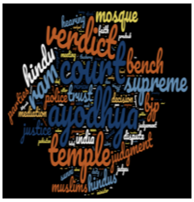
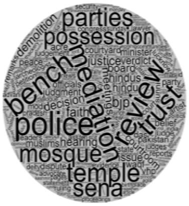
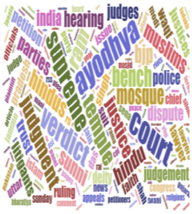
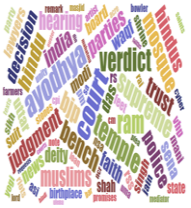
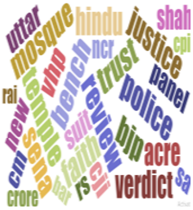
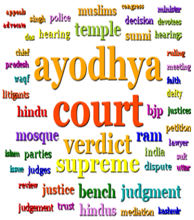
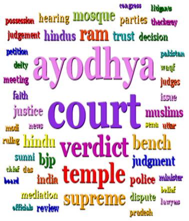
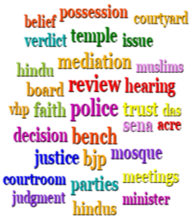
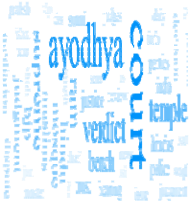
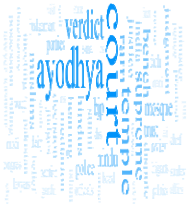
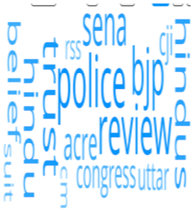
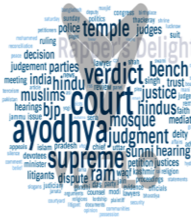
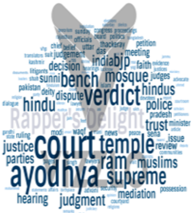
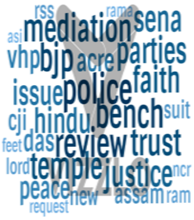
Tools	*KWExtract-k*	*KWExtract-f*	*KWExtract-idf*
E			
F			
G			
H			
I			
J			

Figure 6 Priority window (KWExtract-k) (see online version for colours)

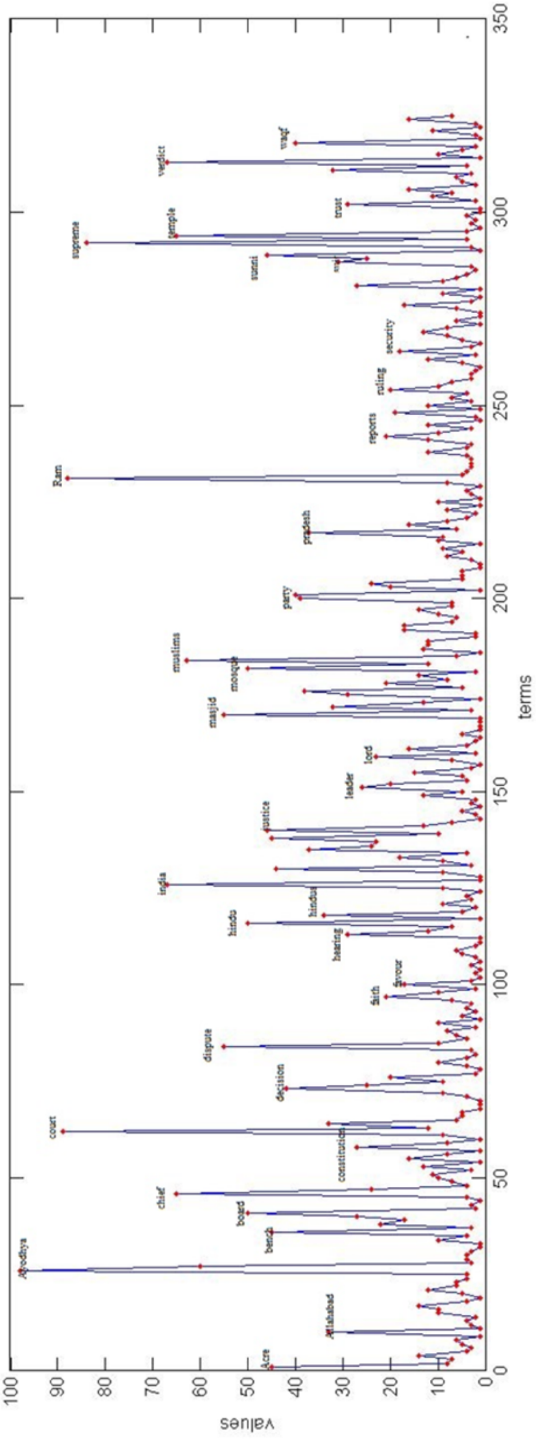


Figure 7 Priority window with frequency (KWExtract-f) (see online version for colours)

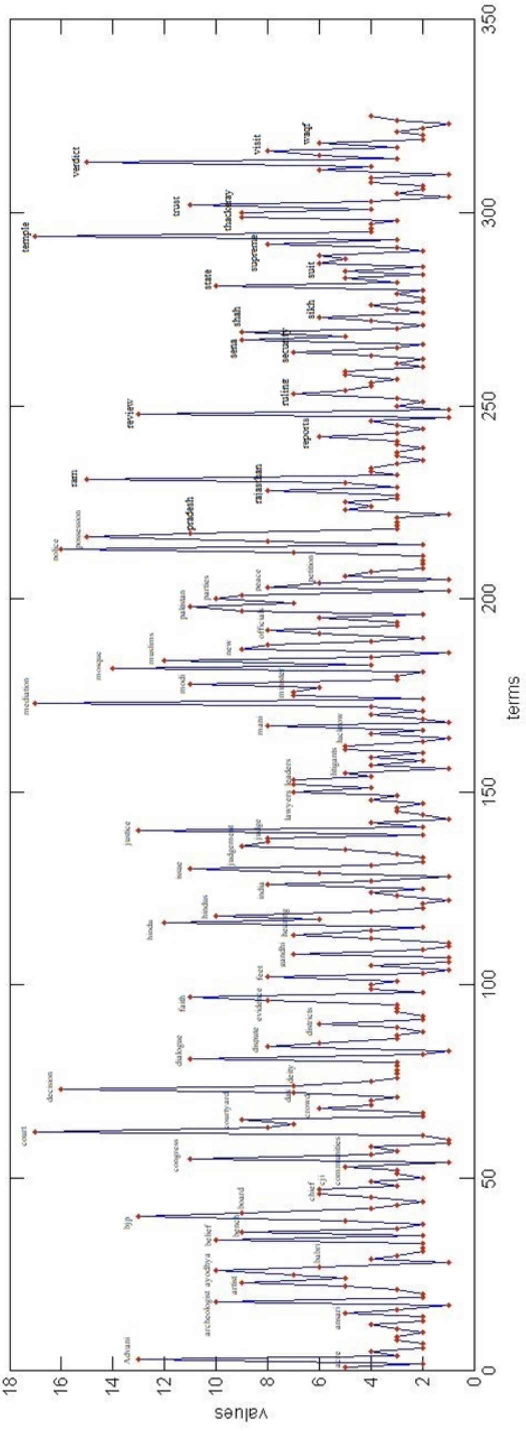
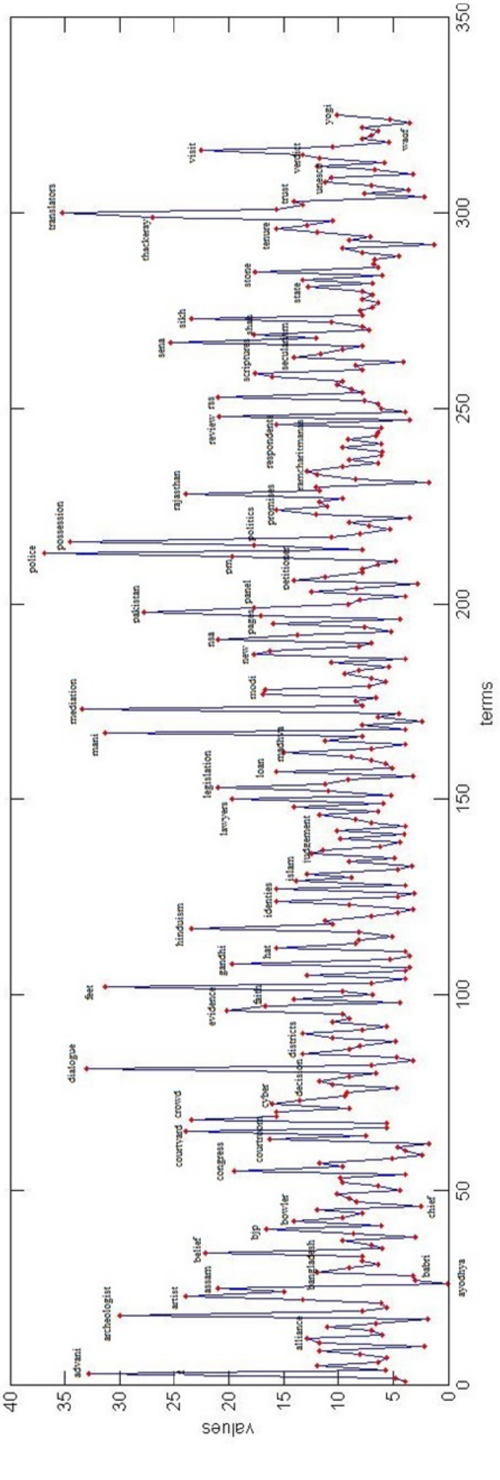


Figure 8 Priority window with inverse document frequency (KWExtract-idf) (see online version for colours)



5.2 Results

The word-clouds obtained by inputting the single text files obtained from a naïve and the proposed approaches are shown in Table 1.

5.2.1 Observation of word-clouds

While visualising the obtained word-clouds using all the three methods, we have identified many differences in words importance in the tools. Manual inspection of the word-clouds reveals that more topic-relevant terms are highlighted in output obtained through KWExtract-f approach than KWExtract-idf approach. Moreover, in wordclouds of KWExtract-f approach words of lesser priority are completely vanished. It has reduced the visual noise and emphasised the important words. So, the advantage of using a frequency-based approach is very clear. The effect of our proposed schema KWExtract-f reduces the words which are not important.

Also, if we want to see particularly which terms have been emphasised and how the priority window approach affected the visualisation of corpus, then we can plot the terms and their final frequency in the single output file. Figure 6 shows the terms on x-axis and their frequency in output file on y-axis for a naïve KWExtract-k approach. Figure 7 plots the same for KWExtract-f approach and the terms on peaks here are different from the peaks of Figure 6. This indicates that KWExtract-f has affected the importance of certain terms by increasing their frequency in output file as compared their frequency in the corpus. These terms are those that have high frequency in many documents. The terms that have been suppressed in Figure 7 as compared to Figure 6 are those that had very high frequency in few documents. Figure 8 is plot for KWExtract-idf approach where even common peaks of Figures 6 and 7 have been suppressed. Hence it is not recommended. Finally, we can conclude that using a priority window based on frequency is able to identify words/terms of evenly distributed high frequency for a good summarised visualisation.

6 Conclusions

Text visualisation is need of several non-technical fields too where researchers and users lack either knowledge or tools to produce effective visualisation. Open-source tools are available to help people in such situations. Still there is a challenge to convert the corpus (that is often unstructured) into a single text file such that its data characteristics are retained, because majority of visualisation tools restrict size of input. This paper has proposed three 'priority window approaches' using frequency and inverse frequencies of words in corpus. Using the proposed methods corpus can be converted to a small text file of only important terms and hence visualisation produced is effective. For demonstration, we produce word-clouds from different tools using the priority window techniques and compare them. It is observed that a combination of word-importance and its frequency in individual documents gives effective output.

References

Alharbi, M. and Laramee, R.S. (2018) 'SoSTextVis: a survey of surveys on text visualization', *Proceedings in the Eurographics Association, EG UK Computer Graphics & Visual Computing*.

Alper, B., Yang, H., Haber, E. and Kandogan, E. (2011) 'OpinionBlocks: visualizing consumer reviews, conference', *IEEE VisWeek Workshop on Interactive Text Analytics for Decision Making*, October.

Antoine, J., Tixier, P., Skianis, K. and Vazirgiannis, M. (2016) 'GoWvis: a web application for graph-of-words-based text visualization and summarization', *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – System Demonstrations*, Berlin, Germany, 7–12 August, pp.151–156.

Archambault, D., Greene, D., Hannon, J., Cunningham, P. and Hurley, N. (2011) 'ThemeCrowds: multiresolution summaries of twitter usage', *Conference PaperSMUC'11*, October, DOI: 10.1145/2065023.2065041.

Awasthi, A., Gupta, K., Bhogal, P.S., Anand, S.S. and Soni, P.S. (2021) 'Natural language processing (NLP) based text summarization – a survey', *6th International Conference on Inventive Computation Technologies (ICICT)*, pp.1310–1317, DOI: 10.1109/ICICT50816.2021.9358703.

Bhargav, S., Choudhury, A., Kaushik, S., Shukla, R. and Dutt, V. (2021) 'A comparison study of abstractive and extractive methods for text summarization', *Advances in Intelligent Systems and Computing*.

Bhoslay, S.S. and Bali, M. (2021) 'Text analytics by business analytics specialization', *Book – Data Geek Text Analytics by Business Analytics Specialization School of Business and Management*, Vol. 3, No. 1, pp.181–196.

Chen, S., Yuan, L. and Yuan, X. (2017) 'Social media visual analytics', *Computer Graphics Forum*, June, Vol. 36, No. 3, pp.563–587, DOI:10.1111/cgf.13211

Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M.X. and Qu, H. (2010) *Word-cloud Visualization*, IEEE Computer Society, 0272-1716/10/$26.00 © IEEE.

Elbatta, M., Arnaud, E., Gignon, M. and Dequen, G. (2021) 'The role of text analytics in healthcare: a review of recent developments and applications', *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) –5: HEALTHINF*, pp.825–832, DOI: 10.5220/0010414508250832

Elmqvist, N., Hlawitschka, M. and Kennedy, J. (2014) 'Visualizing translation variation of Othello: a survey of text visualization and analysis tools', *Supplementary Material, Eurographics Conference on Visualization (EuroVis)*, Eurographics Association.

Gan, Q., Zhu, M., Li, M., Liang, T., Cao, Y. and Zhou, B. (2014) 'Document visualization: an overview of current research', *WIREs Computational Statistics*, Vol. 6, pp.19–36, DOI: 10.1002/wics.1285.

Havre, S., Hetzler, E., Whitney, P. and Nowell, L. (2002) 'ThemeRiver: visualizing thematic changes in large document collections', *IEEE Transactions on Visualization and Computer Graphics*, Vol. 8, No. 1, pp.9–20.

Hearst, M. and Rosner, D.K. (2008) 'Tag clouds: data analysis tool or social signaller?', in *Proceedings of the Hawaii International Conference on System Sciences*, pp.160–160.

Heimerl, F., Lohmann, S., Lange, S. and Ertl, T. (2014) 'Word-cloud explorer: text analytics based on word-clouds', *47th Hawaii International Conference on System Sciences*, pp.1833–1842, DOI: 10.1109/HICSS.2014.231.

Hong, J.W. and Park, S.B. (2019) 'The identification of marketing performance using text mining of airline review data', *Mobile Information Sy*stems, Article ID: 1790429, Vol. 2, No. 2, p.8, Hindaw, https://doi.org/10.1155/2019/1790429.

Iskender, N., Polzehl, T. and Moller, S. (2021) 'Reliability of human evaluation for text summarization: Lessons learned and challenges ahead', *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pp.86–96.

Jänicke, S., Blumenstein, J., Rücker, M., Zeckzer, D. and Scheuermann, G. (2018) 'TagPies: comparative visualization of textual data', *International Conference on Information Visualization Theory and Applications*, January, DOI: 10.5220/0006548000400051.

Kabir, A.I., Karim, R., Newaz, S. and Hossain, M.I. (2018) 'The power of social media analytics: text analytics based on sentiment analysis and word-clouds on R', *Informatica Economical*, April, DOI: 10.12948/issn14531305/22.1.2018.03

Koh, K., Lee, B., Kim, B. and Seo, J. (2010) 'ManiWordle: providing flexible control over wordle', *IEEE Transactions on Visualization and Computer Graphics*, Vol. 16, No. 6, pp.1190–1197.

Kononova, O., He, T., Huo, H., Trewartha, A., Olivetti, E.A. and Ceder, G. (2021) 'Opportunities and challenges of text mining in materials', *iScience*, 19 March, Vol. 24, No. 3, p.102155.

Kucher, K. and Kerren, A. (2014) 'Text visualization browser: a visual survey of text visualization techniques. Poster abstract', *IEEE Information Visualization (Infovis'14)*, Paris, France.

Kucher, K. and Kerren, A. (2019) 'Text visualization revisited: the state of the field in 2019', *Proceedings in the Eurographics Association.*

Kucher, K., Paradis, C. and Kerren, A. (2018) 'DoSVis: document stance visualization', *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP '18) – 3: IVAPP*, SciTePress, Funchal, Madeira – Portugal, pp.168–175.

Kulahcioglu, T. and Melo, G. (2019) 'Paralinguistic recommendations for affective word-clouds', *IUI '19*, ACM, Marina del Rey, CA, USA, 17–20 March, ISBN: 978-1-4503-6272-6/19/03.

Kumar, S., Kar, A.K. and Vigneswaran, L.P. (2021) 'Applications of text mining in services management: a systematic literature review', *International Journal of Information Management Data Insights*, Vol. 1, No. 1, p.100008, ISSN 2667-0968, https://doi.org/10.1016/j.jjimei.2021.100008.

Lee, B., Riche, N.H., Karlson, A.K. and Carpendale, S. (2010) 'SparkClouds: visualizing trends in tag clouds', *IEEE Transactions on Visualization and Computer Graphics*, Vol. 16, No. 6, pp.1182–1189.

Mandal, S.J. (2021*) Deep Learning Powered Text Summarization Framework for Creating a Highly Accurate Summary*, Whitepaper, Data-Matics Global Services Ltd.

Preim, B., Rheingans, P. and Theisel, H. (2013) 'Wordonoi: Visualizing the structure and textual contents of knowledge networks', *Eurographics Conference on Visualization (EuroVis)*, Vol. 32, No. 3.

Rai, A. (2019) *What is Text Mining: Techniques and Applications? 3rdi – 5 Common Techniques Used in Text Analysis Tools* [online] https://www.3rdisearch.com/5-common-techniques-used-in-text-analysis-tools (accessed 3 September 2012).

Sarkar, D. (2019) *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*, 2nd ed., O'Reilly Media Publication, Apress, ISBN-10: 1484243536.

Syed, S., Yousef, T., Al-Khatib, K., Jänicke, S. and Potthast, M. (2021) 'Summary explorer visualizing the state of the art in text summarization', *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp.185–194.

Viegas, F.B., Wattenberg, M. and Feinberg, J. (2009) 'Participatory visualization with Wordle', *IEEE Transactions on Visualization and Computer Graphics*, Vol. 15, No. 6, pp.1137–1144.

Vilaplana, J.N. and Montoro, M.P. (2014) 'How we draw texts: a review of approaches to text visualization and exploration',? *Elprofesional de la informaci'on*, Mayo-Junio, Vol. 23, No. 3, pp.221–235.

Wanner, F., Stoffel, A., Jäckle, D., Kwon, B.C, Weiler, A. and Keim, D.A. (2014) 'State-of-the-art report of visual analysis for event detection in text data streams', *Computer Graphics Forum*, Vol. 33, No. 3, pp.125–139.

Wattenberg, M. (2006) 'Visual exploration of multivariate graphs', in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp.811–819.

Widyassari, A.P., Rustad, S., Shidik, G.F., Noersasongko, E., Syukur, A., Affandy, A., Ignatius, D.R. and Setiadi, M. (2020) 'Review of automatic text summarization techniques & methods', *Journal of King Saud University – Computer and Information Sciences*, Vol. 34, No. 4, pp.1029–1046.

Wu, Y., Cao, N., Gotz, D., Tan, Y.P. and Keim, D.A. (2016) 'A Survey on visual analytics of social media data', *IEEE Transactions on Multimedia*, Vol. 18, No. 11, pp.2135–2148, https://dx.doi.org/10.1109/TMM.2016.2614220.

Yadav, A.K., Maurya, A.K. and Ranvijay, R.S. (2021) 'Extractive text summarization using recent approaches: a survey', *Ingénierie des Systèmesd'Information*, Vol. 26, No. 1, pp.109–121.

Ying, S., Zheng, Y. and Zou, W. (2021) 'LongSumm 2021: session based automatic summarization model for scientific document', *Proceedings of the Second Workshop on Scholarly Document Processing*, pp.97–102.