International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 5, Issue 11, pp 52-56, November-2021
https://doi.org/10.47001/IRJIET/2021.511009

Implemented Text Summarization Tool using Text Rank Algorithm

[1]Abhilasha More, [2]Vipul Dalal

[1,2]Vidyalankar Institute of Technology, Mumbai, India
Email IDs: [1]maurya.abhilasha@gmail.com, [2]vipul.dalal@vit.edu

***Abstract* - Text Synopsis is the most common way of producing the dense perspective on the text by choosing valuable and pertinent data from the first source records. It is a sub subject of Data Mining. Text synopsis is a procedure for understanding the point of any record, to picture huge text archive inside brief term. Synopsis gives adaptability and accommodation. Business pioneers, investigator and scientists should go through immense number of records on an everyday premise to remain ahead and a curiously large part of their time is spent simply choosing what report has importance and what isn't. By discovering significant and important sentences and making outlines, it's feasible to rapidly look at whether a record merits perusing. In this paper, we propose to plan programmed text summarizer to sum up the numerous text reports. The contribution to the framework is the different wellsprings of news stories. Significant sentences from the source record are chosen and arranged in the objective archives or the summed up reports. This is called as the extraction method in programmed text outline. Here, one archive visited is being handled and its outline is created utilizing extraction method through which a weighted graph is built utilizing language preparing.**

This work presents the method of extracting summary of news articles from multiple sources on a selected topic. Multiple source of information is provided as an input to the text summarization system, summary of each single document is extracted, the individual summaries of each document are combined together and a single summary of upto 100 words is generated using text rank algorithm.

Keywords: Text summarization, Natural language processing, weighted graph, abstractive and extractive text summarization.

I. INTRODUCTION

Immense data is accessible to us over the Internet due to the availability of the World Wide Web. There is data accessible through different sources such as News authoritative records, diaries identified with different types, e-magazines, digital books, and so forth. Every day e-news sources furnish us with news containing data on different themes like governmental issues, sports, classifieds and so forth. Typically individuals simply need a significance of what's going on with the news because of absence of time to peruse every one of the wellsprings of information completely. Here, a text synopsis instrument is required that can save a ton of time.

Programmed Text Rundown is gathering or summarizing the text to an outline without being unintelligent. The text rundown can be done on a solitary source archive or from various wellsprings of reports, which can be additionally named single record summarizer or multi report summarizer separately. To peruse these accessible unlimited data structure online assets can be extremely monotonous and tedious for the clients connected with various areas of interest. A text summarizer can be utilized by wide assortment of individuals to peruse a brief archive. For instance, it very well may be utilized by an individual to get up to speed with news in a hurry, or for a legal counselor to triumph ultimately a last moment rundown, an understudy to comprehend an idea and so forth In this way, a summarizer can be utilized as an instrument for any individual to have a consolidated perspective on any article inside less time.

Text synopsis is the sub subject of Natural Language Processing which assists with typifying the huge information accessible on the web. The Text Summing up tool takes as info an archive that can be one free record or a set of reports, which the client can choose from wide assortment of sources like the Internet, client's very own records or digital books. Then, at that point, it performs pre-handling of the information, eliminates stop words and afterward tokenize. A sentence examination is then performed which incorporates sentence scoring and sentence positioning strategy. Subsequent to positioning the sentences, the rundown of these records is created, that can be utilized as a speedy access for the client.

Extractive and Abstractive summarization are the two types of text summarization techniques. In Extractive Text Summarization technique, a couple of chosen words is taken from the text and combined to shape a rundown. It tends to be considered as a highlighter that chooses the fundamental data

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 11, pp 52-56, November-2021

https://doi.org/10.47001/IRJIET/2021.511009

from a source text. In AI, extractive synopsis generally includes gauging the fundamental areas of sentences and utilizing the outcomes to get outlines. Various kinds of calculations and techniques can be utilized to measure the loads of the sentences and afterward rank them as per their significance and comparability with each other and further going along with them to produce a rundown.

1.1 Text Rank Algorithm

Text Rank is a graph-based text processing ranking model that may be used to determine the most relevant phrases and keywords in a text.

With this algorithm the most relevant sentences in text are found and a graph is generated with the vertices representing each phrase in the document and the edges linking sentences based on content overlap, i.e. computing the number of words shared by two sentences.

The sentences are sent into the Page rank algorithm, which finds the most important sentences, based on this network of sentences. We may now extract only the most important sentences from the text when extracting a summary.

The text rank algorithm builds a word network to find relevant keywords. This network is built by examining which words follow one another. If two words follow one another, a link is formed; the link gains weight if these two words appear more frequently in the text.

The Page rank algorithm is applied on top of the resulting network to determine the importance of each word. The top one-third of all of these words are retained and deemed relevant. Following that, a keywords table is created by combining the relevant words that appear next to one another in the text.

The most relevant sentences will be extracted in the example given below:

```
library(textrank)
data(joboffer)
```

```
cat(unique(joboffer$sentence), sep = "\n")
```

Statistical expert / data scientist / analytical developer

BNOSAC (Belgium Network of Open Source Analytical Consultants), is a Belgium consultancy company specialized in data analysis and statistical consultancy using open source tools.

In order to increase and enhance the services provided to our clients, we are on the lookout for an all-round statistical expert, data scientist and analytical developer.

Function

Your main task will be the execution of a diverse range of consultancy services in the field of statistics and data science.

You will be involved in a small team where you handle the consultancy services from the start of the project until the end.

This covers:

Joint meeting with clients on the topic of the analysis.

Acquaintance with the data.

Analysis of the techniques that are required to execute the study.

Mostly standard statistical and biostatistical modelling, predictive analytics & machine learning techniques.

Perform statistical design, modeling and analysis, together with more seniors.

Building the report on the data analysis.

Automating and R/Python package development

Integration of the models into the existing architecture.

Giving advise to the client on the research questions, design or integration.

Next to that, you will help in building data products and help sell them.

These cover text mining, integration of predictive analytics in existing tools and the creation of specific data analysis tools and web services.

You also might be involved in providing data science related courses for clients.

Profile:

You have a master degree in the domain of Statistics, Biostatistics, Mathematics, Commercial or Industrial Engineering, Economics or similar.

You have a strong interest in statistics and data analysis.

You have good communication skills, are fluent in English and know either Dutch or French.

You soak up new knowledge and either just make things work or have the attitude of 'I can do this'.

Besides this, you have attention to detail and adapt to changes quickly.

You have programming experience in R or you really want to switch to using R.

You have a sound knowledge of another data analysis language (Python, SQL, javascript) and you don't care in which relational database, Excel, bigdata or noSQL store your data is located.

Interested in robotics is a plus.

Offer:

A half or full time employment depending on your personal situation.

The ability to get involved in a whole range of sectors and topics and the flexibility to shape your own future.

The usage of a diverse range of statistical & data science techniques.

Support in getting up to speed quickly in the usage of R.

An environment in which you can develop your talent and make your own proposals the standard way to go.

Liberty in managing your open source projects during working hours.

Figure 1: Example

The textrank algorithm (keyword extraction / sentence ranking) requires as input the identification of words which are relevant in your domain. This is normally done by doing Parts of Speech tagging which can be done using a broad range of R packages.

```
head(joboffer[, c("sentence_id", "lemma", "upos")], 10)
```


International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 11, pp 52-56, November-2021

https://doi.org/10.47001/IRJIET/2021.511009

	sentence_id	lemma	upos
1	1	Statistical	ADJ
2	1	expert	NOUN
3	1	/	PUNCT
4	1	data	NOUN
5	1	scientist	NOUN
6	1	/	PUNCT
7	1	analytical	ADJ
8	1	developer	NOUN
9	2	BNOSAC	PROPN
10	2	(	PUNCT

Figure 2: POS

We get that job offer data frame as follows:

```
job_rawtxt <- readLines(system.file(package = "textrank", "extdata",
"joboffer.txt"))

job_rawtxt <- paste(job_rawtxt, collapse = "\n")


library(udpipe)

tagger <- udpipe_download_model("english")

tagger <- udpipe_load_model(tagger$file_model)

joboffer <- udpipe_annotate(tagger, job_rawtxt)

joboffer <- as.data.frame(joboffer)
```

Figure 3: Library

Vector of words is provided for extracting keywords from the job description to search the relevant keywords. In this case , we consider various parts of speech like noun, verbs and adjectives.

```
keyw <- textrank_keywords(joboffer$lemma,

                          relevant = joboffer$upos %in% c("NOUN", "VERB",
"ADJ"))

subset(keyw$keywords, ngram > 1 & freq > 1)
```

	keyword	ngram	freq
4	data-analysis	2	4
9	data-science	2	3
14	consultancy-service	2	2

1.2 Sentence ranking with Text Rank

```
head(joboffer[, c("sentence_id", "lemma", "upos")], 10)
```

	sentence_id	lemma	upos
1	1	Statistical	ADJ
2	1	expert	NOUN
3	1	/	PUNCT
4	1	data	NOUN
5	1	scientist	NOUN
6	1	/	PUNCT
7	1	analytical	ADJ
8	1	developer	NOUN
9	2	BNOSAC	PROPN
10	2	(	PUNCT

For applying textrank for sentence ranking, we need to feed the function textrank_sentences 2 inputs: - a data.frame with sentences and - a data.frame with words which are part of each sentence.

```
library(udpipe)

joboffer$textrank_id <- unique_identifier(joboffer, c("doc_id",
"paragraph_id", "sentence_id"))

sentences <- unique(joboffer[, c("textrank_id", "sentence")])

terminology <- subset(joboffer, upos %in% c("NOUN", "ADJ"))

terminology <- terminology[, c("textrank_id", "lemma")]

head(terminology)
```

	textrank_id	lemma
1	1	Statistical
2	1	expert
4	1	data
5	1	scientist
7	1	analytical
8	1	developer

While applying the textrank algorithm for a sentence it finds nouns/adjectives, etc. Which are the same in sentences and next applies Google Pagerank on the sentence network? The result is an object of class textrank_sentences which contains the sentences, the links between the sentences and the result of Google's Pagerank.

```
## Textrank for finding the most relevant sentences

tr <- textrank_sentences(data = sentences, terminology = terminology)

names(tr)

[1] "sentences"      "sentences_dist" "pagerank"

plot(sort(tr$pagerank$vector, decreasing = TRUE),

     type = "b", ylab = "Pagerank", main = "Textrank")
```

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 11, pp 52-56, November-2021

https://doi.org/10.47001/IRJIET/2021.511009

With the help of summary function, we can extract the top n most relevant sentences. By default it gives the sentences in order of Pagerank importance but we can also get the n most important sentences and keep the sentence order as provided in the original sentences data. Frame.

```
s <- summary(tr, n = 4)

s <- summary(tr, n = 4, keep.sentence.order = TRUE)

cat(s, sep = "\n")
```

Output for the implemented text algorithm is as given below:

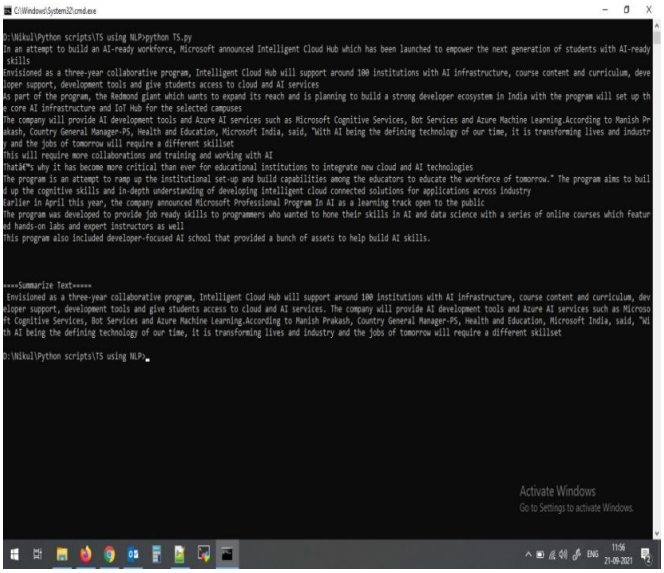

Figure 4: Output

II. LITERATURE SURVEY

[1] "Keyword Extraction from Scientific Articles in Bahasa Indonesia using TextRank Algorithm", D. Gunawan, F. Purnamasari, et al, suggests the use of text rank algorithm. This paper makes use of medical articles. The data used in this article is generated from various fields, namely technology, economics, law, culture, chemistry, physics, and mathematics. The quantity of medical articles which might be used in here is a hundred and fifty articles.

[2] "Implemented Text Rank based Automatic Text Summarization using Keyword Extraction" Anurag Kumar Yadav, et al, proposes the methodology of single and multiple document text summarizations through various approaches such as ranking algorithm and machine learning. It shows the recent trends in the summarization field with these approaches. An approach of extracting news from web source has been proposed in this paper.

III. RESULT

The implemented text rank algorithm works efficiently and provides a summary of 100 words, provided the document in .docx format. Multiple documents of the same news generates accurate summary. The efficiency lowers in the case where multiple documents of different genres of news is provided to the text summarizer.

IV. CONCLUSION

Text summarization system is proving beneficial for capturing information in various sectors in these fast paced lives. The summarization technique used in this work is very easy to implement as we used the basic algorithm, the Text rank algorithm for generating a summary of 100 words. This tool helps us summarize single or multiple documents. It works efficiently, provided the document provided as input belongs to the same genre and same news.

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my guide Dr Vipul Dalal, who gave me the opportunity to work on this wonderful project on the topic Text Summarization, Which helped and engaged me in doing a lot of research and I learnt a many new things. This work has helped me increase my knowledge and skills.

REFERENCES

[1] Prayana Trisna, I Nyoman & Nurwidyantoro, Arif. (2020). Single document keywords extraction in Bahasa Indonesia using phrase chunking. TELKOMNIKA (Telecommunication Computing Electronics and Control). 18. 1917. 10.12928/telkomnika.v18i4.14389.

[2] Yadav, Anurag & Kumar, Mukesh & Pathre, Ayonija. (2020). Implemented Text Rank based Automatic Text Summarization using Keyword Extraction. International Research Journal of Innovations in Engineering and Technology. 04. 20-25. 10.47001/IRJIET/2020.411003.

[3] A Kazemi, V Pérez-Rosas, R Mihalcea - arXiv preprint arXiv:2011.01026, 2020 - arxiv.org

[4] Elayeb, B., Chouigui, A., Bounhas, M. et al. Automatic Arabic Text Summarization Using Analogical Proportions. Cogn Comput 12, 1043–1069 (2020). https://doi.org/10.1007/s12559-020-09748-y

[5] Belwal, R.C., Rai, S. & Gupta, A. A new graph-based extractive text summarization using keywords or topic modeling. J Ambient Intell Human Comput 12, 8975–8990 (2021). https://doi.org/10.1007/s12652-020-02591-x

ISSN (online): 2581-3048
Volume 5, Issue 11, pp 52-56, November-2021
https://doi.org/10.47001/IRJIET/2021.511009

[6] Zhenrong Deng, Fuxin Ma, Rushi Lan, Wenming Huang, Xiaonan Luo, A Two-stage Chinese text summarization algorithm using keyword information and adversarial learning, Neurocomputing, Volume 425, 2021, Pages 117-126, ISSN 0925-2312.
[7] Taner Uçkan, Ali Karcı, Extractive multi-document text summarization based on graph independent sets.
[8] Egyptian Informatics Journal, Volume 21, Issue 3, 2020, Pages 145-157, ISSN 1110-8665, https://doi.org/10.1016/j.eij.2019.12.002.

AUTHORS BIOGRAPHY

Ms. Abhilasha More, Lecturer, IT Dept. Shri Bhagubhai Mafatlal Polytechnic, Research Scholar, Vidyalankar Institute of Technology, Mumbai , Maharashtra, India.

Dr. Vipul Dalal, Associate Professor, Vidyalankar Institute of Technolgy, Mumbai, Maharashtra, India.

Citation of this Article:

Abhilasha More, Vipul Dalal, "Implemented Text Summarization Tool using Text Rank Algorithm" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 5, Issue 11, pp 52-56, November 2021. Article DOI https://doi.org/10.47001/IRJIET/2021.511009

****

