Document Summarization using NLP techniques

Dona Hasini Gammune
Department of Mathematical Sciences
University of Texas at Dallas
Richardson, USA
dvg190000

Mananage Sanjaya Kumara
Department of Mathematical Sciences
University of Texas at Dallas
Richardson, USA
mmk190004

Shradha Upadhyay
Department of Mathematical Sciences
University of Texas at Dallas
Richardson, USA
sxu140730

Sahra Premani
Department of Mathematical Sciences
University of Texas at Dallas
Richardson, USA snp200002

***Abstract*—In this paper, an algorithm is developed for extractive text summarization using feature vectorization method called Term frequency-inverse document frequency (TF-IDF). This Term frequency-inverse document frequency method coded from scratch reflects the importance of a term to a document in the collection of texts. The proposed approach is comprised on the four different steps: 1) loading Data-Text processing stop word remover, 2) calculating TF-IDF, 3) scoring each sentence, 4) summary generation using highest scored sentences. In-built Spark libraries are used for data loading, stop words removal and data cleaning. The proposed approach was assessed using the BBC news summary dataset which consists of 2225 documents from BBC news website corresponding to stories in five topical areas from 2004-2005. The proposed algorithm enables the user to select the number of sentences present in the summary and is efficient in generating meaningful extractive summaries.**

Keywords— PySpark, text-summarization, Term Frequency-Inverse Document Frequency (TF-IDF), NLP, Summary generation using highest scored sentences.

I. INTRODUCTION

Natural language processing is used in many disciplines such as computer science, and mathematics. Natural language processing is useful for common use cases such as question answering, paraphrasing or summarizing, sentiment analysis, natural language BI, language modeling, and disambiguation. [1] The requirement for automatic text summarization has significantly increases as the amount of textual data rises. To identify and sort most important information in to summary is an important task. One of the processes to create summary with the major points of the document is Text summarization process. Text Summarization technique can automatically generate the important information from big amount of data and compresses in to shorter version. Two main summarization techniques based on output type are extractive summarization and abstractive summarization. To generate a summary with extractive summarization is, selecting important information from the source document where the same sentences can be seen in summary. Abstractive method of summarization presents a shorter version of the whole information of the document using Natural Language Processing which does not include the same sentences from input text for the generated summary [2]. Abstractive summarizer model forms its own sentences for summary. Extractive summarizer forms a summary by selecting meaningful information in sentences from input text data. There are many applications of text summarizer such as media monitoring, search marketing, internal document workflow, and financial research [3]. In this paper, we are trying to achieve the task of the extractive text summarization of a data set which consist of 2225 documents from BBC news website. In-built Spark libraries are used for data preprocessing and an algorithm is coded for the term frequency inverse document frequency (TF-IDF) and sentence scoring calculations.

II. BACKGROUND WORK

S. Zaware, D. Patadiya, et al. [4] conducted research on "Text Summarization using TF-IDF and Textrank algorithm". TF-IDF method was used for summarization. They have discussed that to manually summarize a large document, a combination of TFIDF and text rank algorithm with some NLP methods efficiently summarizes data and will perform better than the other systems. This study gives an idea of relevant methods that can be used for extractive text summarization.

S. Rauniyar Rahul and Monika.[5] conducted survey on "Deep learning based various methods analysis of Text Summarization". The study provides an idea about evaluating different text summarization methods.

A. Mishra and S. Vishwakarma.[6] conducted "Analysis of TF-IDF Model and its Variant for Document Retrieval". It is about a system with quality of storing, extracting, and keeping information. The experiments were conducted with Term Frequency-Inverse Document Frequency (TF-IDF) and its variants. They have concluded that by using the TF-IDF the information retrieval systems can demonstrate to be more efficient and provide the highest precision results.

Shahzad Qaiser and Ramsha Ali.[7] conducted study on "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents". They have concluded that TF-IDF is a very useful algorithm with some limitation. They have also discussed the

significance, related work, exceptions and solutions of the algorithm.

III. THEORETICAL AND CONCEPTUAL STUDY OF THE TECHNIQUE//ALGORITHM

A. System Model

In figure 1, a brief summary is presented for document summarization using natural language processing technique to study importance of the words to document. This approach is comprised of following steps: text processing, calculating TF-IDF value for words, scoring each sentence, and summary generation using highest scored sentences.

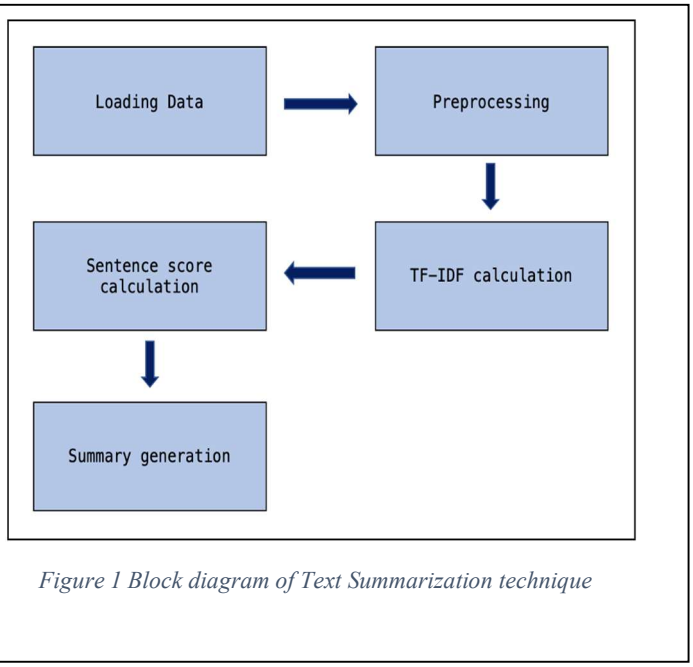

Figure 1 Block diagram of Text Summarization technique

Fig 1. Block diagram of Text Summarization technique

B. Text preprocessing

Text processing is a first and required step. Feature transformation processes modify features of the data. It is useful to transform the data into machine readable format [8]. It removes unnecessary information like punctuations, extra spaces, etc. Some of the preprocessing feature transformation processes are tokenization, stop word removing, normalization, filtering and stemming [9].

- Tokenization: Tokenization is a process of splitting sentences in to single words, and removing all punctuations. [9]
- Stop word removal: Stop words are words appearing too frequently in the sentences but do not have specific meaning such as “a”, “the”, “is”, etc. [6]. Stop word remover takes output of tokenizer and remove all the stops words from it [8].
- Stemming: this process reduces words by removing it to their stem words. It is useful to lower redundancy, and to decrease the size of document [6].

C. TF-IDF Method,and Equations

A feature extraction method used in text mining is to show the importance of a term to a document in the corpus is called Term frequency-inverse document frequency (TF-IDF). It denotes corpus by D, document by d, and term by t. [8]

Term frequency TF (t, d) is the number of times that term t appears in document d. It is useful to measure the importance of the terms that appear very often but have very little information about the document. Only term frequency calculation is not enough. [8]

$$\text{TF}(t,d)$$

Equation 1

Document frequency DF (t, D) is the number of documents which contain term t. And Inverse document frequency is a numerical calculation of how much important information a term provides.[8]

$$\text{IDF}(t, D) = log\frac{|D|+1}{DF(t,D)+1}$$

Equation 2

To measure TF-IDF the product of (TF * IDF) is calculated. TF-IDF values are calculated using equation below. It will choose relevant and important content and prepare a summary. [8]

$$\text{TF-IDF}(t,d,D) = \text{TF}(t,d) * \text{IDF}(t,D)$$

Equation 3

TF-IDF used in text summarization method which is the extractive method. TF-IDF value assigns some weight to each word in the document. The weight will be lower when the frequency will be less, and higher weight when frequency will be higher. [2]

Next step is to calculate sentence score. The sentence score will be calculated using two different approaches. Two different approaches are following,

Let $\text{TF} - \text{IDF}_{i,j}$ be the TF-IDF value of i^{th} word in sentence j.

Let n_j be the number of non-stop words in sentence j.

1. Summation of TF-IDF

$$\text{Score} = \sum_{i=1}^{n_j} \text{TFIDF}_{i,j}$$

Equation 4

The sentence score will be calculated by calculating the summation TF-IDF values of the words in a sentence using above equation 4.

2. Average of the word scores

$$\text{Score} = \frac{\sum_{i=1}^{n_j} \text{TFIDF}_{i,j}}{n_j}$$

Equation 5

The score will be calculated by dividing summation of TF-IDF values of the words by number of non-stop words in that sentence using equation 5.

The sentences will be arranged by descending sentence score value. The algorithm will extract the set percentage of important sentences having higher sentence score, and with sentences with low sentence score will be discarded. The original text summary will be generated.

D. Implementation

In this section, one article from the BBC News Summary dataset was used to evaluate the proposed method. In Method, we are following text summarization algorithm steps shown in the figure 2.

Algorithm Text Summarization Algorithm

Input: Data file

Output: Summary file

Summarization_TF_IDF(Input_file, num_sentences_for_summary)

{

Step 1: Preprocessing:

1. Splitting into sentences
2. Sentence indexing
3. Removing punctuations
4. Tokenizing
5. Removing stop-words
6. Matching stem words

Step 2: TF-IDF calculation

Step 3: Sentence scores calculation:

Let $\omega_{i,j}$ be the TF-IDF value of i^{th} word of sentence j.

Let n_j be the number of non-stop words in sentence j.

Sentence score of sentence j is;

Using sum

$$Sentence\_Score_j = \sum_{i=1}^{n_j} \omega_{i,j}$$

Using average

$$Sentence\_Score_j = \frac{\sum_{i=1}^{n_j} \omega_{i,j}}{n_j}$$

Step 4: Summary generation:

1. Sorting sentences with sentence score.
2. Generating the summary with number of sentences specified.

}

Figure 2 Text summarization algorithm steps ASK about references for this table

The news article "Google's toolbar sparks concern" was used to produce a summary of the text by following steps shown in the figure 2.

- Step 1: *INPUT* (snippet of news article used for method evaluation)

```
Google's toolbar sparks concern

Search engine firm Google has released a trial tool which is concerning some net users
because it directs people to pre-selected commercial websites.

The AutoLink feature comes with Google's latest toolbar and provides links in a webpage
to Amazon.com if it finds a book's ISBN number on the site. It also links to Google's
map service, if there is an address, or to car firm Carfax, if there is a licence plate.
Google said the feature, available only in the US, "adds useful links". But some users
are concerned that Google's dominant position in the search engine market place could
mean it would be giving a competitive edge to firms like Amazon.

AutoLink works by creating a link to a website based on information contained in a
webpage - even if there is no link specified and whether or not the publisher of the
page has given permission.
```

Figure 3 Snippet of Input

- Step 2: Preprocessing (The input data is preprocessed by removing punctuations, tokenizing, and removing stop-words.)

Table 1 Preprocessing output

Sentence ID	Preprocessed words
1	[google, toolbar, spark, concern]
2	[search, engine, firm, google, release, trial, tool, concern, net, user, direct, people, preselect, commercia, website]
3	[autolink, feature, come, google, latest, toolbar, provide, link, webpage, amazoncom, find, book, isbn, number, site]
4	[also, link, google, map, service, address, car, firm, carfax, license, plate]
5	[google, said, feature, avail, add, use, link]
6	[user, concern, google, domain, position, search, engine, market, place, mean, give, competitor, edge, firm, like, amazon]

- Step 3: Calculation of TF-IDF values by using the formula from equation 1, equation 2, and equation 3.

$$\text{Score} = \frac{\sum_{i=1}^{n_j} \text{TFIDF}_{i,j}}{n_j}$$

Table 2 Calculated TF-IDF values

Sentence ID	Tokenized words	TF-IDF
11	media	2.6848453616444100
13	never	2.6848453616444100
22	use	2.595904278307630
22	service	1.4807253789884900
9	library	1.3424226808222100
15	new	1.3424226808222100
9	online	1.3424226808222100
19	sign	1.3424226808222100
9	list	1.3424226808222100
1	spark	1.3424226808222100
10	advertise	1.3424226808222100

- Step 4: Sentence Score calculation using two different methods. One is summation of TF-IDF using equation 4, and another approach is dividing summation of TF-IDF values of the words by number of non-stop words in that sentence using equation 5.

Summation of TF-IDF

$$\text{Score} = \sum_{i=1}^{n_j} \text{TFIDF}_{i,j}$$

Average of the word scores

Table 3 Sentence Score Calculation using two different methods, Top 10 sentences scores using the methods sum and average.

Using Sum		Using Average	
Sentence Id	Average of sentence scores	Sentence Id	Average of sentence scores
11	1.2991	22	22.4975
20	1.1727	11	22.0844
22	1.1249	20	19.9358
7	1.1178	7	17.8847
16	1.0789	6	16.1718
4	1.0481	2	13.1762
13	1.0166	3	13.0660
6	1.0107	9	12.8715
9	0.9901	19	12.7113
10	0.9567	16	11.8679

- Ten sentences' scores using the two different methods which are sum and average result generated is shown in Table 3.

- Step 5: Sentences are ordered in descending order of their sentence score. The set default 25% of highest ranked sentences are selected for the summary. And sentences will be arranged in the order as per input original text. Finally, with set default percent summary will be generated from original input text.

IV. RESULTS AND ANALYSIS.

Experimental results and observations are presented in tabular and graphical formats.

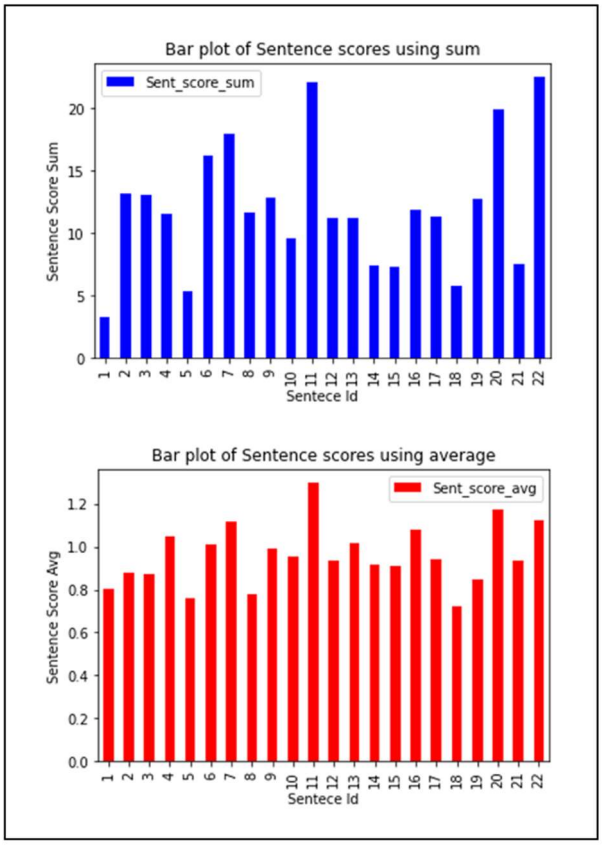

Figure 4 Bar plot of sentence scores using sum and average Methods

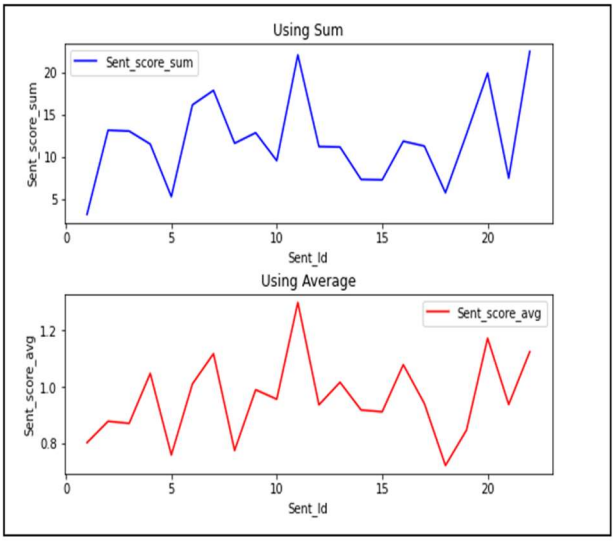

Figure 5 Line plot of sentence scores using sum and average Methods

Figure 4 shows output results of bar plot using calculation performed by using equation 4, and equation 5. The top panel shows the Bar plot of sentence score of each sentence calculated using the sum of TF-IDF values of non-stop words in the sentence. The lower panel shows the Bar plot of sentence score of each sentence calculated using the average of TF-IDF values of non-stop words in the sentence.

Figure 5 shows output results of bar plot using calculation performed by using equation 4, and equation 5. The top panel shows the variation of sentence score of each sentence is calculated using the sum of TF-IDF values of non-stop words in the sentence and the lower panel shows the variation of the sentence score of each sentence calculated using the average of TF-IDF values of non-stop words in the sentence.

From figure 4 and figure 5, sentences 11, 20, and 22 are the most significant sentences for summary generation. As per Table 3, if the summary is generated in 10 sentences both methods will generate summaries which are having a 70% of similarity with each other.

There is a limitation of using the sum of TF-IDF values of non-stop words in the sentence as the sentence score. Generally, longer sentences without significant words gets higher score because they have higher numbers of words than a shorter sentence. This can be resolve by using the average of TF-IDF values of non-stop words in the sentence as the sentence score by normalizing scores [10].

Table 4 Original Text Document

Original document example

Google's toolbar sparks concern

Search engine firm Google has released a trial tool which is concerning some net users because it directs people to pre-selected commercial websites.

The AutoLink feature comes with Google's latest toolbar and provides links in a webpage to Amazon.com if it finds a book's ISBN number on the site. It also links to Google's map service, if there is an address, or to car firm Carfax, if there is a licence plate. Google said the feature, available only in the US, "adds useful links". But some users are concerned that Google's dominant position in the search engine market place could mean it would be giving a competitive edge to firms like Amazon.

AutoLink works by creating a link to a website based on information contained in a webpage - even if there is no link specified and whether or not the publisher of the page has given permission.

If a user clicks the AutoLink feature in the Google toolbar then a webpage with a book's unique ISBN number would link directly to Amazon's website. It could mean online libraries that list ISBN book numbers find they are directing users to Amazon.com whether they like it or not. Websites which have paid for advertising on their pages may also be directing people to rival services. Dan Gillmor, founder of Grassroots Media, which supports citizen-based media, said the tool was a "bad idea, and an unfortunate move by a company that is looking to continue its hypergrowth". In a statement Google said the feature was still only in beta, ie trial, stage and that the company welcomed feedback from users. It said: "The user can choose never to click on the AutoLink button, and web pages she views will never be modified. "In addition, the user can choose to disable the AutoLink feature entirely at any time."

The new tool has been compared to the Smart Tags feature from Microsoft by some users. It was widely criticised by net users and later dropped by Microsoft after concerns over trademark use were raised. Smart Tags allowed Microsoft to link any word on a web page to another site chosen by the company. Google said none of the companies which received AutoLinks had paid for the service. Some users said AutoLink would only be fair if websites had to sign up to allow the feature to work on their pages or if they received revenue for any "click through" to a commercial site. Cory Doctorow, European outreach coordinator for digital civil liberties group Electronic Fronter Foundation, said that Google should not be penalised for its market dominance. "Of course Google should be allowed to direct people to whatever proxies it chooses. "But as an end user I would want to know - 'Can I choose to use this service?, 'How much is Google being paid?', 'Can I substitute my own companies for the ones chosen by Google?'." Mr Doctorow said the only objection would be if users were forced into using AutoLink or "tricked into using the service".

Table 5 Generated final summary output

Summary

Using sum

Search engine firm Google has released a trial tool which is concerning some net users because it directs people to pre-selected commercial websites.. But some users are concerned that Google's dominant position in the search engine market place could mean it would be giving a competitive edge to firms like Amazon.. AutoLink works by creating a link to a website based on information contained in a webpage - even if there is no link specified and whether or not the publisher of the page has given permission.. Dan Gillmor, founder of Grassroots Media, which supports citizen-based media, said the tool was a "bad idea, and an unfortunate move by a company that is looking to continue its hypergrowth". Cory Doctorow, European outreach coordinator for digital civil liberties group Electronic Fronter Foundation, said that Google should not be penalised for its market dominance. "But as an end user I would want to know - 'Can I choose to use this service?, 'How much is Google being paid?', 'Can I substitute my own companies for the ones chosen by Google?'." Mr Doctorow said the only objection would be if users were forced into using AutoLink or "tricked into using the service".

Using average

It also links to Google's map service, if there is an address, or to car firm Carfax, if there is a licence plate. AutoLink works by creating a link to a website based on information contained in a webpage - even if there is no link specified and whether or not the publisher of the page has given permission.. Dan Gillmor, founder of Grassroots Media, which supports citizen-based media, said the tool was a "bad idea, and an unfortunate move by a company that is looking to continue its hypergrowth". It was widely criticised by net users and later dropped by Microsoft after concerns over trademark use were raised. Cory Doctorow, European outreach coordinator for digital civil liberties group Electronic Fronter Foundation, said that Google should not be penalised for its market dominance. "But as an end user I would want to know - 'Can I choose to use this service?, 'How much is Google being paid?', 'Can I substitute my own companies for the ones chosen by Google?'." Mr Doctorow said the only objection would be if users were forced into using AutoLink or "tricked into using the service".

Table 4 and Table 5 shows original text input and the number of sentences resulted from summary after following all the steps from figure 2.

Table 5 shows the output summaries generated by the two methods, using the sum of TF-IDF vales as the sentence score and using average of TF-IDF values as the sentence score. The sentences that are included in both of the summaries are quite similar. Hence, we can say that both methods can be useful to generate document summaries.

V. CONCLUSION AND FUTURE WORK

In this paper, we have developed a simple extractive text summarization algorithm using the sentence scores. We have considered two approaches in calculating the sentence scores, one using the sum of TF-IDF values and the other using the average of TF-IDF values of non-stop words in the sentence. The implemented algorithm provides an effective solution in generating extractive summary. The summary is generated from the list of sentences which have the highest sentence scores. The set default 25% of highest ranked sentences are selected for the summary. Also, the developed algorithm has the option for the user to select the number or the proportion of sentences they need for the summary. We have generated summary for the “” news report using both the approaches. The sentences that are included in both the summaries are mostly similar. Hence, we can say that both methods can be useful to generate document summaries.

In document summarization, both the sum and the average of the TF-IDF values of the non-stop words can be used to calculate the sentence score in a document. However, using average of the TF-IDF values has some advantages over using the sum. Most of the times, longer sentences get higher score as they contain a greater number of words than a shorter sentence, although the actual relevance of the words is the same. But averaging the sum, normalizes the scores. The sum can be greatly affected by the outlier words which are either extremely rare or extremely common in the document. The average is robust to outliers so that we can outcome the issues with the outliers. By the same time, average score is more intuitive and can be easily interpretable.

TF-IDF only consider term frequency of the word. An algorithm can be improved by considering more parameter such as sentence position, similarity with the title [11].

REFERENCES

[1] V.Kocaman, and D.Talby “Spark NLP: Natural language Understanding at scale” Software Impacts Volume8, May 2021, 100058

[2] U.Mavani,D.Shinde, A.Pednekar, and S.Hamdare “Natural language processing based text summarization and querying model,”IEEE Proceedings of the 8th International Conference on Advanced Intelligent Systems and Informatics 2022, vol.152, pp.627, 2023.

[3] I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand and P. K. Soni, "Natural Language Processing (NLP) based Text Summarization - A Survey," *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, 2021, pp. 1310-1317, doi: 10.1109/ICICT50816.2021.9358703.

[4] S. Zaware, D. Patadiya, A. Gaikwad, S. Gulhane and A. Thakare, "Text Summarization using TF-IDF and Textrank algorithm," *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2021, pp. 1399-1407, doi: 10.1109/ICOEI51242.2021.9453071.Sdsd

[5] S. Rauniyar Rahul and Monika, "A Survey on Deep Learning based Various Methods Analysis of T ext Summarization", *2020 International Conference on Inventive Computation Technologies (ICICT)*, pp. 113-116, 2020.

[6] A. Mishra and S. Vishwakarma, "Analysis of TF-IDF Model and its Variant for Document Retrieval", *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 772-776, 2015.

[7] Shahzad Qaiser and Ramsha Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents", *International Journal of Computer Applications (0975–8887)*, vol. 181, no. 1, July 2018.

[8] ApacheSparkMLlib:MainGuide https://spark.apache.org/docs/latest/mllib-feature-extraction.html#tf-idf

[9] S. M. H. Dadgar, M. S. Araghi and M. M. Farahani, "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification," *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, Coimbatore, India, 2016, pp. 112-116, doi: 10.1109/ICETECH.2016.7569223.

[10] Salton, G., & Buckley, C. "Term-weighting approaches in automatic text retrieval"(1988) Information processing & management, 24(5), 513-523.

[11] T.Sri Ramu Raju, Bhargav Allarpu "Text Summarization using Sentence Scoring Method" April 2017 International Research Journal of Engineering and Technology, Dept Of CSE Engineering, GITAM University, Andhra Pradesh, India e-ISSN:2395-005, p-ISSN:2395-0072