Recent Trends in Intensive Computing
M. Rajesh et al. (Eds.)

doi:10.3233/APC210179

Text Summarizing Using NLP

G. Vijay Kumar [1], Arvind Yadav, B. Vishnupriya, M. Naga Lahari, J. Smriti, D. Samved Reddy
Department of Computer Science &Engineering, Koneru Lakshmaiah Education Foundation, Greenfields, Vaddeswaram, Guntur-522502, India

Abstract. In this era everything is digitalized we can find a large amount of digital data for different purposes on the internet and relatively it's very hard to summarize this data manually. Automatic Text Summarization (ATS) is the subsequent big one that could simply summarize the source data and give us a short version that could preserve the content and the overall meaning. While the concept of ATS is started long back in 1950's, this field is still struggling to give the best and efficient summaries. ATS proceeds towards 2 methods, Extractive and Abstractive Summarization. The Extractive and Abstractive methods had a process to improve text summarization technique. Text Summarization is implemented with NLP due to packages and methods in Python. Different approaches are present for summarizing the text and having few algorithms with which we can implement it. Text Rank is what to extractive text summarization and it is an unsupervised learning. Text Rank algorithm also uses undirected graphs, weighted graphs. keyword extraction, sentence extraction. So, in this paper, a model is made to get better result in text summarization with Genism library in NLP. This method improves the overall meaning of the phrase and the person reading it can understand in a better way.

Keywords. Text Rank, Text summarization, NLP, Extractive, Abstractive.

1. Introduction

The whole idea of automatic text summarization is to collect the necessary and crisp points from a large amount of data. There is a lot of information that is available on the internet and it also keeps growing every day and having to collect the main data from it becomes hard since it takes a lot of time. The use of automatic text summarization makes it easier for the users to collect the important data from huge information. Some of the graph base ranking algorithms are Text Rank [1], Hyperlinked Induced Topic Search [2], Positional Power Function [3] and so on. In this paper we are going to implement Text Rank algorithm. Noting down the important points manually from large amount of data can be a very stressful job. So, automatic text summarization takes out the crucial words and sends them back in a way that the readers find it easy. This, automatic text summarization is a small piece of NLP which cut downs the information and sends to the readers. It also arranges the information and sends back the sentences that are useful to create a crisp summary.

[1] G Vijay Kumar,Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation
Email: gvijay_73@kluniversity.in.

The words that occur the most no of times are considered the most worth. The top most words are also arranged and then a summary is created. The Extractive methodology chooses the principal significant lines from the information text and utilizes them to think of the outline. The abstractive methodology addresses the information text in a type then generates the outline with the desired output of words and sentences that disagree from the first text sentences. Extractive systems extract vital text units (such as sentences, paragraphs etc.) within the input document. The theoretic approach is practically identical to the way that human summarizers 1st perceives the most ideas of a document, so generate new sentences that aren't seen within the original document. The general design of an associate ATS system has the subsequent tasks: Pre-processing, Processing, Post-Processing. Text summarization is in this field as a conclusion that monitors are needed to grasp what humans have composed and generated human-readable outputs. human language technology also can be viewed as a study of computer science (AI). Therefore, several existing AI algorithms and strategies, as well as neural network models, are used for finding human language technology connected issues. With the present analysis, researchers typically believe 2 kinds of approaches for text summarization as shown in Figure 1 extractive summarization and abstractive summarization.

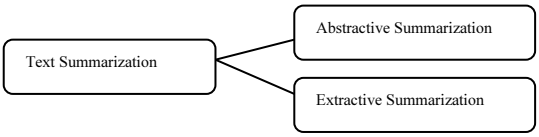

Figure 1. Types of Summarizations

2. Related work

Pratibha Devi Hosur et al [4] suggest the system by implementing unsupervised learning during automatic text summarization. This paper expresses the overall depiction of summarization of text using NLP which includes input text document, pre-processing, lesk algorithm and finally generating the summary. The lesk algorithm outputs, calculations, conclusion and proposed system. Narendra Andhale et al [5] This paper discussed about abstractive and extractive summarizations. They described how long texts are summarized in less time. and they focused on the extractive summarization methods. Deepali K Gaikwad et al [6] This paper expresses regarding how the necessary information is extracted from the long text document and forms the summary. Regular patterns [7, 8] be useful in Text summarization to extract useful keywords. They mainly discussed about abstractive method and extractive method and their approaches. Neelima Bhatia et al [9], in this study paper researched the famous and considerable effort done in the area of unit and numerous archive outlines. The creators examined the technique-based methodologies for summarization of text. These technique-based methodologies incorporate term-based recurrence strategy, diagram-based technique, time-based strategy, division and combining based strategy, semantic reliance strategy, theme-based methodologies, talk based methodologies, Latent Semantic based methodologies, approaches dependent on lexical chain, approaches dependent on fluffy rationale.

N. Moratanch et al [10] clarified about the strategy of summarization of text is that removed data is gotten as synopsis report and introduced as a small outline to the client. In the work creators examined about word level highlights, sentence level highlights and different extractive content synopsis techniques. The creators in this paper proposed a portion of the assessment measurements like human assessment, Rogue score, Recall, Precision, F-measure, Compression proportion. Shohreh Rad Rahimi et al [11] claimed that NLP explores are with more interest in summing up literary data. In this paper, creators characterized text synopsis as interaction of naturally making and lessening type of given report and holding its data content source into more limited variant with right importance. In this paper the creators likewise clarified about the connection between text mining and text outline. At last, this paper examines about different ways to deal with text outline, for example, Statistical methodologies, Lexical Chain based methodologies, Cluster based methodologies, Fuzzy rationale-based methodologies.

3. Problem Definition

In our busy schedule, it is very difficult for us to go through the entire article or document. So, we prefer to read summary. In this paper we are going to summarize the large text in to a short summary which reduces reading time for users.

4. Methodology

NLP is a part of Artificial Intelligence reasoning that manages analyzing, understanding, finding and producing the dialects that people use in a characteristic manner to make interface with PCs in both composed and spoken settings utilizing common human dialects rather than codes.

4.1. TextRank Algorithm

Text rank algorithm is a diagram-based positioning model for text processing which can be used in order to find the most applicable sentences in text and also to find keywords. Text rank algorithm is similar to page rank algorithm. Page rank algorithm is used to mark Webpages in web search conclusions and in web usage mining. In text rank algorithm, in position of Webpages sentences are taken.

1. Identify content units that best characterize the current task, and add them as vertices in the diagram.
2. Identify the relations that append the content units, and in the chart utilizing these relations draw edges between vertices. Edges can be un-weighted or weighted and undirected or coordinated.
3. And at that point loop the diagram-based positioning algorithm until union.
4. Based on their last score mastermind the vertices. For positioning and determination choices Use the qualities appended to every vertex.
5. At last, the highest-level sentences will shape a synopsis.

5. Proposed System Results

As shown in Figure 2, Source document is the input text given. Preprocessing: Tokenization is the technique used to split the text in to tokens (words or paragraphs or sentences). Stop words is used to reduce the size of text, we have a dictionary in pre-processing which is made up of stop words. It compares the words in given text and then remove the matched words. Hence removal of stop words will increase the performance. feature extraction: word frequency means most common word that occur

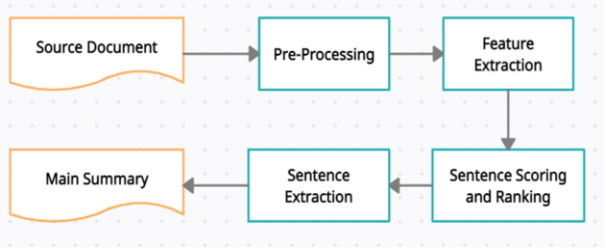

Figure 2. Flow Chart

in text are measure of information. It is determined as the quantity of occurrence of word by complete number of words in the archive. Too long or too short sentences are eliminated utilizing length of sentences. It is determined as number of words in the sentence by number of words in longest sentence. Sentence scoring and ranking: it calculates the score for each sentence and rank them. Sentence extraction: The main target of this is to identify best in the text. The target of this is to rank complete sentences. Main summary: place the sentences in order and generate the resultant summary.

5.1. TextRank Model

Graph based algorithms is the most required strategy of determining the powerful of a vertex in the actual graph, elicited from over all information gathered from the entire graph. The fundamental idea we have implemented here is voting and recommendation. Based on the votes casted, the score is related with vertex. We implement “random surfer model” as the probability that skip from one vertex to some other vertex. The score of a graph, starting from arbitrary values and the computation iterates. The score of a vertex is based on the importance of vertex and the last qualities are not affected by beginning qualities. Figure 3, 4, 5, 6 shows the results for TextRank algorithm and for single document we use textrank and for multi-document we use lexrank.

5.2. UnDirected graphs

Basically, we apply recursive graph-based ranking algorithm on directed graphs, as the out-degree and in-degree is equal it is also enforced for un directed graph In convergence curve as the connectivity of the graph increases then fewer iterations take place and the convergence curve for in-directed or directed graph practically overlap.

5.3. Weighted graphs

As the main definition of PageRank for graph based learning algorithm is we have to assume un weighted graph and as the graphs are constructed from nlp, textrank may include multiple or partial link between units. Based on the weight of edge textrank incorporate the power of connectivity which we can see in Figure 7.

$$WS(V_i) = (1-d) + d * + \sum_{V_j \in In\,(V_i)} \frac{w_{ji}}{\sum_{Vk \in Out(V_j=0} w_{jk}} WS(V_j)$$

When we compute the score related with the vertex in a graph then the latest anon takes in to account edge weights. The final vertex ranking and scores differ as compared to original measure and the number of iterations is nearly same for unweighted or weighted graphs.

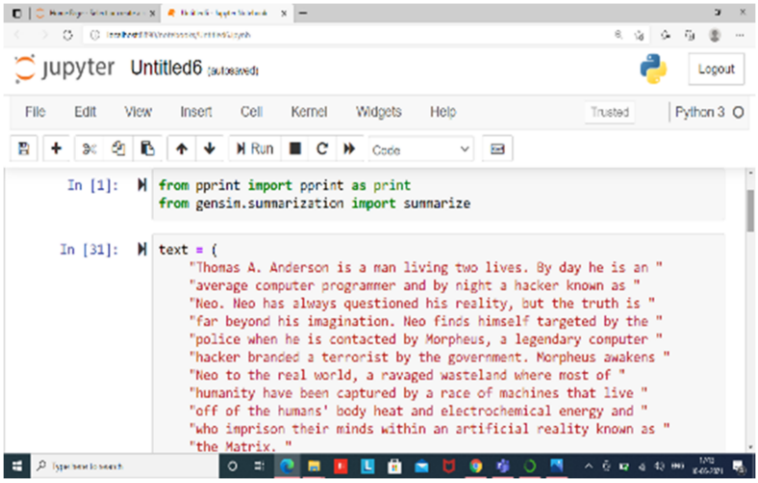

Figure 3. Text Paragraph

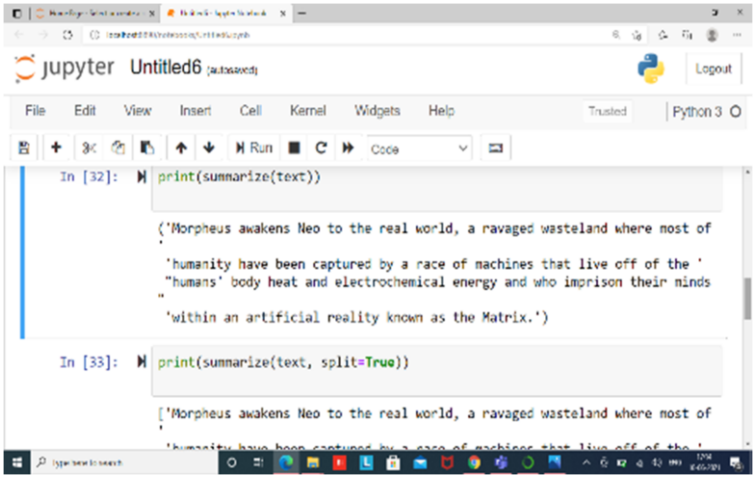

Figure 4. Summarizing Text Paragraph

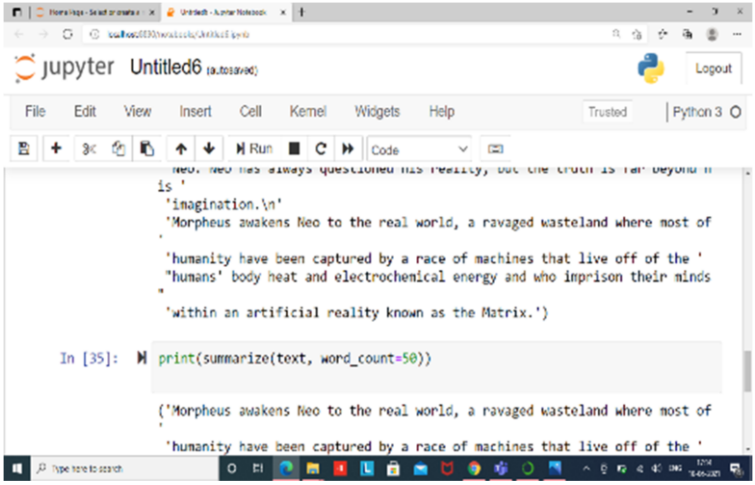

Figure 5. Summarizing Text with word count = 50

Figure 6. Summarizing Text with ratio = 0.5

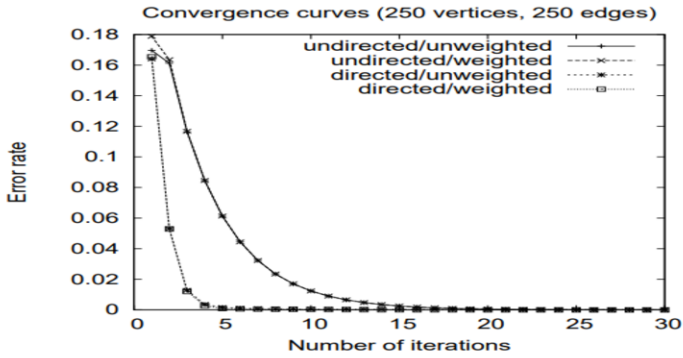

Figure 7. Convergence Curves for weighted graphs

In this, the module automatically summarizes the given input text and it finally it picks up the important sentences. It can also extract keywords as shown in execution.

6. Conclusion

The paper demonstrates that we use advanced techniques to apply on the document for text summarization using extractive summarization method called TextRank algorithm. At first, we loaded necessary libraries and related function in python and then code was implemented to summarize the text. Afterwards, a model is proposed with slight expansions to improve by showing the outline text. The techniques that are presented in this paper to get better result in text summarization with Genism library in NLP. With this the overall meaning of the document can be understand easily.

References

[1] Hans Peter Luhn, “The automatic creation of literature abstracts”, IBM Journal.

[2] Kleinberg J. M., “Authoritative sources in a hyperlinked environment”. Journal of the ACM, Volume 46 issue 5, pp.604–632, Sep 1999.

[3] Herings, G. van der Laan, and D. Talman, “Measuring the power of nodes in digraphs”, Technical report, Tinbergen Institute, 2001.

[4] Pratibha Devihosur, Naseer R. "Automatic Text Summarization Using Natural Language Processing" International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 08, Aug-2017

[5] Deepali K. Gaikwad, C. Namrata Mahender, "A Review Paper on Text Summarization", “International Journal of Advanced Research in Computer and Communication Engineering”. Vol.5, Issue 3, March 2016.

[6] G. Vijay Kumar, M. Sreedevi, NVS Pavan Kumar, “Mining Regular Patterns in Transactional Databases using Vertical Format”.“International Journal of Advanced Research in Computer Science”, Volume 2, Issue 5, 2011.

[7] G. Vijay Kumar and V. Valli Kumari, “Sliding Window Technique to Mine Regular Frequent Patterns in Data Streams using Vertical Format”, IEEE International Conference on Computational Intelligence and Computing Research, 2012.

[8] Neelima Bhatia and Arunima Jaiswal, "Automatic Text Summarization and its Methods-AReview", 6th International Conference. Cloud System and Big Data Engineering, 2016.

[9] Potharaju, S. P., & Sreedevi, M. (2017). A Novel Clustering Based Candidate Feature Selection Framework Using Correlation Coefficient for Improving Classification Performance. Journal of Engineering Science & Technology Review, 10(6).

[10] Shohreh Rad Rahimi, Ali Toofan zadeh Mozhdehi and Mohamad Abdolahi, "An Overview on Extractive Text Summarization".“IEEE 4th International Conference on Knowledge Based Engineering and Innovation” (KBEI) Dec. 22nd, 2017, Iran University of Science and Technology – Tehran, Iran.