TEXT SUMMARIZATION USING THE TEXTRANK ALGORITHM

Lince Salu Varghese
Department of Computer Applications
Amal Jyothi College of Engineering,
Koovapally
Kottayam, Kerala.
lincesaluvarghesre2022@mca.ajce.in

Mr. Rony Tom
Asst. Professor in Computer Applications
Amal Jyothi College of Engineering,
Koovapally
Kottayam, Kerala.
ronytom@amaljyothi.ac.in

***Abstract*— Automatic text summarization is a technique for producing a succinct and accurate summary. The machine learning algorithm may be trained to understand texts and identify the areas that contain key facts and information before constructing the required summary phrases. We are using the text rank-based Automatic Text Summarization in this study work to create quality summaries and keywords, which are essential for text summarization. The suggested summarizing system's performance is evaluated.**

Keywords—Automatic Text Summarization, TextRank Algorithm, PageRank Algorithm

I.INTRODUCTION

Text Summarization is one of those Natural Language Processing (NLP) technologies that will undoubtedly have a significant impact on our daily lives. One of the most difficult and fascinating tasks in Natural Language Processing is automatic text summarization. It's a method for extracting a brief and meaningful summary of text from a variety of sources, including books, news stories, blog posts, research papers, emails, and tweets. The availability of vast amount of textual data has sparked a surge in demand for automatic text summarizing systems. In this article text summarization is discussed and how to use the TextRank algorithm and how to put it into practice in python.

II.APPROACHES TO TEXT SUMMARIZATION

Automatic text summarization has been a source of fascination since the 1950s. "The automated production of literary abstracts," a research paper by Hans Peter Luhn published in the late 1950s, used factors like word frequency to pick keywords from the text for summarising.

Another notable study, conducted by Harold P Edmundson in the late 1960s, extracted relevant sentences for text summary using approaches such as the existence of cue words, terms from the title occurring in the text, and sentence location. Many important and intriguing works on the topic of automatic text summarization have been published since then.

The two most common types of text summary are extractive and abstractive summarization.

Extractive Summarization: This approach works by extracting many components from a text, such as phrases and sentences, then stacking them together to generate a summary. As a result, finding the proper phrases for summary is critical in an extraction approach.

To create a completely new summary, abstractive summarization employs advanced natural language processing algorithms. It's possible that some of the details in this summary aren't included in the original text. This study will concentrate on the technique of extractive summarization.

III.PAGERANK ALGORITHM

The PageRank algorithm provides a probability distribution that is used to predict whether a user will wind up on a given website after clicking on random links. Any number of documents can be used to compute PageRank. At the outset of the computational approach, some research articles assume that the distribution is evenly distributed throughout all documents in the collection. To update predicted PageRank values to more closely approximate the theoretical real value, the PageRank calculations require multiple runs over the collection, referred to as "iterations."

Assume a four-page document containing the letters A, B, C, and D on each page. Links between pages are disregarded, as are outbound links from a single page to another single page. For all pages, PageRank is set to the same value. Because the entire number of pages on the web at the time was equal to the sum of PageRank across all pages in the original version of PageRank, in this case, each page would start with a value of one. The rest of this section, as well as later versions of PageRank, assumes that each page has a probability distribution between and, with 0.25 as the beginning value.

On the next cycle, the PageRank sent from a specific page to the targets of its outbound links is split evenly across all outbound connections. If the system's only connections to A were from pages B, C, and D each link would transmit 0.25 PageRank to A on the following iteration, for a total of 0.75.

$$PR(A) = PR(B) + PR(C) + PR(D).$$

Consider what would happen if page B was linked to pages C and A, page C to page A, and page D to all three sites. As a consequence, on the first repeat, page B would communicate half of its current value, or 0.125 to page A, and the other half, or 0.125, to page C. Because it had three outbound links, Page C's whole current value, 0.25, would be transferred to A's sole existing value, or roughly 0.083. At the end of this cycle, Page A will have a PageRank of around 0.458, and the page it connects to, A. D, will transmit one-third of its PageRank.

$$PR(A) = PR(B)/2 + PR(C)/1 + PR(D)/3$$

DOI: 10.5281/zenodo.6369622
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering KanjiraKottayamppally,

To put it another way, the PageRank of an outgoing link is capable the PageRank score of the document divided by the quantity of outward-bound links L ( )

$$PR(A) = PR(B)/L(B) + PR(C)/L(C) + PR(D)/L(D)$$

1.

IV. TEXTRANK ALGORITHM

Instead of pages, we employ phrases in the TextRank algorithm. The chance of a web page change is determined by the similarity of the two texts. The similarity scores are kept in a square matrix that resembles PageRank's.

TextRank is an unsupervised extractive text summarising approach. The flow of the TextRank algorithm is :

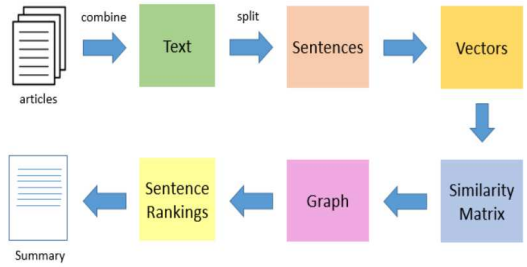

Fig 1: Flow of TextRank Algorithm

The first step would be to concatenate all of the articles' text. The content would then be separated into distinct sentences.

In the following phase, we'll look for vector representations (word embeddings) for each sentence.

The similarity between sentence vectors is then calculated and stored in a matrix.

2. The similarity matrix is reworked into a graph with sentences as vertices and similarity scores as edges to determine the rank of the sentence.

3. Finally, the final summary is made out of a selection of top-ranked sentences.

V. IMPLEMENTATION OF THE TEXTRANK ALGORITHM

In this article TextRank algorithm is used to create a clean and succinct summary from a collection of scraped articles. Please bear in mind that this is a single-domain-multiple-documents summarising project, which means we'll be using a variety of articles as input and creating a single bullet-point summary.

Using Jupyter Notebook lets implement the TextRank algorithm.

A. Import the necessary libraries

```
In [4]: import numpy as np
        import pandas as pd
        import nltk
        nltk.download('punkt') # one time execution
        import re

        [nltk_data] Downloading package punkt to
        [nltk_data]     C:\Users\Lince\AppData\Roaming\nltk_data...
        [nltk_data]   Package punkt is already up-to-date!
```

.

B. Examine the Information

```
In [5]: df = pd.read_csv(r"D:\tennis_articles.csv",encoding = 'unicode_escape')

In [6]: df.head()
```

Out[6]:

	article_id	article_title	article_text	source
0	1	I do not have friends in tennis, says Maria Sh...	Maria Sharapova has basically no friends as te...	https://www.tennisworldusa.org/tennis/news/Mar...
1	2	Federer defeats Medvedev to advance to 14th Sw...	BASEL, Switzerland (AP) □ Roger Federer advanc...	http://www.tennis.com/pro-game/2018/10/copil-s...
2	3	Tennis: Roger Federer ignored deadline set by ...	Roger Federer has revealed that organisers of ...	https://scroll.in/field/899938/tennis-roger-fe...
3	4	Nishikori to face off against Anderson in Vien...	Kei Nishikori will try to end his long losing ...	http://www.tennis.com/pro-game/2018/10/nishiko...
4	5	Roger Federer has made this huge change to ten...	Federer, 37, first broke through on tour over ...	https://www.express.co.uk/sport/tennis/1036101...

Article id, article text, and source are the three columns in our dataset. The 'article text' column has the text of the articles, which is what we're most interested in. Let's print a few of the variable's values to observe how they appear.

```
In [7]: df['article_text'][0]
```

Out[7]: "Maria Sharapova has basically no friends as tennis players on the WTA Tour. The Russian player has no problems in openly speaking about it and in a recent interview she said: 'I don't really hide any feelings too much. I think everyone knows this is my job here. When I'm on the courts or when I'm on the court playing, I'm a competitor and I want to beat every single person whether they're in the locker room or across the net. So I'm not the one to strike up a conversation about the weather and know that in the next few minutes I have to go and try to win a tennis match. I'm a pretty competitive girl. I say my hellos, but I'm not sending any players flowers as well. Uhm, I'm not really friendly or close to many players. I have not a lot of friends away from the courts.' When she said she is not really close to a lot of players, is that something strategic that she is doing? Is it different on the men's tour than the women's tour? 'No, not at all. I think just because you're in the same sport doesn't mean that you have to be friends with everyone just because you're categorized, you're a tennis player, so you're going to get along with tennis players. I think every person has different interests. I have friends that have completely different jobs and interests, and I've met them in very different parts of my life. I think everyone just thinks because we're tennis players we should be the greatest of friends. But ultimately tennis is just a very small part of what we do. There are so many other things that we're interested in, that we do.' ALSO READ: Maria Sharapova reveals how tennis keeps her motivated."

```
In [8]: df['article_text'][1]
```

Out[8]: "BASEL, Switzerland (AP) \x97 Roger Federer advanced to the 14th Swiss Indoors final of his career by beating seventh-seeded Daniil Medvedev 6-1, 6-4 on Saturday. Seeking a ninth title at his hometown event, and a 99th overall, Federer will play 93th-ranked Marius Copil on Sunday. Federer dominated the 20th-ranked Medvedev and had his first match-point chance to break serve again at 5-1. He then dropped his serve to love, and let another match point slip in Medvedev's next service game by netting a backhand. He clinched on his fourth chance when Medvedev netted from the baseline. Copil upset expectations of a Federer final against Alexander Zverev in a 6-3, 6-7 (6), 6-4 win over the fifth-ranked German in the earlier semifinal. The Romanian aims for a first title after arriving at Basel without a career win over a top-10 opponent. Copil has two after also beating No. 6 Marin Cilic in the second round. Copil fired 26 aces past Zverev and never dropped serve, clinching after 2 1/2 hours with a forehand volley winner to break Zverev for the second time in the semifinal. He came through two rounds of qualifying last weekend to reach the Basel main draw, including beating Zverev's older brother, Mischa. Federer had an easier time than in his only previous match against Medvedev, a three-setter at Shanghai two weeks ago."

```
In [9]: df['article_text'][2]
```

Out[9]: 'Roger Federer has revealed that organisers of the re-launched and condensed Davis Cup gave him three days to decide if he would commit to the controversial competition. Speaking at the Swiss Indoors tournament where he will play in Sunday\x92s final against Romanian qualifier Marius Copil, the world number three said that given the impossibly short time frame to make a decision, he opted out of any commitment. \x93They only left me three days to decide,\x94 Federer said. \x93I didn\x92t to have time to consult with all the people I had to consult. \x93I could not make a decision in that time, so I told them to do what they wanted.\x94 The 20-time Grand Slam champion has voiced doubts about the wisdom of the one-week format to be introduced by organisers Kosmos, who have promised the International Tennis Federation up to $3 billion in prize money over the next quarter-century. The competition is set to feature 18 countries in the November 18-24 finals in Madrid next year, and will replace the classic home-and-away ties played four times per year for decades. Kosmos is headed by Barcelona footballer Gerard Pique, who is hoping fellow Spaniard Rafael Nadal will play in the upcoming event. Novak Djokovic has said he will give precedence to the ATP\x92s intended re-launch of the defunct World Team Cup in January 2020, at various Australian venues. Major players feel that a big event in late November combined with one in January before the Australian Open will mean too much tennis and too little rest. Federer said earlier this month in Shanghai in that his chances of playing the Davis Cup were all but non-existent. \x93I highly doubt it, of course. We\x92ll see what happens,\x94 he said. \x93I don\x92t think this was designed for me, anyhow. This was designed for the future generation of players.\x94 Argentina and Britain received wild cards to the new-look event, and will compete along with the four 2018 semi-finalists and the 12 teams who win qualifying rounds next February. \x93I don\x92t like being under that kind of pressure,\x94 Federer said of the deadline Kosmos handed him.'

We now have two options: manually summarize each item or construct a single summary for all articles. We'll go with the latter for our purposes.

C. Split Text into Sentences

The material must now be broken down into distinct sentences. To do so, we'll utilize the nltk library's sent tokenize () method.

DOI: 10.5281/zenodo.6369622
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering KanjiraKottayamppally,

```
In [10]: from nltk.tokenize import sent_tokenize
         sentences = []
         for s in df['article_text']:
           sentences.append(sent_tokenize(s))

         sentences = [y for x in sentences for y in x] # flatten list

In [11]: sentences[:5]

Out[11]: ['Maria Sharapova has basically no friends as tennis players on the WTA Tour.',
          "The Russian player has no problems in openly speaking about it and in a recen
         t interview she said: 'I don't really hide any feelings too much.",
          'I think everyone knows this is my job here.',
          "When I'm on the courts or when I'm on the court playing, I'm a competitor and
         I want to beat every single person whether they're in the locker room or across
         the net.",
          "So I'm not the one to strike up a conversation about the weather and know tha
         t in the next few minutes I have to go and try to win a tennis match."]
```

Word embedding's is a sort of vector representation of words used in GloVe. These word embedding's will be used to vectorize our texts. We could have created features for our sentences using the Bag-of-Words or TF-IDF techniques, but these algorithms neglect the order of the words.

```
# Extract word vectors
word_embeddings = {}
f = open('glove.6B.100d.txt', encoding='utf-8')
for line in f:
    values = line.split()
    word = values[0]
    coefs = np.asarray(values[1:], dtype='float32')
    word_embeddings[word] = coefs
f.close()
```

```
sentence_vectors = []
for i in clean_sentences:
  if len(i) != 0:
    v = sum([word_embeddings.get(w, np.zeros((100,))) for w in i.split()])/(len(i.split())+0.001)
  else:
    v = np.zeros((100,))
  sentence_vectors.append(v)
```

```
len(sentence_vectors)
```

For our phrases, we'll need to make vectors. To arrive at a consolidated vector for a phrase, we will first retrieve vectors (each with 100 elements) for the component words in the sentence, the mean/average of the vectors is then calculated.

The next step is to find similarities between the sentences and we do this using the cosine similarity approach. For this task create an empty similarity matrix and fill it with cosine similarities between the texts..

D. Applying PageRank Algorithm

Let's turn the similarity matrix sim mat into a graph before moving on. The edges of this network will indicate the similarity scores between the sentences, while the nodes will represent the phrases. To arrive at the sentence ranks on this network, we'll use the PageRank method.

```
import networkx as nx

nx_graph = nx.from_numpy_array(sim_mat)
scores = nx.pagerank(nx_graph)
```

To generate a summary, extract the top N sentences depending on their rankings.

```
# Specify number of sentences to form the summary
sn = 10

# Generate summary
for i in range(sn):
  print(ranked_sentences[i][1])
```

```
When I'm on the courts or when I'm on the court playing, I'm a competitor and I want to beat every single person whether they're in the locker room or a
cross the net.So I'm not the one to strike up a conversation about the weather and know that in the next few minutes I have to go and try to win a tenni
s match.
Major players feel that a big event in late November combined with one in January before the Australian Open will mean too much tennis and too little re
st.
Speaking at the Swiss Indoors tournament where he will play in Sundays final against Romanian qualifier Marius Copil, the world number three said that g
iven the impossibly short time frame to make a decision, he opted out of any commitment.
"I felt like the best weeks that I had to get to know players when I was playing were the Fed Cup weeks or the Olympic weeks, not necessarily during the
tournaments.
Currently in ninth place, Nishikori with a win could move to within 125 points of the cut for the eight-man event in London next month.
He used his first break point to close out the first set before going up 3-0 in the second and wrapping up the win on his first match point.
The Spaniard broke Anderson twice in the second but didn't get another chance on the South African's serve in the final set.
"We also had the impression that at this stage it might be better to play matches than to train.
The competition is set to feature 18 countries in the November 18-24 finals in Madrid next year, and will replace the classic home-and-away ties played
four times per year for decades.
Federer said earlier this month in Shanghai in that his chances of playing the Davis Cup were all but non-existent.
```

VI. CONCLUSIONS

With the advancement of the Internet, a vast quantity of information is now available. Summarizing large quantities of

text is extremely difficult for humans. In this age of information overload, automated summarization systems are in high demand. There is an information overload as a result of the fast increase of knowledge and the usage of the Internet.

This difficulty can be handled if there are reliable text summarizers that provide a document summary for the user's convenience. As a result, a system must be developed that allows a user to quickly access and obtain a summary document.

A summary of a document using extractive or abstractive approaches is one such answer. Extractive text summarization is simpler to construct.

In this work, we focused on extractive techniques for automatic text summarising. We've gone through a handful of the more common methods. It gives a good overview of recent developments and advancements in automatic summarization methods, as well as the most up-to-date information in this field.

VII. REFERENCES

[1] Egyptian a informatics Journal, Extractic text summarization using modifiedpagerankalgorithm, https://www.sciencedirect.com/science/article/pii/S1110866519301355

[2] Analytics Vidhya, An introduction to text summarization using text rank algorithm(with python implementation), https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/

[3] Data Science in your pocket, Text summarization using text rankhttps://medium.com/data-science-in-your-pocket/text-summarization-using-textrank-in-nlp-4bce52c5b390

[4] Opengenus, Text rank for text summarization, https://iq.opengenus.org/textrank-for-text-summarization/

[5] Research Gate, Graph-based text summarization using modified text rank, https://www.researchgate.net/publication/327136473_Graph-Based_Text_Summarization_Using_Modified_TextRank

DOI: 10.5281/zenodo.6369622
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering KanjiraKottayamppally,