

Document Summarization Using NLP

CS 6307

under Prof. Anurag Nagar

Sai Abhishek Thota
Computer Engineering
University of Texas at Dallas
Dallas, Texas
saiabhishek.thota@utdallas.edu

Saicharan Reddy Ragannagari
Computer Engineering
University of Texas at Dallas
Dallas, Texas
saicharanreddy.ragannagari@utdallas.edu

Sai Shantanu Nagelli
Computer Engineering
University of Texas at Dallas
Dallas, Texas
saishantan.nagelli@utdallas.edu

Surekha Kumari
Computer Engineering
University of Texas at Dallas
Dallas, Texas
surekha.kumari@utdallas.edu

Abstract—In this paper we discuss the implementation of a method to summarize documents irrespective of its size. With abundance of data and information, it is essential that there be a method to automatically summarize documents in order to save people the effort and time in selecting relevant information. Automating summarizing is crucial to the amount of information that can be processed when compared with manually summarizing and this method is also wary of human biases. This paper further discusses the algorithm used and implementation steps.

Index Terms—implementation, summarize, algorithm [1]

I. INTRODUCTION

We have worked on the Python implementation of the text summarization algorithm, which seeks to extract the most important sentence from a given paragraph and generate summaries. The algorithm uses (NLP) techniques such as word frequency, sentence relevance, and graph-based algorithms to identify the most salient information in the text. That extraction approach is the most widely used method of summarizing text, given that it provides greater clarity and interpretation compared with abstract methods, which generate summaries based on paraphrasing and rewording. However, the whole meaning of the original text cannot be taken into account by extractive methods. We achieve this by using libraries such as spaCy and NLP library NLTK, we built a code around it to paraphrase long paragraphs into meaningful sentences to form a summary. This paper elaborates on the libraries used and algorithms implemented to achieve the desired result.

Index Terms—NLP, NLTK, spaCy

II. BACKGROUND WORK

A. NLP

NLP stands for Natural Language Processing, which is a subfield of Artificial Intelligence used to train machines

to understand Natural languages used by humans. It helps machines understand human language to perform redundant tasks post automation.

B. NLTK

NLTK stands for Natural Language ToolKit which is a python library used to aid NLP.

C. SpaCy

spaCy is a highly advanced, open source library used to perform NLP. It uses the best algorithms and is faster than NLTK when it comes to word tokenizing.

D. Stopwords

Stopwords are commonly used words which are irrelevant to the machine when constituting meaning from the text. It generally consists of articles, prepositions, conjunctions, pronouns, etc which are present in a text to aid human understanding. As machines do not need typical grammar, stopwords are filtered out for NLP.

III. THEORETICAL AND CONCEPTUAL STUDY

Text summarization, which can be used for a variety of applications in different areas, such as news summaries, documents, and chatbot responses, is an extensively discussed problem within the field of natural language processing NLP. The summarization of texts is made up of two main approaches: extractive and abstract. The best sentences or phrases in the original texts are selected by an extract summarization algorithm that uses them to produce a summary. These algorithms typically use statistical methods such as word frequency, sentence relevance, and graph-based algorithms to identify the most salient information. Generally, extraction

A. Code Description

Our Code implements text summarization algorithm which uses frequency-based approach to identify and extract meaningful sentences from the provided information. This algorithm uses several libraries such as pySpark, NLTK, spaCy and spark-nlp to achieve desired goal. The code is designed in order to first process the provided input with removing stop words (commonly used words that are irrelevant to the machine when constituting meaning data from the text. It generally consists of articles, prepositions, conjunctions, pronouns, etc which are present in a text to aid human understanding. As machines do not need typical grammar, stopwords are filtered out for NLP). We then use urllib specifically to store the above text so that we have the output in String format rather than table for ease of further steps. We further convert our string to spaCy readable format with Doctext to provide as input to spaCy operations. We use spaCy for dependency parsing and named entity recognition. We then calculate frequency of each word by assigning score to each sentence. The individual scores are then used to find max score to find top 30% words used to find most relevant sentences. The summary is finally generated using these sentences to capture important information.

We selected this algorithm under the assumption that most commonly words used are important and carry higher weight and should be present in the summary. When calculating individual frequency of the important words, the algorithm can identify words crucial to the document and can create condensed summary with relevant information.

Our output contains a summary of input document with all crucial information identified with NLP. This is achieved by selecting most relevant words and combining them together to form meaningful conclusions. This way of processing document is helpful when there is a need for quick extraction of information from huge documents or several documents that need to be preprocessed in smaller time frames.

our one potential limitation is that we rely solely on frequency of the words and may not provide precise summary in case of complex or nuanced language documents. Another limitation could be because the algorithm only extracts sentences and does not generate new sentences or paraphrase existing sentences, which may limit its ability to produce summaries that are highly readable or stylistically similar to the original text.

For this project, we are trying to sum up a paragraph of any size in its smaller form with no loss of integrity. Consequently, we used a summarizing algorithm from the extracted text. It is in its work that it finds the most important and relevant sentences of a paragraph, converting them into their shorter form.

Based on the project flow described, we have been able to find a frequency for each word in paragraphs 1, 2, and 3:

Fig. 1. word frequency

Then normalized frequency score is calculated for every word by dividing each frequency of a word by the maximum frequency. Following is the output for the line scores:

Line score is the sum of normalized frequency scores in a sentence. We are now going to get sentences with the top-line score. We wanted our algorithm to get 30% as a summary of the original document, so we selected the 30 largest function in the heapq library. nlargest will return the list of sentences with the amount of selected percentage we want.

```

1 line_scores = []
2 for data in line_tokens:
3     for words in data:
4         if words.text.lower() in wfreqs.keys():
5             if data not in line_scores.keys():
6                 line_scores[data] = wfreq(words.text.lower())
7             else:
8                 line_scores[data] += wfreq(words.text.lower())
9
10 line_scores = printing score of each line

Out[15]: (Near the ancient Indian kingdom of Mahishmati, an injured woman named Sivagami exits a cave underneath a mountain waterfall, carrying an infant.: 0.9666666666666667)
She kills two soldiers pursuing her and attempts to cross a raging river, but slips and is washed away in the current.: 0.7333333333333333
Before drowning, she holds the baby aloft and prays to Lord Shiva, explaining that she doesn't care about her life and wishes that the baby, Mahendra Baahubali must live.: 0.7
The child is saved by the people of the local Amburi tribe, who reside near the river and worship Lord Shiva.: 0.6666666666666667
The wife of the tribe's chieftain, Sanga, decides to adopt the boy and names him Sivudu.: 1.7666666666666667
Sivudu grows up to be an ambitious and mischievous child, obsessed with ascending the mountain.: 0.5
Despite Sangha's pleas, he tries many times to scale the cliffs but always fails.: 1.3888888888888889
As a young man, he is shown to possess superhuman strength when he lifts a Lingam of Lord Shiva and places it at the foot of the mountain.: 0.5333333333333333
A mask then falls from the cliffs, and realizing it possesses feminine features, Sivudu finally succeeds in scaling the mountain in order to find the woman the mask bel ongs.: 1.1666666666666667
Upon reaching the top, he sees a beautiful woman named Avantika fighting Mahishmati soldiers.: 0.7
He discovers that she is a member of a local resistance group dedicated to overthrowing the tyrannical king of Mahishmati, Lord Bhallaladeva, and recruiting royal captive Princess Devasena.: 0.9666666666666667
Sivudu is immediately smitten with Avantika and secretly follows her, even managing to draw a tattoo on her hand while she sleeps.: 0.40000000000000006
When she discovers Sivudu she attacks him, but he outmaneuvers her and returns her mask.: 0.2666666666666667

Out[16]:
The child is saved by the people of the local Amburi tribe, who reside near the river and worsh
The wife of the tribe's chieftain, Sanga, decides to adopt the boy and names him Sivudu.:
1.7666666666666667,
Sivudu grows up to be an ambitious and mischievous child, obsessed with ascending the mountain.
Despite Sangha's pleas, he tries many times to scale the cliffs but always fails.: 1.3888888888888889
As a young man, he is shown to possess superhuman strength when he lifts a Lingam of Lord Shiva
A mask then falls from the cliffs, and realizing it possesses feminine features, Sivudu finally ongs.: 1.1666666666666667,
Upon reaching the top, he sees a beautiful woman named Avantika fighting Mahishmati soldiers.:
He discovers that she is a member of a local resistance group dedicated to overthrowing the tyr Princess Devasena.: 0.9666666666666667,
Sivudu is immediately smitten with Avantika and secretly follows her, even managing to draw a t When she discovers Sivudu she attacks him, but he outmaneuvers her and returns her mask.: 0.2666666666666667

```

Fig. 2. Summary

```

1 summarizedData = nlargest(line_length, line_scores, key = line_scores.get)
2 summarizeData

Out[16]: (Amarendra is accepted as guard at the royal palace of Kuntala while Bhallaladeva perceives about Amarendra's acts and upon viewing Devasena's portrait, lusts f or her and acquires a greater love for Sivagami, who assures Bhallaladeva's marriage with Devasena alludes to the fact that Amarendra has already fallen in love with her.: 1.7666666666666667,
Lord Bijjaladeva, Vikramadeva's brother and the next in line for the throne, is denied the position due to his scheming nature, and as such Bijjaladeva's wife, Lady Siv agami, assumes power with the intention of raising both her son Bhallaladeva and the orphaned Baahubali in an equal manner to select the next heir to the throne.,
In Kuntala, an attack by the Pandur's defeats on the royal palace exposes Amarendra and Kattappa's bravery and they nullify the attack with assistance from Devasena's maternal cousin Kumar Varma who overcomes his cowardice.,
that Amarendra, Devasena and their unborn baby's lives shall be threatened by Bhallaladeva's machinations and manipulates him into entering the palace at stealth of nig ht for assassinating Bhallaladeva to do good.,
In the present day, Sivudu's adoptive parents, impressed by Kattappa's story, wish to meet Baahubali.,
In the process, he is recognized by a worker, who then leads the other workers in chanting "Baahubali", to Sivudu's confusion and Bhallaladeva's displeasure.,
With Kattappa's and Avantika's assistance, the army lays siege to Mahishmati.,
While Bhallaladeva kills Isidhar, Baahubali's valour and concern for the people of his kingdom convinces Sivagami to make Baahubali her apparent.,
After vanquishing the Kalakeyas, Amarendra is declared as the heir apparent to the throne while Bhallaladeva is announced as the kingdom's future commander-in-chief.,
He also orders his men to tear the head of Bhallaladeva's statue out of the palace walls, where it is kept to the great waterfall.,
The wife of the tribe's chieftain, Sanga, decides to adopt the boy and names him Sivudu.,
Bhallaladeva sends his son, Bhadra, and the royal family's loyal slave Kattappa to recapture Devasena.,
He falls into subterfuge at Sivudu's Feet, proclaiming him to be "Baahubali".)

```

Fig. 3. line frequency calculation

The summary of amount of words in the actual paragraph and the amount of words in the summarized paragraph is shown below

VII. CONCLUSION AND FUTURE WORK

Our code is a result of frequency based approach to extract text summaries. Algorithm used identifies the most relevant sentences in a text and then extract those sentences to combine and form a summary of the whole input provided. In conclusion, though we do have limitations, by using extractive text summarization, this approach is quick in extracting most important information from a large body of text, we now know the significance/effectiveness of the NLP in producing exact summaries of huge paragraphs. This algorithm selects the most important sentence in the paragraphs and adds all those into one summary. Our current implementation is limited to summarization only has room for improvement in areas where more precise information is to be extracted. The project could be extended to create presentation from extracted summaries, post automated online reviews.

ACKNOWLEDGMENT

It is my sincere thanks to Professor Anurag Nagar for his invaluable assistance, valuable observations, and constant support in the course of this project. He was instrumental in shaping our understanding of the subject and helping us

```

1 print("Input Document")
2 print(in_file)

Input Document
Near the ancient Indian kingdom of Mahishmati, an injured woman named Sivagami ex
fore drowning, she holds the baby aloft and prays to Lord Shiva, explaining that i
Lord Shiva. The wife of the tribe's chieftain, Sangha, decides to adopt the boy and

Sivudu grows up to be an ambitious and mischievous child, obsessed with ascending
Siva and places it at the foot of the mountain. A mask then falls from the cliffs
oman named Avantika fighting Mahishmati soldiers. He discovers that she is a membr
n with Avantika and secretly follows her, even managing to draw a tattoo on her h
feelings and later they have sex.

After she quietly faints Sivudu and leaves, she gets attacked by more soldiers. At e
city on Bhallaladeva's birthday and assists in erecting a gigantic statue of the
rates the royal palace disguised as a soldier and distracts Bhallaladeva and his j
s Bhadra as both the Amburi tribe and resistance warriors arrive. Kattappa lunges

The next morning, Kattappa reveals to Sivudu that Sivudu is actually Mahendra Baal
ther died giving birth to him. Lord Bijjaladeva, Vikramadeva's brother and the ne
llaladeva and the orphaned Baahubali in an equal manner to select the next heir t
comes beloved by the kingdom.

Command took 0.09 seconds -- by sxn210883@utdallas.edu at 4/30/2023, 8:15:12 PM on Project

Cmnd 25

1 print("Summarized text of Input Document")
2 print(summarizedData)

Summarized text of Input Document
Amarendra is accepted as guard at the royal palace of Kuntala while Bhallaladeva p
the fact that Amarendra has already fallen in love with her. Lord Bijjaladeva, Vik
n of raising both her son Bhallaladeva and the orphaned Baahubali in an equal mann
ce from Devasena's maternal cousin Kumar Varma who overcomes his cowardice. that
hallaladeva to do good. In the present day, Sivudu's adoptive parents, impressed t
a's displeasure. With Kattappa's and Avantika's assistance, the army lays siege to

After vanquishing the Kalakeyas, Amarendra is declared as the heir apparent to th
s swept to the great waterfall. The wife of the tribe's chieftain, Sangha, decides

Bhallaladeva sends his son, Bhadra, and the royal family's loyal slave Kattappa t

Kattappa is revealed to be the side companion of Mahendra's father.

It breaks as it falls and crashes against the cliff's walls and lands near the li
laladeva's treachery and exposes it to Shivagami, who regrets while Devasena deliv
river after being hit by an arrow shot by Bhallaladeva which leads her to a passag
he tries many times to scale the cliffs but always fails. Devasena, during her vi

```

Fig. 4. final output

```

Command took 0.09 seconds -- by sxn210883@utdallas.edu at 4/30/2023, 8:15:12 PM on Project

Cmnd 26

1 len(in_file)

Out[22]: 10897

Command took 0.09 seconds -- by sxn210883@utdallas.edu at 4/30/2023, 8:15:12 PM on Project

Cmnd 27

1 len(summarizedData)

Out[23]: 3227

Command took 0.10 seconds -- by sxn210883@utdallas.edu at 4/30/2023, 8:15:12 PM on Project

```

Fig. 5. length of summary

achieve our objectives through his expertise in Natural Language Processing. Our heartfelt thanks go out to our teacher's assistant, Truong Quang Pham, who has been doing everything he can to let us know our questions and help us keep track of the project.

We would like to thank our team for their contribution, SaiCharan Reddy, Sai Abhishek Thota, Sai Shantan Nagelli, and Surekha Kumari, who have played a major role in this project. This project's successful achievement was largely made possible through their dedication, hard work, and creativity. Lastly, we'd like to say our deepest thank you to each and everyone who has supported us in this endeavor directly or indirectly. Encouragement and feedback have been the source of motivation for us, and we appreciate your support.

REFERENCES

- [1] Dan W. Patterson, Introduction to Artificial Intelligence and Expert System , PHI, 2001, Chapter 12
- [2] <https://en.m.wikipedia.org/wiki/India>
- [3] <https://en.m.wikipedia.org/wiki/Dallas>
- [4] https://en.m.wikipedia.org/wiki/Spider-Man:_Far_From_Home
- [5] https://en.m.wikipedia.org/wiki/Baahubali:_The_Beginning
- [6] [https://en.m.wikipedia.org/wiki/Titanic_\(1997_film\)](https://en.m.wikipedia.org/wiki/Titanic_(1997_film))