NATURAL LANGUAGE PROCESSING

COMP 7780 - SPRING 2022

Name: Rahul Marru

UID: U00843883

Title: Text Summarization

Problem Definition:

In this project I doing that in general, if any person wants to know about any information, they need to read the entire text. But this is not possible in all cases. We cannot get an idea just by reading a part of it. There may be different meanings in another part. So, to get a clear understanding we need to read every line of the text. In the present day, people are leading very busy lives and lack patience in reading such tedious texts. Taking all the factors into consideration such as less time, precise text and patience we need to reduce the length of the text without lacking in exact meaning. The main aim of the project is to summarize the text base on all the factors. Through the proposed method of ours, summarization of the text can be easily done and can help people to know everything in particular of what they are about to do. This can avoid people to misuse their information unknowingly.

why it is important:

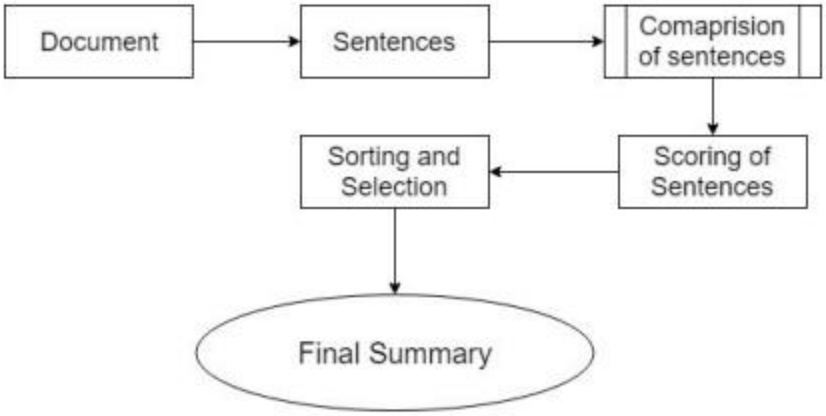
In today's digital world everything is revolving around data. Now every individual require everything to be in a precise manner and wants the work to be done as fast as possible. So, this project is mainly based on "Summarization of Text", which is a process of deflating the text to help people save their time. If the data generated is enormous every individual do not have time and patience to read and understand the given data. In this project we provide solution for the problem faced by user to get an idea about huge data by summarizing it. Text summarization is the process of automatically extracting a compressed version of the document by preserving its information. There are two types of summarization extractive and abstractive here we go with the extractive method which simplify the problem by considering the subsets of the sentence and its aim is to cover the important sentence for understanding the document. We developed this by considering text rank algorithm which follows by tokenization and vectorization. The experiment results shows that our suggested approach provides quality in summarization by maintaining its information.

Approach:

In this project I am using Extractive Text Summarization , which will extract the important sentences and gives the abstract for the document that helps in understanding easily. Text Summarization is implemented using an unsupervised Text Rank Algorithm. Graphical User Interfaces are created using advanced NLTK techniques. The input can be in the form of large chunks of text and any text document which is stored in your local hard disk.

Mainly focusing on unsupervised machine learning is used on the text given by the user. First we need to collect the document which contains text. The text can be in a file or any text document which is stored in your local hard disk or any URL.

Implementation:



Modules Used:

NLTK, Corpus, Tokenizer, TKINTER, Window, TopLevelWindow, Frame, Widget, and TTk.

Following steps by using unsupervised technique:

- Gather data
- Perform Pre-Processing

  * Delimiter removal - (, . ‘ “ ? !)

  * Stop word removal (which are mostly used by user) - the, is, an, a, like etc.

  * Stem word Removal (adjective or adverb of a word) - (ied, ed, mis, ary)

- Weighted word frequency
- calculating vectors for every sentence
- Sentences scores by using text rank algorithm

- Adding the highest sentence scores to summary

Implementing Algorithm Model:

Text Rank Algorithm:

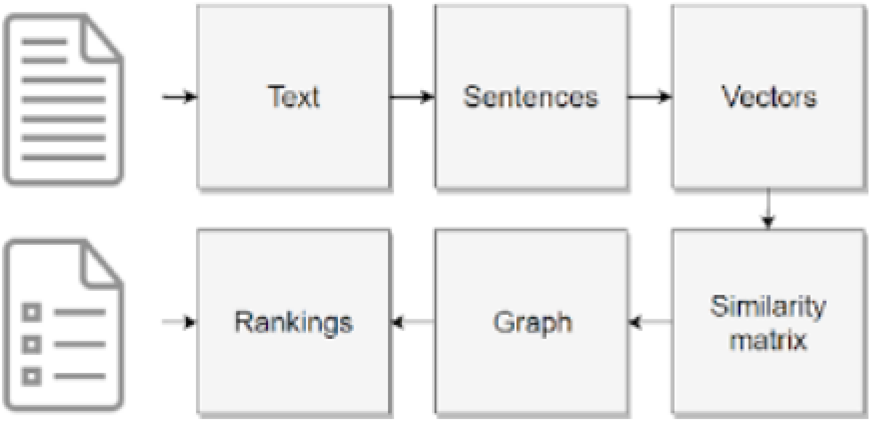
Text Rank is a set of rules primarily based totally on Page Rank, normally used to summarize text. Here we create a cosine matrix similarity in which we've got the similarity of every sentence to every other.

1. The first step is to divide the text document into sentences.

2. In the subsequent step, we are able to find the vector illustration of every sentence.

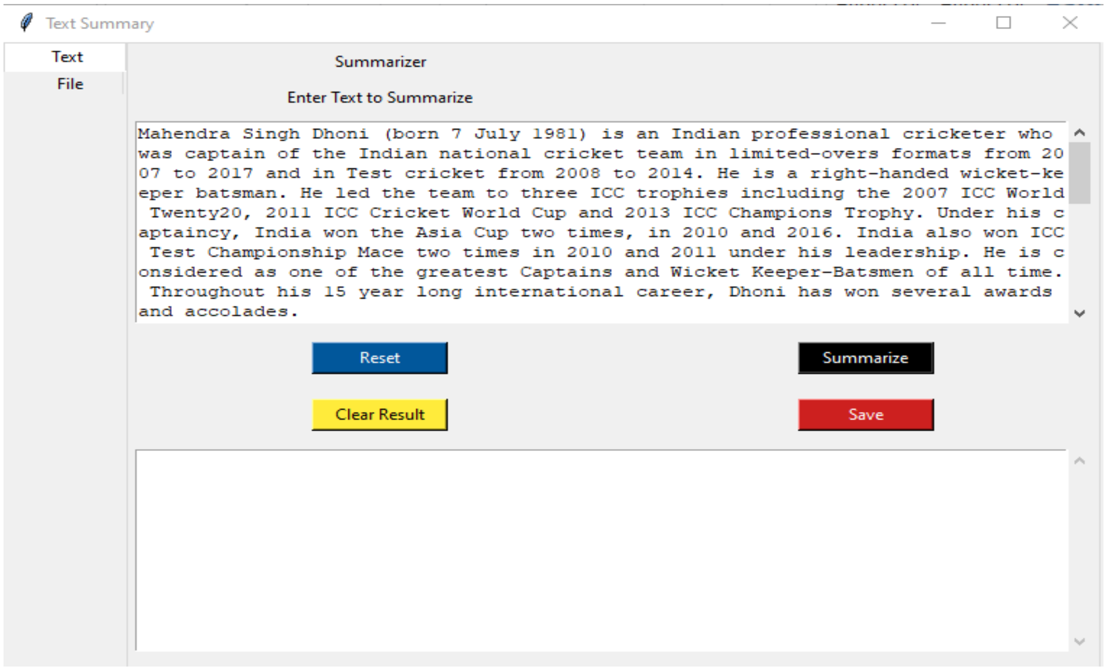
3. Similarities among sentence vectors are calculated and saved in a matrix.

4. The similarity matrix is then transformed right into a graph, with sentences as vertices and similarity ratings as edges, for sentence rank calculation. 5. Finally, the top-ranked sentences will combine and make it as a summary.



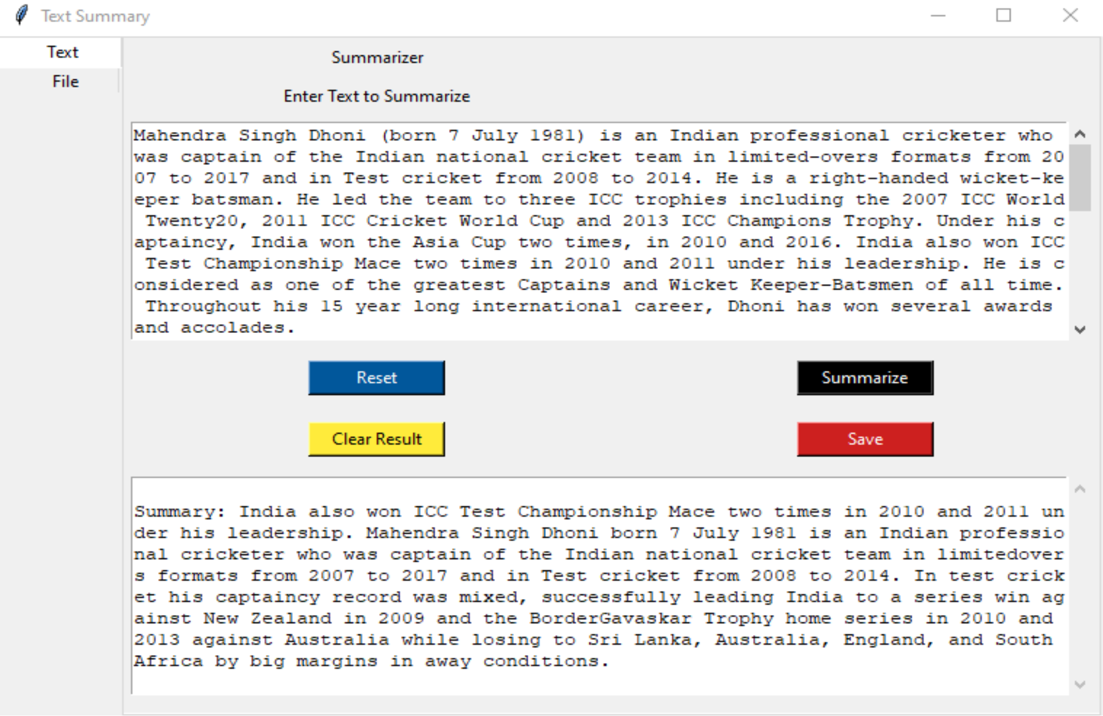
Result:

Case 1: summarize for text



```
Mahendra Singh Dhoni (born 7 July 1981) is an Indian professional cricketer who
was captain of the Indian national cricket team in limited-overs formats from 20
07 to 2017 and in Test cricket from 2008 to 2014. He is a right-handed wicket-ke
eper batsman. He led the team to three ICC trophies including the 2007 ICC World
 Twenty20, 2011 ICC Cricket World Cup and 2013 ICC Champions Trophy. Under his c
aptaincy, India won the Asia Cup two times, in 2010 and 2016. India also won ICC
 Test Championship Mace two times in 2010 and 2011 under his leadership. He is c
onsidered as one of the greatest Captains and Wicket Keeper-Batsmen of all time.
 Throughout his 15 year long international career, Dhoni has won several awards
and accolades.
```

Input screen for text

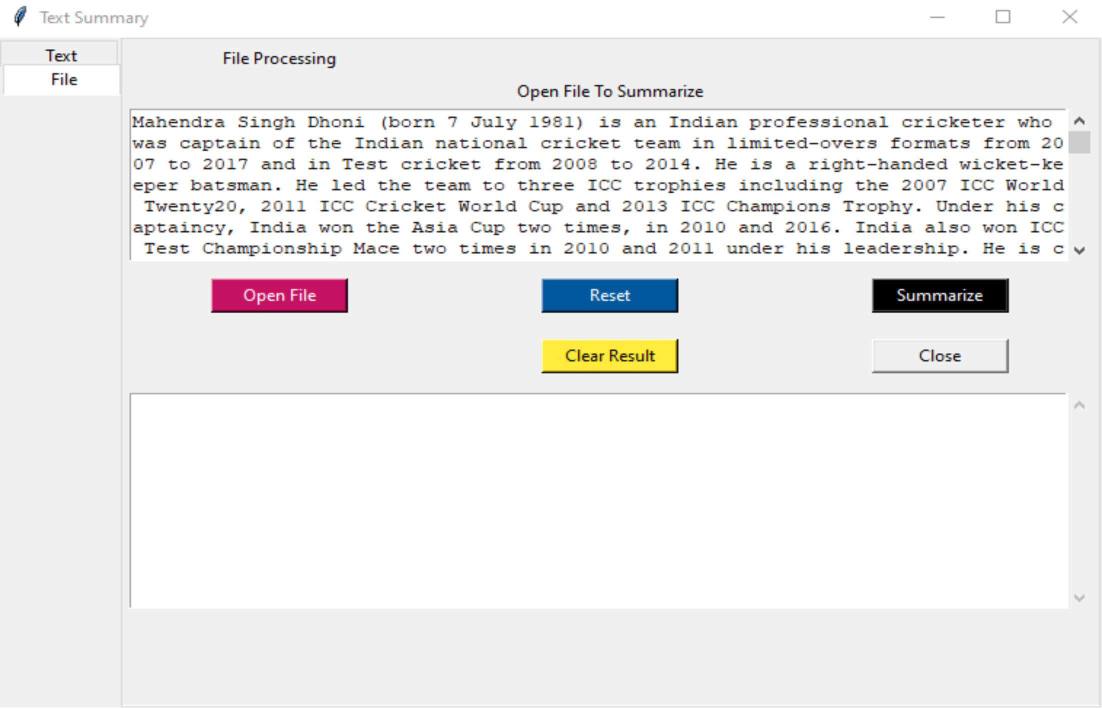


```
Mahendra Singh Dhoni (born 7 July 1981) is an Indian professional cricketer who
was captain of the Indian national cricket team in limited-overs formats from 20
07 to 2017 and in Test cricket from 2008 to 2014. He is a right-handed wicket-ke
eper batsman. He led the team to three ICC trophies including the 2007 ICC World
 Twenty20, 2011 ICC Cricket World Cup and 2013 ICC Champions Trophy. Under his c
aptaincy, India won the Asia Cup two times, in 2010 and 2016. India also won ICC
 Test Championship Mace two times in 2010 and 2011 under his leadership. He is c
onsidered as one of the greatest Captains and Wicket Keeper-Batsmen of all time.
 Throughout his 15 year long international career, Dhoni has won several awards
and accolades.
```

```
Summary: India also won ICC Test Championship Mace two times in 2010 and 2011 un
der his leadership. Mahendra Singh Dhoni born 7 July 1981 is an Indian professio
nal cricketer who was captain of the Indian national cricket team in limitedover
s formats from 2007 to 2017 and in Test cricket from 2008 to 2014. In test crick
et his captaincy record was mixed, successfully leading India to a series win ag
ainst New Zealand in 2009 and the BorderGavaskar Trophy home series in 2010 and
2013 against Australia while losing to Sri Lanka, Australia, England, and South
Africa by big margins in away conditions.
```

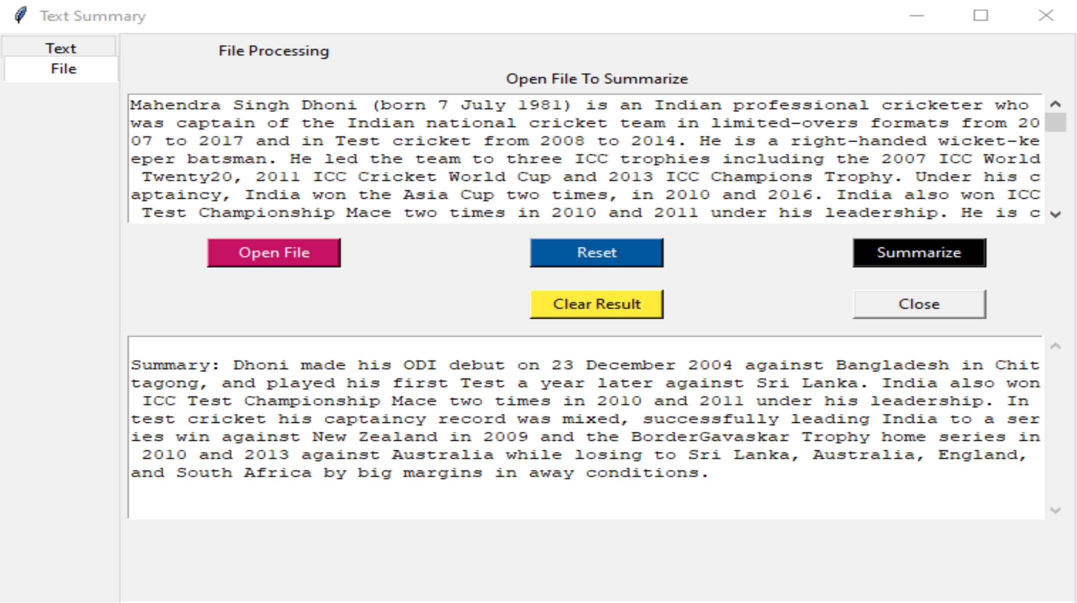
Output screen for text

Case 2: summarize for file



```
Mahendra Singh Dhoni (born 7 July 1981) is an Indian professional cricketer who
was captain of the Indian national cricket team in limited-overs formats from 20
07 to 2017 and in Test cricket from 2008 to 2014. He is a right-handed wicket-ke
eper batsman. He led the team to three ICC trophies including the 2007 ICC World
 Twenty20, 2011 ICC Cricket World Cup and 2013 ICC Champions Trophy. Under his c
aptaincy, India won the Asia Cup two times, in 2010 and 2016. India also won ICC
 Test Championship Mace two times in 2010 and 2011 under his leadership. He is c
```

Input screen for file



```
Mahendra Singh Dhoni (born 7 July 1981) is an Indian professional cricketer who
was captain of the Indian national cricket team in limited-overs formats from 20
07 to 2017 and in Test cricket from 2008 to 2014. He is a right-handed wicket-ke
eper batsman. He led the team to three ICC trophies including the 2007 ICC World
 Twenty20, 2011 ICC Cricket World Cup and 2013 ICC Champions Trophy. Under his c
aptaincy, India won the Asia Cup two times, in 2010 and 2016. India also won ICC
 Test Championship Mace two times in 2010 and 2011 under his leadership. He is c
```

```
Summary: Dhoni made his ODI debut on 23 December 2004 against Bangladesh in Chit
tagong, and played his first Test a year later against Sri Lanka. India also won
 ICC Test Championship Mace two times in 2010 and 2011 under his leadership. In
test cricket his captaincy record was mixed, successfully leading India to a ser
ies win against New Zealand in 2009 and the BorderGavaskar Trophy home series in
 2010 and 2013 against Australia while losing to Sri Lanka, Australia, England,
and South Africa by big margins in away conditions.
```

Output screen for file

Conclusion:

The purpose of summarization of text is to interpret the source text as a condensed edition with the words preserved, to express the contents of a document in a succinct way in order to satisfy customers’ needs. The main aim of this project is to collect the important keywords from the provided agreements or the manuals and give the user a brief and short summary for the input.

Future Scope:

Future work aims at improving the extraction on important keywords, a framing the sentences based on the extracted keywords. This is called abstractive summarization. we’ll discuss the descriptive methodology of overview text where profound thinking plays a major part. Summarization can also be worked on with the help of machine learning, deep learning techniques.Two reasonable recommendations may be to change the topic-oriented outline framework for news and blogs and extend the analysis machine-leaning methods. Summaries of the news reports are often more descriptive and accessible to users. Research on subject modeling and resuming in the area of social networking would be more relevant in the future.