Extractive text summarization of image extracted text

MSc Research Project
Data Analytics

Sufal Addya
Student ID: X18180825

School of Computing
National College of Ireland

Supervisor: Prof. Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sufal Addya
Student ID:	X18180825
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Prof. Christian Horn
Submission Due Date:	28/09/2020
Project Title:	Extractive text summarization of image extracted text
Word Count:	5926
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	28th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Extractive text summarization of image extracted text

Sufal Addya
X18180825

Abstract

Text summarization is a huge field in text analytics, research is tried to propose an unique approach to find text summarization from images. Optical character recognition using PyTesseract with OpenCV perform very well to extract text from images and research applied two unsupervised extractive text summarization algorithms Textrank and TF-IDF algorithms on that text to find a meaningful summary. This proposed sequence of program pipeline produce a very attractive output with can be applied in future to implement in making text summarization application. Here, Tesseract with OpenCV perform outstanding to extract the text and two extractive summarization algorithm produce a meaningful extractive summary successfully but evaluating accuracy of generated summary is a challenging part of this research which needs to overcome in future.

1 Introduction

Data science is a data-driven decision making process. In the early stage of digital evolution the data was mainly generated from PCs, but in the later stage data is producing from plenty of digital devices. For this huge amount of data, humans are flooding with the information and records, because of drastic growth in big-data and internet. To deal with this huge structured and unstructured data there are several approach in data science, in that text analytics is focused on natural language processing and natural language generation. The main aim of this proposed research is text summarization of the extracted data from image which is a combination approach of machine learning and natural language processing techniques.

Text summarization is a technique to find out meaningful summary from a lengthy pieces of text. Today's world humans are surrounded by huge amounts of data in the digital space, automatic text summarization techniques can help to get a short and meaningful summary which can help human to understand the text in less time, also increase the quality and quantity of information in the short piece of summarized text (Babar et al.; 2013). There are many techniques for text summarization in natural language processing (NLP) domain. Main two techniques are,

1. Extraction-based summarization
2. Abstraction-based summarization

Extractive text summarization is a process of summarization which pull the main points from source text and merge them to make a meaningful summary. Abstractive text summarization is a process which paraphrase the source document and shorten the text into a meaningful summary. Extractive text summarization is totally dependent on the original text source that takes key sentences or part of that from original text to make a summary with less grammatical mistakes.

On the other hand, text extraction is also a part of text analytics, which can be done from the image. Automatic text recognition and extraction from an image is a part of natural language processing. Optical character recognition(OCR) is core technique behind the text extraction from an image. OCR technology can collect the text data from any format of an image and can be used for the NLP techniques. Along with this text cleaning is the key process in natural language processing. After extracting data from data image to further processing of the data is coupled by the text cleaning process. To get an accurate output in the natural language processing, text pre-processing part will be in lead role.

The proposed research is based on combination of natural language processing (NLP) and optical character recognition (OCR) techniques. Objective of this research is mainly focused on extractive text summarization using different summarization method. This research is extracting data from an image using the OCR techniques. After getting unstructured text data from an image, proposed research will apply pre-processing and text cleaning techniques to get a structured data for the further implementation of text summarization technique to achieve a meaningful and short text. The proposed research project is divided into three parts which are

1. Text extraction from an image
2. Text pre-processing of that extracted text
3. Applied extractive text summarization techniques on that text to get a summary.

The pipeline of this three section is the key of this research project. Research is using python as a programming language to implement the processes. The proposed research is using Python-tesseract to get extract the text from an image, then natural language toolkit (NLTK) is applied for text pre-processing.

OCR	Text pre-processing	Text summarization algorithm
Pytesseract	Natural language toolkit (NLTK)	Textrank algorithm
OpenCV	Regular expression (RE)	TF-IDF algorithm

Table 1: Table of applied techniques for OCR, text pre-processing and text summarization.

1.1 Research Question

How efficient are the two unsupervised extractive summarization algorithms in summarizing the text from given image?

1.2 Research Objectives and Contribution

The objective of this research is to produce a meaningful summary using unsupervised extractive text summarization algorithms on the image extracted text using Tesseract and OpenCV. This research of extractive text summarization from images can contribute in text analytics and also can go a step ahead in making text summarization application in an unique way to make human life more reliable and time saving in this huge digital data world.

2 Related Work

2.1 Introduction

This chapter of the report is mainly focused on to discuss about the previous works, along with strengths and limitations of the implemented techniques which have already done in the past upon the related topic. Also, this section is trying to present the strengths and limitations of the previous techniques for the chosen topic.

2.2 Optical character recognition

The text extraction approach has been done in different way by the researchers. The optical character recognition process is the most frequent and popular process to do the text extraction from an image. In the year 2007, researchers gave an brief about the Tesseract OCR engine. The open-source Tesseract engine for OCR was developed by the HP to use in different digital devices. The architecture of OCR engine divided into two part, first is text recognition and second is checking the pattern of the text in papers. There are several techniques included in the Tesseract OCR engine which are mentioned by the researchers. Line and word finding, Word recognition, Static character classifier, linguistic analysis, adaptive classifier all are techniques of OCR engine to get a better and more accurate result. In mentioned approach of optical character recognition researchers mentioned about the Tesseract OCR to increase the accuracy of the text extraction engine (Smith; 2007).

In the recent dates the application of OCR is growing rapidly to improve the digital technology. Modern world is improving and the use of image in different sectors are also increasing. Image can carry important data, to extract the data from the image is an challenging part in digital world. Textual data are available from different resources like newspapers, images, notes etc. In 2019, Pawar and their team discussed more briefly about the text extraction techniques in a research paper from the image using the Tesseract OCR engine. OCR engine is a section of artificial intelligence which is an advanced text extraction method from image and different sources. At the initial stage, the OCR was build upon the convolution neural network, after the development of the OCR, now its mainly based on Long Short Term Memory (LSTM) which is part of Recurrent neural network. Tesseract OCR is an open source and best for the handwritten text recognition and extraction. The discussed research is mainly focused on the Tesseract OCR model and study revealed about different approach of OCR model. Those are Connected component based method, Sliding window based method, Hybrid method, Edge based method, Color based method, Texture based method, Corner based method, Stroke based method, Semi automatic ground truth generation based method. Also, researchers discussed about the advantages of the OCR model, as a data image can take more storage than a text and along with this people still prefer the text document as data more than an image, which made the OCR process more valuable in the recent days. Also, OCR can be used to make text data into speech after extracting from the image which can be valuable for the blind people (Pawar et al.; 2019).

The extraction of image text in a perfect manner is the challenging task. Also, different images can be in different format, size and colour which can be a difficult task for the OCR. In a study researchers introduced an approach which is focused on to extract the image more clearer. Here, researchers introduced the method of image processing than cropping and then text extraction. As a result this approach of Tesseract OCR is applied on different type of images and produced accurate text (Chawla et al.; 2020).

Researcher R.R. Palekar and his team mentioned about one of the challenging part in text detection and then extraction from the image in 2017. To develop the result of the Tesseract OCR researchers used OpenCV and OCR together. OpenCV used to do image processing and Tesseract OCR used to extract the text from the image which is comparatively more accurate than the simple OCR engine. After using the following process researchers identified that the text processing before text extraction is an important part. Also, revealed that text processing before text extraction with the different image achieved more accurate result. Researchers used real time car number plate to identify the accuracy of the model. Here, OpenCV with the Tesseract OCR produced perfect result (Palekar et al.; 2017).

2.3 Text summarization

There are mainly two techniques for the text summarization, extractive and abstractive text summarization. In 2010 Vishal and three other researchers introduced different approaches of extractive text summarization technique. There are several features to do the text summarization in the extractive manner discussed by the researchers which were Content word, Title word, Sentence location, Sentence Length, Proper Noun, Upper-case word, Cue-Phrase, Biased Word, Font based, Pronouns, Sentence-to-Sentence Cohesion, Sentence-to-Centroid Cohesion, Occurrence of non-essential information, Discourse analysis feature. Also, in the mentioned study, researchers introduced several method of extractive text summarization. Those mentioned methods were Term Frequency-Inverse Document Frequency (TF-IDF), Cluster based method, Graph theoretic approach, Machine Learning approach, LSA method, Context obtained text summarization method and Text summarization with neural networks. Extractive summarization can be done in two ways supervised and unsupervised, in the machine learning and neural network approach need a large dataset to train the model but unsupervised techniques don't need training part. In the recent days researchers are trying to develop the supervised and unsupervised extractive text summarization techniques. In these papers researchers made a survey upon the extractive text summarization and gave a brief overview of extractive text summarization technique (Gupta and Lehal; 2010) (El-Refaiy et al.; 2018).

The popularity of text summarizers are growing day by day because of the growth in the data. In natural language processing text summarization has a deep impact, which can identify the highest priority data from the text and can make a summary from that selected data. In 2016, Christian and a group of researchers used TF-IDF model to make summary which was extractive based text summarization. The focus area was to produce the summary of text and compared with different online summarizer. The used TF-IDF model for the text summarization produced 67 percent accuracy which was better than the other online sources. Researchers used F-measure score as a standard differentiation method. In mentioned research, researchers followed a pattern to find out the summary. study proposed TF-IDF model in multiple steps, Text processing of given text, then measures the value TF-IDF, then calculated the score of each sentences with the TF-IDF value, then generate the highest valued sentences together to produced a summary. Researchers mentioned about the NLTK package for the text processing and the calculation of TF-IDF score. Study revealed that TF-IDF model can be a powerful extractive text summarization model to generate summary with a better accuracy (Christian et al.;

2016).

Another research was based on the multi-document summarization based on the TF-IDF and centroid based K-means clustering (TF-IDF: OBC). The base concept behind the research was dimentionality reduction of extracted features by K-means using TF-IDF score. Researchers used various similarity measure to find out the similarity between the sentences, after that with the TF-IDF score study create the cluster for further text summarization. The proposed method of text summarization produced a better and appreciative result (ROBINSON and SARAVANAN; 2019).

To move forward with the development of extractive text summarization, researchers are trying to develop different algorithm for the summarization. In a research of extractive text summarization researchers introduced another automatic extractive text summarization algorithm which was Textrank algorithm. In the era of data flood summarization is one of the main part in natural language processing. Researchers gave a brief about the Textrank method, Textrank method is an unsupervised and extractive method of summarization which only took the data from the source text to make summary. In 2019 a group of researchers studied about textrank algorithm for text summarization researchers made graph with the nodes and similarity between sentences. Here, researchers used cousine similarity to produce weightage to different word in the sentence and took the highest weight sentences from that and made summarization by joining that sentences. The result of the textrank algorithm gave a decent summary but researchers also said about some drawback about duplication of sentences and can improve the result with different similarity check algorithms of sentences (Mallick et al.; 2019).

There are several similarity check approach in textrank algorithm, in 2017, karlsson and simon researchers produce semantic folding method to find the similarity for textrank algorithm to get better result in text summarization. The study also added different pre-processing task including stop words removal, part-of-speech tag and stemming to get better result and used gold-standard summarization dataset to find the accuracy of summary with ROUGE-1 method. The approach of textrank algorithm for text summarization using semantic folding gave better result than its state of the art Karlsson (2017).

In 2019 a researcher Neha Joshi developed an unique technique to produce a text summarization from an image. The research used tesseract with python to extract the text from image and produce summary of the text. The mentioned research following a pipeline of text extraction and then text summarization of that text (Joshi; 2019). The mentioned research is the state of the art of following project where this research trying to go a bit forward by applying Tesseract with OpenCV and also applied unsupervised extractive text summarization to produce meaningful summary. These changes can develop the research a step ahead and can full fill the future objective which is making an application of extractive text summary.

3 Architecture and Methodology of Text Summarization

3.1 Introduction

The proposed research is on text analytics which is mainly focused on text summarization of extracted text from an image. Here the process of the research is mainly divided in to three parts, text extraction form an image, text pre-processing and then applying different extractive summarization algorithm on that cleaned text. This research project design is based on pipeline architecture which includes three segment in it. The brief about design specification and architecture is discussed in section 4. Before discussing the architecture, another important part of project is scientific methodology. Research need to be follow a particular technique to achieve the goal. This research is mainly focused on to develop a pipeline of image processing and text analytics sequentially, after a deep research concluding that the waterfall model is a best way to produce a result.

3.2 Technical Requirement and Data Preparation

There are different technologies and Python programming language are used to get realistic text summarization. Along with this different libraries and machine learning algorithms are applied in different stage of this research. Python programming language is used in this whole research. In the application layer, PyTesseract and OpenCV libraries are used to extract text from an image, NLTK and regular expression packages are used to pre-processing and after cleaning the text data different extractive summarization algorithms are applied to get realistic text summarization. After giving input in input queue data processed with three segment and produced result in output queue.

3.3 Approach for Methodology

Every project should follow a proper scientific methodology to achieve a successful target. This text summarization project is following Waterfall methodology to get a related and meaningful summary. Waterfall methodology is a project management approach which follows a linear method to achieve the goal. The proposed project is trying to build a pipeline for OCR and text analytics which can be applied to make software about text summarization for extracted text from image. After a brief research and understanding of project requirement, concluding that Waterfall methodology is the most related and applicable methodology. Waterfall methodology is a linear project management method with several steps, which are Requirement, Design, Implementation, Verification and Maintenance.

1. Requirement: The main characteristic of Waterfall methodology is to understand the requirements of the project. Requirement of client and project need to be understood in this initial stage of followed methodology, after this stage no other customer interaction would be valid till the completion of the project. Now, this proposed research goal is finding a meaningful text summarization of image extracted text. It means there are two goals need to be followed, text extraction from an image and after cleaning the text need to find a text summary of that text. Now after understanding requirements need to follow the design of the project which is second step of proposed methodology.

2. Project Design: This section refers architecture of the proposed project. Design can be divided into two parts, logical design, and practical design. Logical design is mainly based on theoretical and physical design is the practical implementation of the project. In this project design the practical implementation is divided into three parts. After understanding the project requirements, the main part is project design how can fulfil the project requirement. To understand the design flow of the project, refer Figure 1.

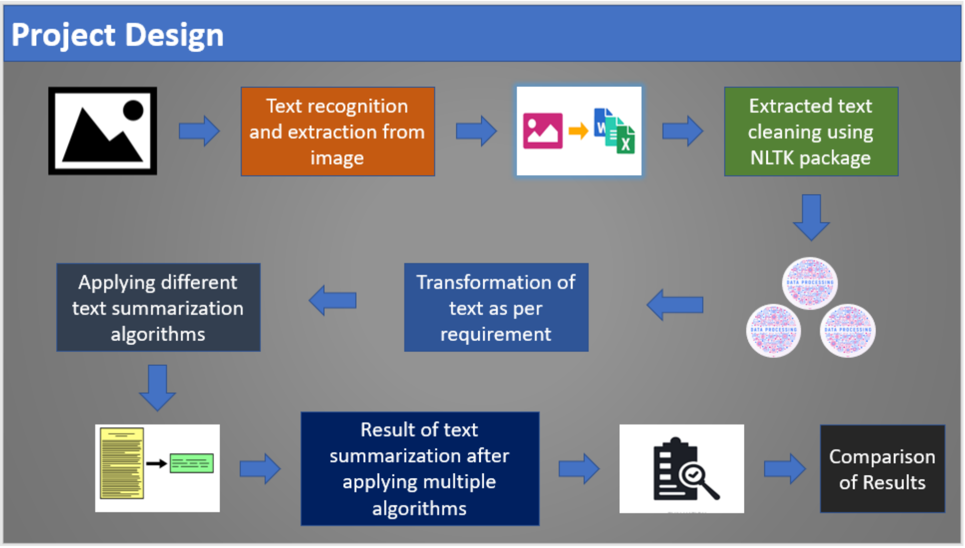

Project Design
Text recognition and extraction from image
Extracted text cleaning using NLTK package
Transformation of text as per requirement
Applying different text summarization algorithms
Result of text summarization after applying multiple algorithms
Comparison of Results

Figure 1: Design of the proposed research

3. Implementation: After validating the design of the project, the third most important part is implementation. This base of this stage is technical, where all the algorithms and different methods is executed to achieve the goal of the project. In this project, as per requirement the total implementation done in python and image to text which is done by optical character recognition, extracted text cleaning using natural language toolkit (NLTK) and then different unsupervised extractive text summarization algorithms like Textrank, TF-IDF and are implemented to get a summary. These are the implementation part of this project.

4. Verification: This stage of methodology comes after the implementation. Verification section is used to verify the justification of the implementation by using different techniques. This is also a major part of testing any project who has already implemented with certain techniques. The verification of this project is done using different type of images and compared result in between different machine learning algorithms. This section is also a part of testing the implemented part.

5. Maintenance: Last and final part of the project methodology is maintaining any project or any application. To maintain any project crucial part is keep updated. In this project this part is not fully functional. But the update of text summarization algorithms and different text cleaning and modification of OCR come under the maintenance.

The aim of this project is to make an interactive application by using OCR and text analytics. This project followed the waterfall methodology which is appropriate for software development life cycle. Research follows each and every step of the proposed methodology to achieve successful result (Petersen et al.; 2009).

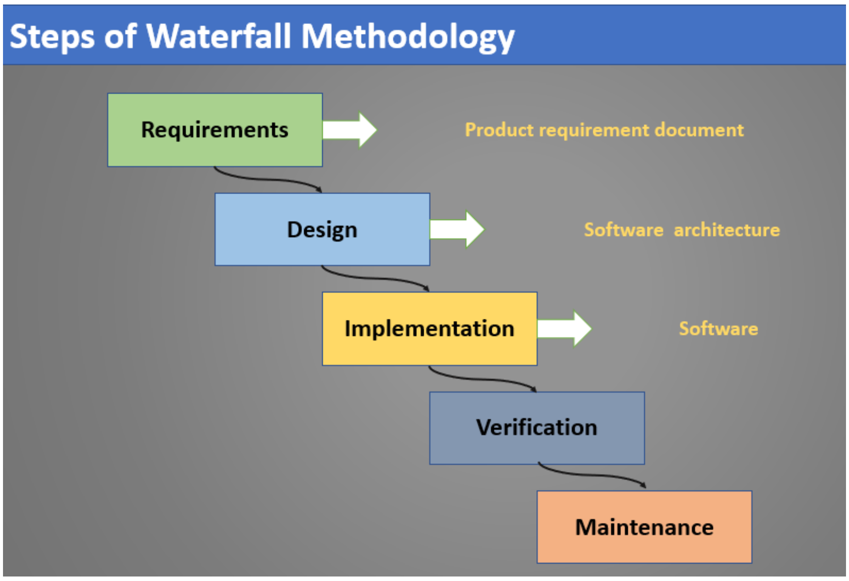


Figure 2: Waterfall methodology

3.4 Conclusion

After analysing and discussing different step of the methodology concluding that this proposed research using waterfall methodology to achieve a proper result. Along with this project is based on Python programming language and project divided into two major sections which are optical character recognition and text analytics. Applied algorithms and techniques are PyTesseract, OpenCV, NLTK, regular expression, Textrank and TFIDF algorithms.

4 Design Specification

This research project is mainly combination of image processing and text analytics. The goal of this project is to provide a short and meaningful extractive summary of given text. Design of this project follows a sequence to make a pipeline which contains extraction of text from an image then, after pre-processing of that text different extractive text summarization models would be applied to produce a meaningful and short summary. After understanding the project flow and design of total project concluding that the project is following a pipeline architecture which allows system to execute different segment in decomposed manner. Pipeline architecture is a faster and cost-effective computer architecture system without duplicating cost of the total workflow. The basic concept of pipeline architecture is, it takes input from input queue and produce output in its output queue (Bienia and Li; 2010) (Ramamoorthy and Li; 1977). This research project also takes an image as an input and produces an extractive text summarization as an output after applying different techniques in between that pipeline on that input image. Here the pipeline of this project is divided into three segments in a sequence which are text extraction from an image (OCR), text pre-processing of extracted text and applying extractive text summarization algorithm to produce a summary. These three parts of this project is managed in a sequential manner in a pipeline architecture and each part is dependent on previous section output. Figure 3 shows the flow of pipeline architecture to understand the project architecture.

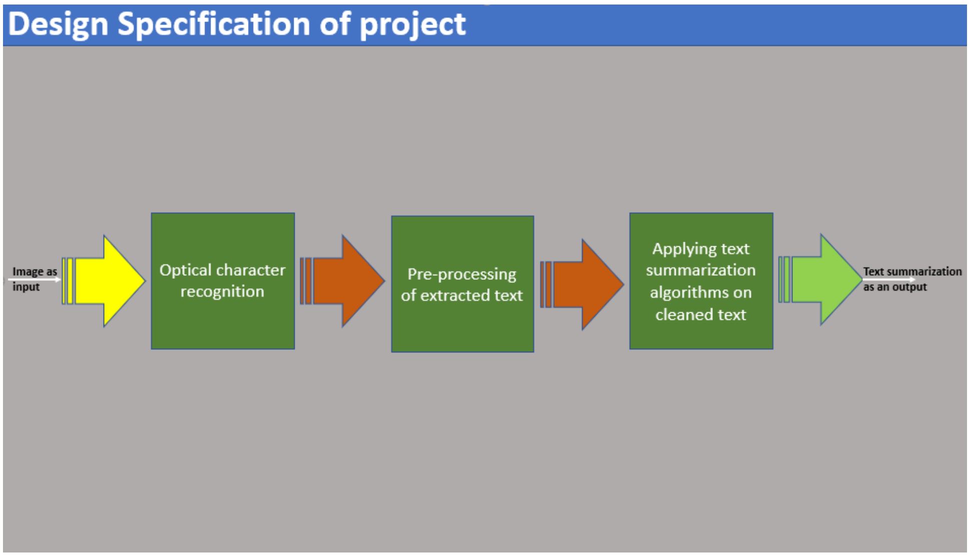

Design Specification of project

Image as input
Optical character recognition
Pre-processing of extracted text
Applying text summarization algorithms on cleaned text
Text summarization as an output

Figure 3: Pipeline architecture

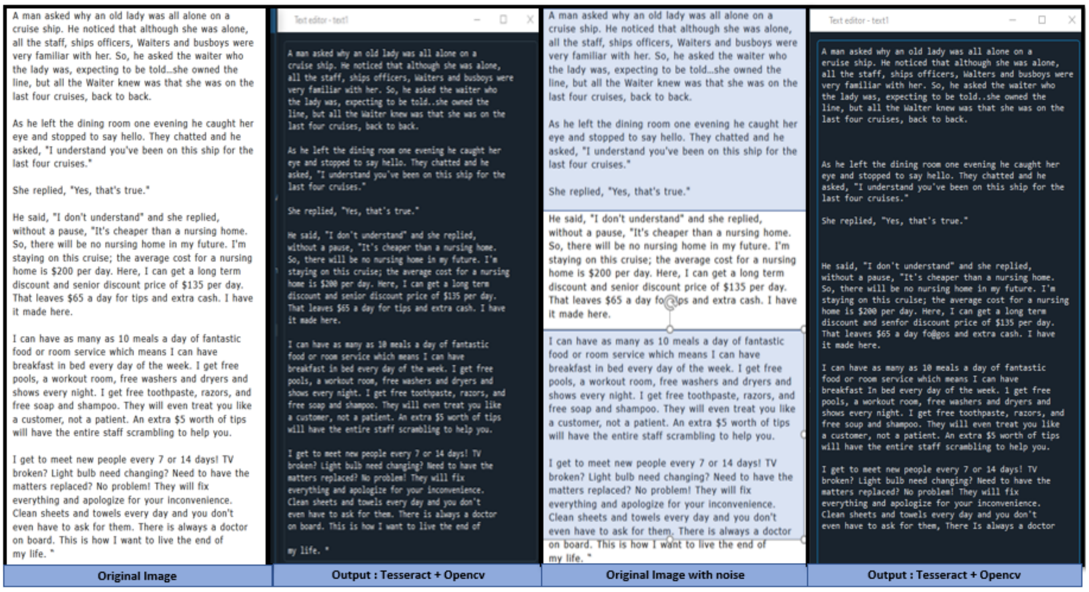
5 Implementation

This research project of extractive text summarization is a combination of two different parts which already described in section 4. This section of the research project is providing a brief description about the execution and implementation of different algorithms and techniques to find out meaningful extractive text summary of extracted text from image. Along with the summary this project is trying to compare extractive summarization algorithms which can be applied after the OCR technique.

5.1 Implementation of OCR

The initial part of this project is based on optical character recognition (OCR) which is a technology to convert any image, PDF and document into an editable text file. The OCR part of this project is applied on the images which can be in different formats. The optical character recognition technique is implemented by python programming language, Tesseract and OpenCV. Here, Tesseract is a text recognition engine and PyTesseract is Tesseract binding for python which is used in this project for image processing. Tesseract can produce adoptable output in general image extraction but in worse scenario where an image can be noisy and in deform order, then project need a tuning process of Tesseract to develop the quality of output which was generated by Tesseract. In this situation OpenCV can act as an important part to tune the output of Tesseract technique. To overcome these noisy image and improper format of image problem, this project implementing PyTesseract with opencv, which can tune PyTesseract engine and can help to produce a good output. OpenCV is used for the text detection, noise reduction from an image and can develop the output from image to help Tesseract for extracting the text from an image. So, implementation of first part of the pipeline which is text recognition and extraction from the image is implemented with combination of PyTesseract and OpenCV.

Also, to tune the result of OCR, lots of operation have been done like grayscale conversion, noise removal on the selected image before giving to the optical character recognition for further processing.


Original Image
Text editor - text1
Output : Tesseract + Opencv
Original Image with noise
Text editor - text1
Output : Tesseract + Opencv

A man asked why an old lady was all alone on a cruise ship. He noticed that although she was alone, all the staff, ships officers, Waiters and busboys were very familiar with her. So, he asked the waiter who the lady was, expecting to be told...she owned the line, but all the Waiter knew was that she was on the last four cruises, back to back.

As he left the dining room one evening he caught her eye and stopped to say hello. They chatted and he asked, "I understand you've been on this ship for the last four cruises."

She replied, "Yes, that's true."

He said, "I don't understand" and she replied, without a pause, "It's cheaper than a nursing home. So, there will be no nursing home in my future. I'm staying on this cruise; the average cost for a nursing home is $200 per day. Here, I can get a long term discount and senior discount price of $135 per day. That leaves $65 a day for tips and extra cash. I have it made here.

I can have as many as 10 meals a day of fantastic food or room service which means I can have breakfast in bed every day of the week. I get free pools, a workout room, free washers and dryers and shows every night. I get free toothpaste, razors, and free soap and shampoo. They will even treat you like a customer, not a patient. An extra $5 worth of tips will have the entire staff scrambling to help you.

I get to meet new people every 7 or 14 days! TV broken? Light bulb need changing? Need to have the matters replaced? No problem! They will fix everything and apologize for your inconvenience. Clean sheets and towels every day and you don't even have to ask for them. There is always a doctor on board. This is how I want to live the end of my life. "

```
A man asked why an old lady was all alone on a
cruise ship. He noticed that although she was alone,
all the staff, ships officers, Waiters and busboys were
very familiar with her. So, he asked the waiter who
the lady was, expecting to be told..she owned the
line, but all the Waiter knew was that she was on the
last four cruises, back to back.

As he left the dining room one evening he caught her
eye and stopped to say hello. They chatted and he
asked, "I understand you've been on this ship for the
last four cruises."

She replied, "Yes, that's true."

He said, "I don't understand" and she replied,
without a pause, "It's cheaper than a nursing home.
So, there will be no nursing home in my future. I'm
staying on this cruise; the average cost for a nursing
home is $200 per day. Here, I can get a long term
discount and senior discount price of $135 per day.
That leaves $65 a day for tips and extra cash. I have
it made here.

I can have as many as 10 meals a day of fantastic
food or room service which means I can have
breakfast in bed every day of the week. I get free
pools, a workout room, free washers and dryers and
shows every night. I get free toothpaste, razors, and
free soap and shampoo. They will even treat you like
a customer, not a patient. An extra $5 worth of tips
will have the entire staff scrambling to help you.

I get to meet new people every 7 or 14 days! TV
broken? Light bulb need changing? Need to have the
matters replaced? No problem! They will fix
everything and apologize for your inconvenience.
Clean sheets and towels every day and you don't
even have to ask for them. There is always a doctor
on board. This is how I want to live the end of

my life. "
```

A man asked why an old lady was all alone on a cruise ship. He noticed that although she was alone, all the staff, ships officers, Waiters and busboys were very familiar with her. So, he asked the waiter who the lady was, expecting to be told...she owned the line, but all the Waiter knew was that she was on the last four cruises, back to back.

As he left the dining room one evening he caught her eye and stopped to say hello. They chatted and he asked, "I understand you've been on this ship for the last four cruises."

She replied, "Yes, that's true."

He said, "I don't understand" and she replied, without a pause, "It's cheaper than a nursing home. So, there will be no nursing home in my future. I'm staying on this cruise; the average cost for a nursing home is $200 per day. Here, I can get a long term discount and senior discount price of $135 per day. That leaves $65 a day fo ips and extra cash. I have it made here.

I can have as many as 10 meals a day of fantastic food or room service which means I can have breakfast in bed every day of the week. I get free pools, a workout room, free washers and dryers and shows every night. I get free toothpaste, razors, and free soap and shampoo. They will even treat you like a customer, not a patient. An extra $5 worth of tips will have the entire staff scrambling to help you.

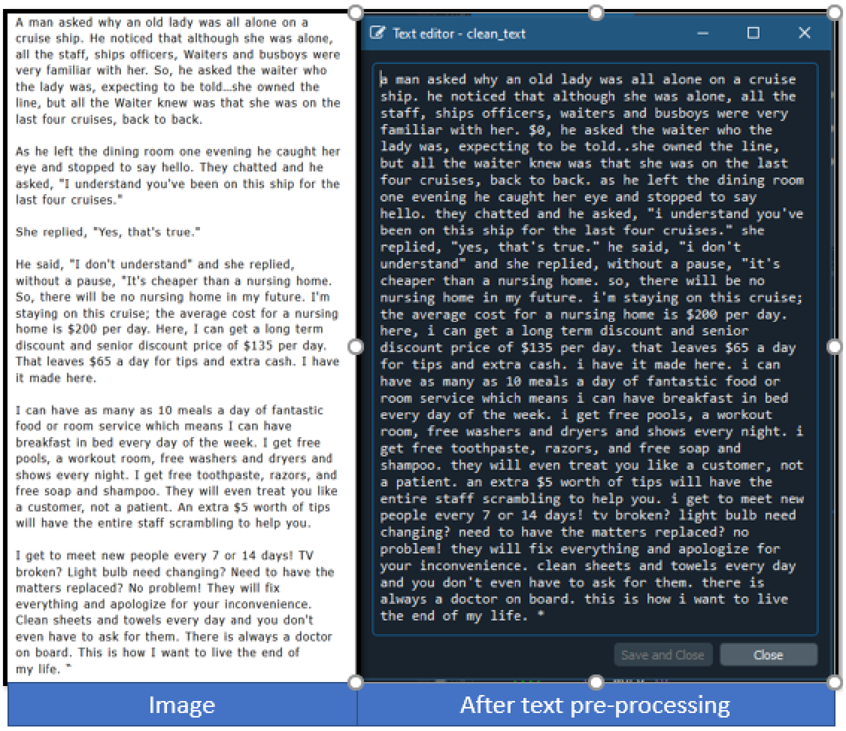
I get to meet new people every 7 or 14 days! TV broken? Light bulb need changing? Need to have the matters replaced? No problem! They will fix everything and apologize for your inconvenience. Clean sheets and towels every day and you don't even have to ask for them. There is always a doctor on board. This is how I want to live the end of my life. "

```
A man asked why an old lady was all alone on a
eruise ship. He noticed that although she was alone,
all the staff, ships officers, Walters and busboys were
very familiar with her. So, he asked the waiter who
the lady was, expecting to be told..she owned the
line, but all the Walter knew was that she was on the
last four cruises, back to back.



As he left the dining room one evening he caught her
eye and stopped to say hello. They chatted and he
asked, "I understand you've been on this ship for the
last four cruises."

She replied, "Yes, that's true."



He said, "I don't understand" and she replied,
without a pause, "It's cheaper than a nursing home.
So, there will be no nursing home in my future. I'm
staying on this crulse; the average cost for a nursing
home is $200 per day. Here, I can get a long term
discount and senfor discount price of $135 per day.
That leaves $65 a day fo@gos and extra cash. I have
it made here.

I can have as many as 10 meals a day of fantastic
food or room service which means I can have
breakfast In bed every day of the week. 1 get free
pools, a workout room, free washers and dryers and
shows every night. I get free toothpaste, razors, and
free soup and shampoo. They will even treat you like
a customer, not a patient. An extra $5 worth of tips
will have the entire staff scrambling to help you.

I get to meet new people every 7 or 14 days! TV
broken? Light bulb need changing? Need to have the
matters replaced? No problem! They will fix
everything and apologize for your inconvenience.
Clean sheets and towels every day and you don't
even have to ask for them, There Is always a doctor
```

Figure 4: Applied Tesseract and OpenCV on cleaned and noisy images

5.2 Implementation of text pre-processing

The second phrase of implementation implemented after the optical character recognition technique which is text cleaning and pre-processing. Principle part of this research was text pre-processing after extraction of text from an image. Image text could be in any format and any style, so before applying any text summarization algorithm text pre-processing is important part. This text cleaning and pre-processing was established in between two major sections of the project. Research applied natural language toolkit (NLTK) and regular expression for text pre-processing and cleaning before applying text summarization algorithms. There were several text pre-processing techniques applied in this section which were extra space removal, proper alignment selection, all lowercase alphabets transformation, joining of sentences were done by regular expression along with this word tokenize, sentence tokenization were applied using the NLTK. After implemented different text pre-processing techniques, processed text would go for text summarization algorithm implementation.



A man asked why an old lady was all alone on a cruise ship. He noticed that although she was alone, all the staff, ships officers, Waiters and busboys were very familiar with her. So, he asked the waiter who the lady was, expecting to be told...she owned the line, but all the Waiter knew was that she was on the last four cruises, back to back.

As he left the dining room one evening he caught her eye and stopped to say hello. They chatted and he asked, "I understand you've been on this ship for the last four cruises."

She replied, "Yes, that's true."

He said, "I don't understand" and she replied, without a pause, "It's cheaper than a nursing home. So, there will be no nursing home in my future. I'm staying on this cruise; the average cost for a nursing home is $200 per day. Here, I can get a long term discount and senior discount price of $135 per day. That leaves $65 a day for tips and extra cash. I have it made here.

I can have as many as 10 meals a day of fantastic food or room service which means I can have breakfast in bed every day of the week. I get free pools, a workout room, free washers and dryers and shows every night. I get free toothpaste, razors, and free soap and shampoo. They will even treat you like a customer, not a patient. An extra $5 worth of tips will have the entire staff scrambling to help you.

I get to meet new people every 7 or 14 days! TV broken? Light bulb need changing? Need to have the matters replaced? No problem! They will fix everything and apologize for your inconvenience. Clean sheets and towels every day and you don't even have to ask for them. There is always a doctor on board. This is how I want to live the end of my life. "

```
a man asked why an old lady was all alone on a cruise
ship. he noticed that although she was alone, all the
staff, ships officers, waiters and busboys were very
familiar with her. $0, he asked the waiter who the
lady was, expecting to be told..she owned the line,
but all the waiter knew was that she was on the last
four cruises, back to back. as he left the dining room
one evening he caught her eye and stopped to say
hello. they chatted and he asked, "i understand you've
been on this ship for the last four cruises." she
replied, "yes, that's true." he said, "i don't
understand" and she replied, without a pause, "it's
cheaper than a nursing home. so, there will be no
nursing home in my future. i'm staying on this cruise;
the average cost for a nursing home is $200 per day.
here, i can get a long term discount and senior
discount price of $135 per day. that leaves $65 a day
for tips and extra cash. i have it made here. i can
have as many as 10 meals a day of fantastic food or
room service which means i can have breakfast in bed
every day of the week. i get free pools, a workout
room, free washers and dryers and shows every night. i
get free toothpaste, razors, and free soap and
shampoo. they will even treat you like a customer, not
a patient. an extra $5 worth of tips will have the
entire staff scrambling to help you. i get to meet new
people every 7 or 14 days! tv broken? light bulb need
changing? need to have the matters replaced? no
problem! they will fix everything and apologize for
your inconvenience. clean sheets and towels every day
and you don't even have to ask for them. there is
always a doctor on board. this is how i want to live
the end of my life. "
```

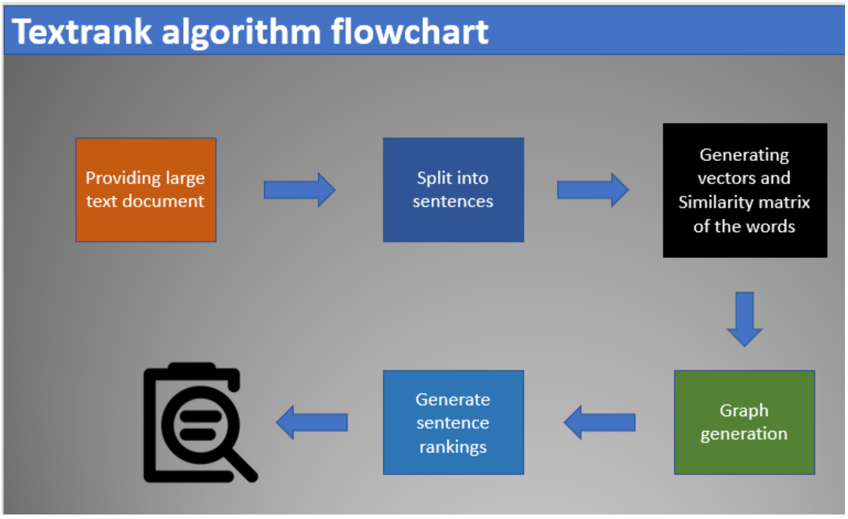
Figure 5: Implementation of text cleaning on extracted text from an image

5.3 Implementation of text summarization algorithms

After processing the text last and final part of the project is implementation different unsupervised extractive text summarization algorithms. This project was based on extractive text summarization which is mainly focused on the important sentences finding from given text and providing a meaningful summary. There were several extractive summarization techniques to find a meaningful summary. This research project applied two text summarization algorithm which were Textrank algorithm and TF-IDF text summarization algorithm. Providing step by step implementation and discussing about the algorithms.

5.3.1 Textrank algorithm

Textrank algorithm is extractive and unsupervised text summarization technique which can provide a meaningful summary from the given text. This algorithm of extractive text summarization were used word embedding to find vector in between each and every text. Here, research is being used Glove word embedding. Along with this textrank used to made a similarity matrix between the words and generate score for sentences and after that choose top ranked sentences which was an extractive summary of given text.



Textrank algorithm flowchart

Providing large text document

Split into sentences

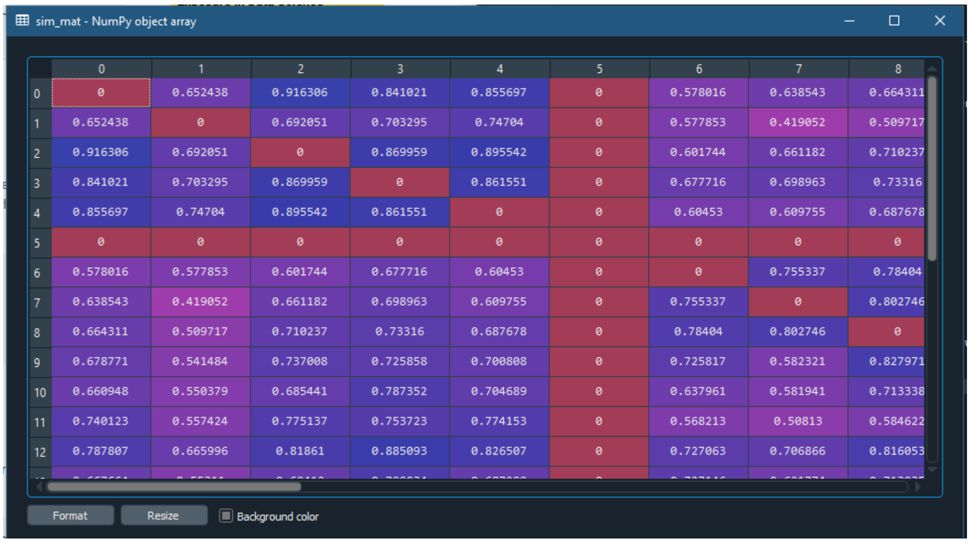
Generating vectors and Similarity matrix of the words

Graph generation

Generate sentence rankings

Figure 6: Flowchart of Textrank algorithm

In this Textrank algorithm after cleaning and splitting the text into sentences next step was vector generation of sentences using word embedding and then made similarity matrix to find similarity between sentences. After generating similarity matrix research implement Textrank algorithm to get summary based on the sentence score. Also, took highest score sentences to make summary of project. Providing image of the similarity vector matrix which helped to apply text summarization algorithm. Here, similarity metric used to determine how similar two entities are irrespective of their size and produce a score in a matrix to apply the Textrank algorithm. Figure 7 shows the similarity matrix of different word in sentences.



sim_mat - NumPy object array

	0	1	2	3	4	5	6	7	8
0	0	0.652438	0.916306	0.841021	0.855697	0	0.578016	0.638543	0.664311
1	0.652438	0	0.692051	0.703295	0.74704	0	0.577853	0.419052	0.509717
2	0.916306	0.692051	0	0.869959	0.895542	0	0.601744	0.661182	0.710237
3	0.841021	0.703295	0.869959	0	0.861551	0	0.677716	0.698963	0.73316
4	0.855697	0.74704	0.895542	0.861551	0	0	0.60453	0.609755	0.687678
5	0	0	0	0	0	0	0	0	0
6	0.578016	0.577853	0.601744	0.677716	0.60453	0	0	0.755337	0.78404
7	0.638543	0.419052	0.661182	0.698963	0.609755	0	0.755337	0	0.802746
8	0.664311	0.509717	0.710237	0.73316	0.687678	0	0.78404	0.802746	0
9	0.678771	0.541484	0.737008	0.725858	0.700808	0	0.725817	0.582321	0.827971
10	0.660948	0.550379	0.685441	0.787352	0.704689	0	0.637961	0.581941	0.713338
11	0.740123	0.557424	0.775137	0.753723	0.774153	0	0.568213	0.50813	0.584622
12	0.787807	0.665996	0.81861	0.885093	0.826507	0	0.727063	0.706866	0.816053

Figure 7: Similarity matrix of Textrank algorithm

5.3.2 TF-IDF algorithm

The applied second text summarization algorithm is based on the product of Term Frequency and Inverse Document Frequency algorithm to find a extractive text summary. This second algorithm was used to the term frequency and inverse document frequency.

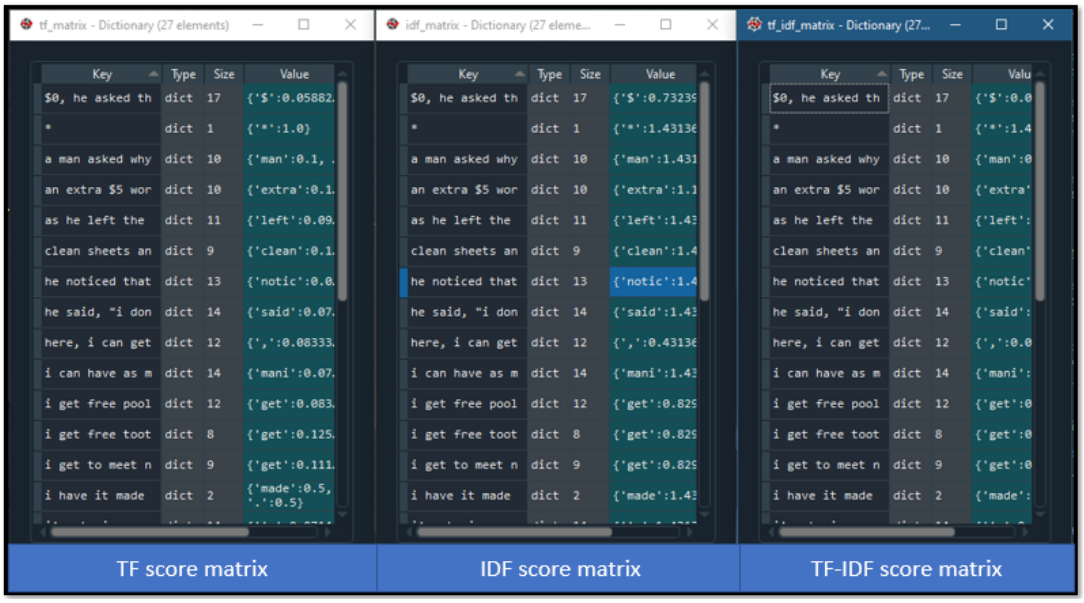
Term frequency Term frequency is a calculation, which measure how often a word appears in a document, divided by the total word of the document.

TF(p) = (Number of times term p appears in a single document)/(Total number of terms in that document)

Inverse document frequency The concept of term frequency is how common the word is and inverse document frequency is how unique a word is.

IDF(p) = ln(Total number of documents/Number of documents with term p in it)

Research was tried to implemented the TF-IDF model to find text summary. TF-IDF model calculated frequency of the word and in a document and made matrix using term frequency and matrix for the inverse term frequency. After calculating the score separately, model made matrix with each of them and implemented to generate the extractive summary. Along with these score, this algorithm need threshold calculation for getting the proper summary, here research find average of the sentences score as a threshold value which can be adjustable to get different size summary. Providing an image which showed that the matrix based on different TF-IDF value. Along with these three matrix TF-IDF model used to generate sentence score to find a proper extractive summary by adjusting the threshold value.


tf_matrix - Dictionary (27 elements)
idf_matrix - Dictionary (27 eleme...
tf_idf_matrix - Dictionary (27...

Key	Type	Size	Value
$0, he asked th	dict	17	{'$':0.05882.
*	dict	1	{'*':1.0}
a man asked why	dict	10	{'man':0.1, .
an extra $5 wor	dict	10	{'extra':0.1.
as he left the	dict	11	{'left':0.09.
clean sheets an	dict	9	{'clean':0.1.
he noticed that	dict	13	{'notic':0.0.
he said, "i don	dict	14	{'said':0.07.
here, i can get	dict	12	{',':0.08333.
i can have as m	dict	14	{'mani':0.07.
i get free pool	dict	12	{'get':0.083.
i get free toot	dict	8	{'get':0.125.
i get to meet n	dict	9	{'get':0.111.
i have it made	dict	2	{'made':0.5, '.':0.5}

TF score matrix

Key	Type	Size	Value
$0, he asked th	dict	17	{'$':0.73239
*	dict	1	{'*':1.43136
a man asked why	dict	10	{'man':1.431
an extra $5 wor	dict	10	{'extra':1.1
as he left the	dict	11	{'left':1.43
clean sheets an	dict	9	{'clean':1.4
he noticed that	dict	13	{'notic':1.4
he said, "i don	dict	14	{'said':1.43
here, i can get	dict	12	{',':0.43136
i can have as m	dict	14	{'mani':1.43
i get free pool	dict	12	{'get':0.829
i get free toot	dict	8	{'get':0.829
i get to meet n	dict	9	{'get':0.829
i have it made	dict	2	{'made':1.43

IDF score matrix

Key	Type	Size	Valu
$0, he asked th	dict	17	{'$':0.0
*	dict	1	{'*':1.4
a man asked why	dict	10	{'man':0
an extra $5 wor	dict	10	{'extra'
as he left the	dict	11	{'left':
clean sheets an	dict	9	{'clean'
he noticed that	dict	13	{'notic'
he said, "i don	dict	14	{'said':
here, i can get	dict	12	{',':0.0
i can have as m	dict	14	{'mani':
i get free pool	dict	12	{'get':0
i get free toot	dict	8	{'get':0
i get to meet n	dict	9	{'get':0
i have it made	dict	2	{'made':

TF-IDF score matrix

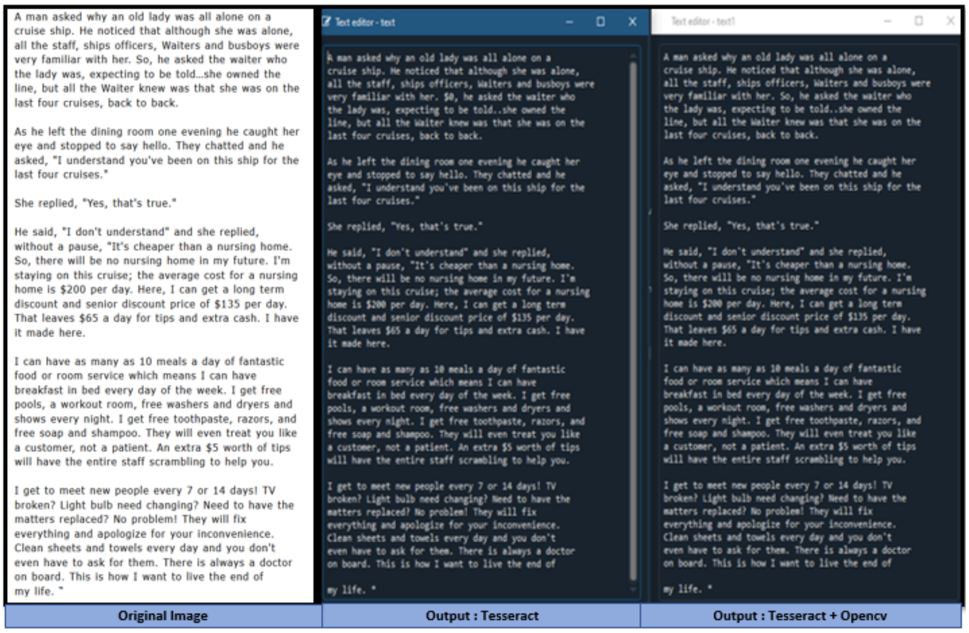
Figure 8: TF, IDF and TF-IDF matrix

6 Evaluation and Result

This section of the project report is providing a brief about evaluation and result of the applied techniques and algorithms to achieve the goal. As discussed this project implemented Tesseract and OpenCV for optical character recognition (OCR) and after cleaning two extractive text summarization algorithm were applied in this project, the output and evaluation of algorithms are being discussed in this section of the project report.

6.1 Applied OCR techniques- text extraction from an image

After applied optical character recognition techniques on the images research provided very efficient and accurate output. Tesseract and Tesseract with OpenCV both techniques were applied to extract the text from an image. Along with this to check ability of the both techniques, research applied both approach on noisy images and get very convincing results. As per the project requirement this optical character recognition section was an important part to find text summarization. Now, providing result of both Tesseract and Tesseract with OpenCV techniques after applying on the original and noisy images to get similar text as present in images.



Original Image

A man asked why an old lady was all alone on a cruise ship. He noticed that although she was alone, all the staff, ships officers, Waiters and busboys were very familiar with her. So, he asked the waiter who the lady was, expecting to be told...she owned the line, but all the Waiter knew was that she was on the last four cruises, back to back.

As he left the dining room one evening he caught her eye and stopped to say hello. They chatted and he asked, "I understand you've been on this ship for the last four cruises."

She replied, "Yes, that's true."

He said, "I don't understand" and she replied, without a pause, "It's cheaper than a nursing home. So, there will be no nursing home in my future. I'm staying on this cruise; the average cost for a nursing home is $200 per day. Here, I can get a long term discount and senior discount price of $135 per day. That leaves $65 a day for tips and extra cash. I have it made here.

I can have as many as 10 meals a day of fantastic food or room service which means I can have breakfast in bed every day of the week. I get free pools, a workout room, free washers and dryers and shows every night. I get free toothpaste, razors, and free soap and shampoo. They will even treat you like a customer, not a patient. An extra $5 worth of tips will have the entire staff scrambling to help you.

I get to meet new people every 7 or 14 days! TV broken? Light bulb need changing? Need to have the matters replaced? No problem! They will fix everything and apologize for your inconvenience. Clean sheets and towels every day and you don't even have to ask for them. There is always a doctor on board. This is how I want to live the end of my life. "

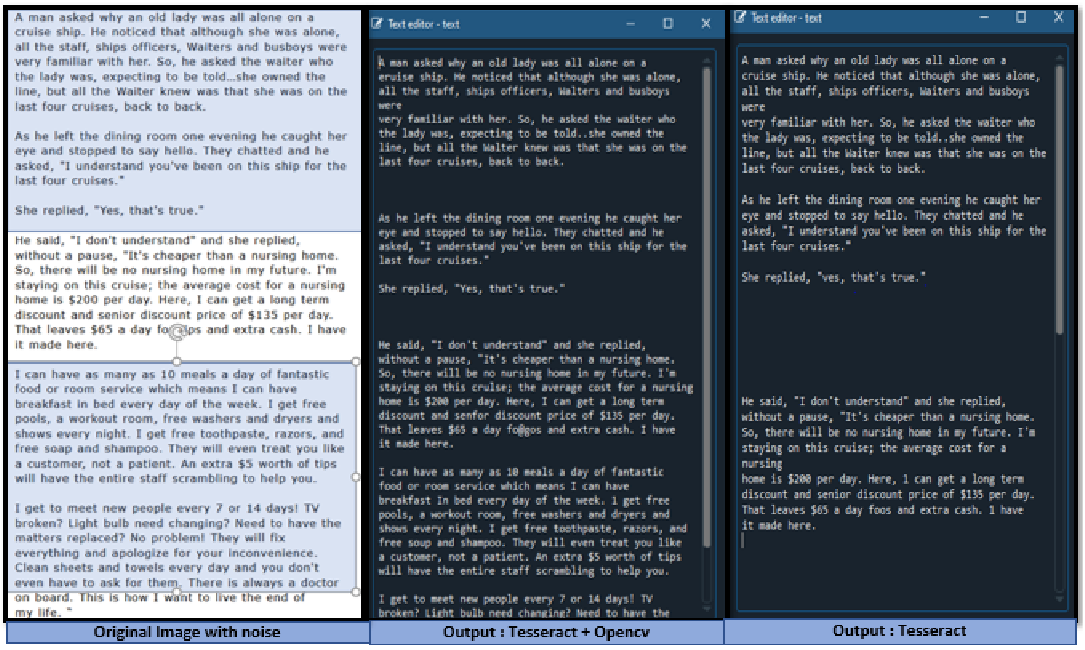
Output : Tesseract

```
A man asked why an old lady was all alone on a
cruise ship. He noticed that although she was alone,
all the staff, ships officers, Waiters and busboys were
very familiar with her. $0, he asked the waiter who
the lady was, expecting to be told..she owned the
line, but all the Waiter knew was that she was on the
last four cruises, back to back.

As he left the dining room one evening he caught her
eye and stopped to say hello. They chatted and he
asked, "I understand you've been on this ship for the
last four cruises."

She replied, "Yes, that's true."

He said, "I don't understand" and she replied,
without a pause, "It's cheaper than a nursing home.
So, there will be no nursing home in my future. I'm
staying on this cruise; the average cost for a nursing
home is $200 per day. Here, I can get a long term
discount and senior discount price of $135 per day.
That leaves $65 a day for tips and extra cash. I have
it made here.

I can have as many as 10 meals a day of fantastic
food or room service which means I can have
breakfast in bed every day of the week. I get free
pools, a workout room, free washers and dryers and
shows every night. I get free toothpaste, razors, and
free soap and shampoo. They will even treat you like
a customer, not a patient. An extra $5 worth of tips
will have the entire staff scrambling to help you.

I get to meet new people every 7 or 14 days! TV
broken? Light bulb need changing? Need to have the
matters replaced? No problem! They will fix
everything and apologize for your inconvenience.
Clean sheets and towels every day and you don't
even have to ask for them. There is always a doctor
on board. This is how I want to live the end of

my life. "
```

Output : Tesseract + Opencv

```
A man asked why an old lady was all alone on a
cruise ship. He noticed that although she was alone,
all the staff, ships officers, Waiters and busboys were
very familiar with her. So, he asked the waiter who
the lady was, expecting to be told..she owned the
line, but all the Waiter knew was that she was on the
last four cruises, back to back.

As he left the dining room one evening he caught her
eye and stopped to say hello. They chatted and he
asked, "I understand you've been on this ship for the
last four cruises."

She replied, "Yes, that's true."

He said, "I don't understand" and she replied,
without a pause, "It's cheaper than a nursing home.
So, there will be no nursing home in my future. I'm
staying on this cruise; the average cost for a nursing
home is $200 per day. Here, I can get a long term
discount and senior discount price of $135 per day.
That leaves $65 a day for tips and extra cash. I have
it made here.

I can have as many as 10 meals a day of fantastic
food or room service which means I can have
breakfast in bed every day of the week. I get free
pools, a workout room, free washers and dryers and
shows every night. I get free toothpaste, razors, and
free soap and shampoo. They will even treat you like
a customer, not a patient. An extra $5 worth of tips
will have the entire staff scrambling to help you.

I get to meet new people every 7 or 14 days! TV
broken? Light bulb need changing? Need to have the
matters replaced? No problem! They will fix
everything and apologize for your inconvenience.
Clean sheets and towels every day and you don't
even have to ask for them. There is always a doctor
on board. This is how I want to live the end of

my life. "
```

Figure 9: Applied Tesseract and Tesseract with OpenCV on a clear image



A man asked why an old lady was all alone on a cruise ship. He noticed that although she was alone, all the staff, ships officers, Waiters and busboys were very familiar with her. So, he asked the waiter who the lady was, expecting to be told...she owned the line, but all the Waiter knew was that she was on the last four cruises, back to back.

As he left the dining room one evening he caught her eye and stopped to say hello. They chatted and he asked, "I understand you've been on this ship for the last four cruises."

She replied, "Yes, that's true."

He said, "I don't understand" and she replied, without a pause, "It's cheaper than a nursing home. So, there will be no nursing home in my future. I'm staying on this cruise; the average cost for a nursing home is $200 per day. Here, I can get a long term discount and senior discount price of $135 per day. That leaves $65 a day fo ps and extra cash. I have it made here.

I can have as many as 10 meals a day of fantastic food or room service which means I can have breakfast in bed every day of the week. I get free pools, a workout room, free washers and dryers and shows every night. I get free toothpaste, razors, and free soap and shampoo. They will even treat you like a customer, not a patient. An extra $5 worth of tips will have the entire staff scrambling to help you.

I get to meet new people every 7 or 14 days! TV broken? Light bulb need changing? Need to have the matters replaced? No problem! They will fix everything and apologize for your inconvenience. Clean sheets and towels every day and you don't even have to ask for them. There is always a doctor on board. This is how I want to live the end of my life. "

Original Image with noise

```
A man asked why an old lady was all alone on a
eruise ship. He noticed that although she was alone,
all the staff, ships officers, Walters and busboys
were
very familiar with her. So, he asked the walter who
the lady was, expecting to be told..she owned the
line, but all the Walter knew was that she was on the
last four cruises, back to back.



As he left the dining room one evening he caught her
eye and stopped to say hello. They chatted and he
asked, "I understand you've been on this ship for the
last four cruises."

She replied, "Yes, that's true."



He said, "I don't understand" and she replied,
without a pause, "It's cheaper than a nursing home.
So, there will be no nursing home in my future. I'm
staying on this cruise; the average cost for a nursing
home is $200 per day. Here, I can get a long term
discount and senfor discount price of $135 per day.
That leaves $65 a day fo gos and extra cash. I have
it made here.

I can have as many as 10 meals a day of fantastic
food or room service which means I can have
breakfast In bed every day of the week. 1 get free
pools, a workout room, free washers and dryers and
shows every night. I get free toothpaste, razors, and
free soup and shampoo. They will even treat you like
a customer, not a patient. An extra $5 worth of tips
will have the entire staff scrambling to help you.

I get to meet new people every 7 or 14 days! TV
broken? Light bulb need changing? Need to have the
```

Output : Tesseract + Opencv

```
A man asked why an old lady was all alone on a
cruise ship. He noticed that although she was alone,
all the staff, ships officers, Waiters and busboys
were
very familiar with her. So, he asked the waiter who
the lady was, expecting to be told..she owned the
line, but all the Waiter knew was that she was on the
last four cruises, back to back.

As he left the dining room one evening he caught her
eye and stopped to say hello. They chatted and he
asked, "I understand you've been on this ship for the
last four cruises."

She replied, "ves, that's true."






He said, "I don't understand" and she replied,
without a pause, "It's cheaper than a nursing home.
So, there will be no nursing home in my future. I'm
staying on this cruise; the average cost for a
nursing
home is $200 per day. Here, 1 can get a long term
discount and senior discount price of $135 per day.
That leaves $65 a day foos and extra cash. 1 have
it made here.
```

Output : Tesseract

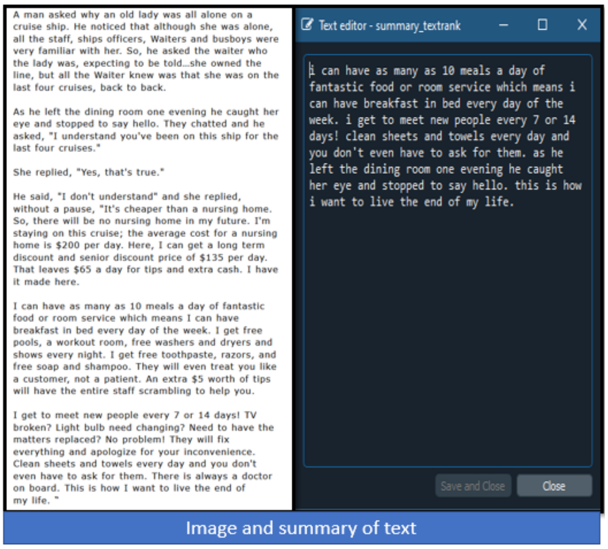
Figure 10: Applied Tesseract and Tesseract with OpenCV on a noisy image

6.2 Applied text summarization algorithms

After text extraction and text pre-processing the final part of project was implementation of text summarization algorithms which were textrank algorithm and TF-IDF algorithms. These two algorithm is unsupervised and extractive text summarization algorithm. This section is being provided evaluation and results of the applied algorithms to generate a summary of the given image.

6.2.1 Textrank algorithm

After applying textrank algorithm on extracted text to generate summary. This is an unsupervised text summarization algorithm which produced result based on the word score and sentence score. This is an extractive text summarization algorithm which can produce a very relevant extractive summary. Here, research is showed the summary of an image by using textrank algorithm, here in result of the summary is based on top five sentences of the total text which scored by sentence vector and similarity matrix. Result is providing multiple summary of multiple sentences using textrank algorithm.



A man asked why an old lady was all alone on a cruise ship. He noticed that although she was alone, all the staff, ships officers, Waiters and busboys were very familiar with her. So, he asked the waiter who the lady was, expecting to be told...she owned the line, but all the Waiter knew was that she was on the last four cruises, back to back.

As he left the dining room one evening he caught her eye and stopped to say hello. They chatted and he asked, "I understand you've been on this ship for the last four cruises."

She replied, "Yes, that's true."

He said, "I don't understand" and she replied, without a pause, "It's cheaper than a nursing home. So, there will be no nursing home in my future. I'm staying on this cruise; the average cost for a nursing home is $200 per day. Here, I can get a long term discount and senior discount price of $135 per day. That leaves $65 a day for tips and extra cash. I have it made here.

I can have as many as 10 meals a day of fantastic food or room service which means I can have breakfast in bed every day of the week. I get free pools, a workout room, free washers and dryers and shows every night. I get free toothpaste, razors, and free soap and shampoo. They will even treat you like a customer, not a patient. An extra $5 worth of tips will have the entire staff scrambling to help you.

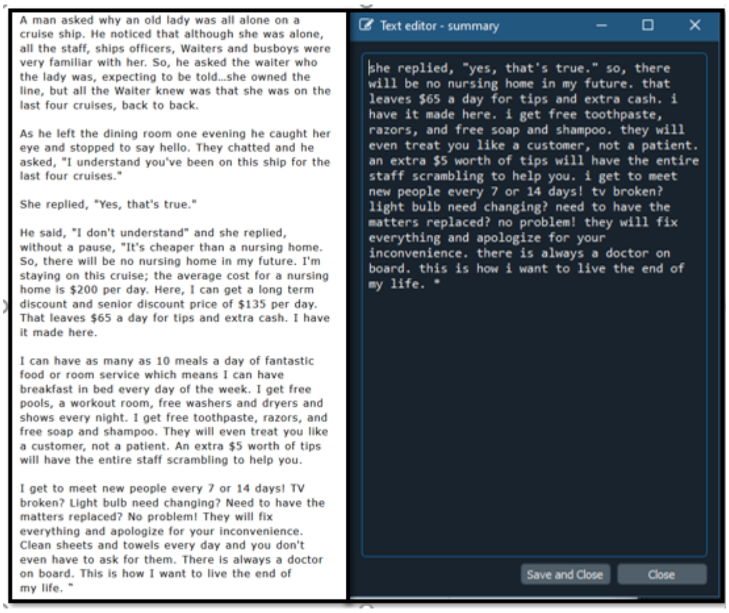
I get to meet new people every 7 or 14 days! TV broken? Light bulb need changing? Need to have the matters replaced? No problem! They will fix everything and apologize for your inconvenience. Clean sheets and towels every day and you don't even have to ask for them. There is always a doctor on board. This is how I want to live the end of my life. "

i can have as many as 10 meals a day of fantastic food or room service which means i can have breakfast in bed every day of the week. i get to meet new people every 7 or 14 days! clean sheets and towels every day and you don't even have to ask for them. as he left the dining room one evening he caught her eye and stopped to say hello. this is how i want to live the end of my life.

Figure 11: Summary using textrank algorithm

6.2.2 TF-IDF algorithm

This extractive summarization is based on TF-IDF score and along with threshold value of the text. Providing result of applied TF-IDF text summarization model which refer figure 12 and providing another output using different image with applying two extractive summarization algorithms. Refer figure 13



A man asked why an old lady was all alone on a cruise ship. He noticed that although she was alone, all the staff, ships officers, Waiters and busboys were very familiar with her. So, he asked the waiter who the lady was, expecting to be told...she owned the line, but all the Waiter knew was that she was on the last four cruises, back to back.

As he left the dining room one evening he caught her eye and stopped to say hello. They chatted and he asked, "I understand you've been on this ship for the last four cruises."

She replied, "Yes, that's true."

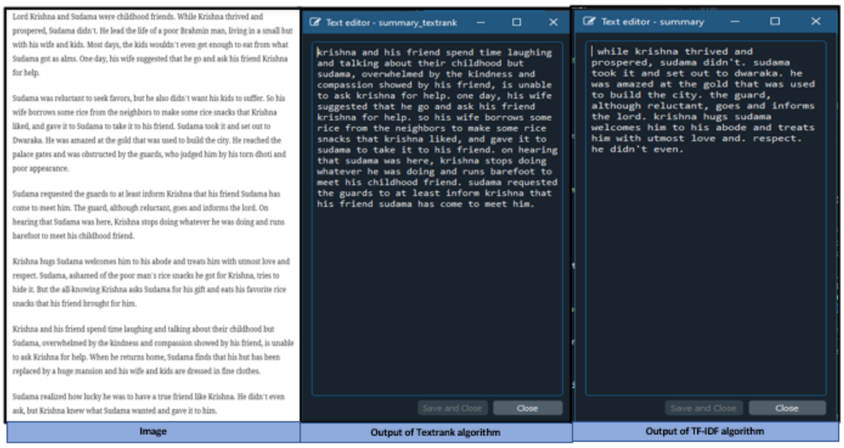
He said, "I don't understand" and she replied, without a pause, "It's cheaper than a nursing home. So, there will be no nursing home in my future. I'm staying on this cruise; the average cost for a nursing home is $200 per day. Here, I can get a long term discount and senior discount price of $135 per day. That leaves $65 a day for tips and extra cash. I have it made here.

I can have as many as 10 meals a day of fantastic food or room service which means I can have breakfast in bed every day of the week. I get free pools, a workout room, free washers and dryers and shows every night. I get free toothpaste, razors, and free soap and shampoo. They will even treat you like a customer, not a patient. An extra $5 worth of tips will have the entire staff scrambling to help you.

I get to meet new people every 7 or 14 days! TV broken? Light bulb need changing? Need to have the matters replaced? No problem! They will fix everything and apologize for your inconvenience. Clean sheets and towels every day and you don't even have to ask for them. There is always a doctor on board. This is how I want to live the end of my life. "

she replied, "yes, that's true." so, there will be no nursing home in my future. that leaves $65 a day for tips and extra cash. i have it made here. i get free toothpaste, razors, and free soap and shampoo. they will even treat you like a customer, not a patient. an extra $5 worth of tips will have the entire staff scrambling to help you. i get to meet new people every 7 or 14 days! tv broken? light bulb need changing? need to have the matters replaced? no problem! they will fix everything and apologize for your inconvenience. there is always a doctor on board. this is how i want to live the end of my life. "

Figure 12: Summary using TF-IDF algorithm with 0.5 threshold


Text editor - summary_textrank
Save and Close
Close
Text editor - summary
Save and Close
Close
Image
Output of Textrank algorithm
Output of TF-IDF algorithm

Lord Krishna and Sudama were childhood friends. While Krishna thrived and prospered, Sudama didn't. He lead the life of a poor Brahmin man, living in a small hut with his wife and kids. Most days, the kids wouldn't even get enough to eat from what Sudama got as alms. One day, his wife suggested that he go and ask his friend Krishna for help.

Sudama was reluctant to seek favors, but he also didn't want his kids to suffer. So his wife borrows some rice from the neighbors to make some rice snacks that Krishna liked, and gave it to Sudama to take it to his friend. Sudama took it and set out to Dwaraka. He was amazed at the gold that was used to build the city. He reached the palace gates and was obstructed by the guards, who judged him by his torn dhoti and poor appearance.

Sudama requested the guards to at least inform Krishna that his friend Sudama has come to meet him. The guard, although reluctant, goes and informs the lord. On hearing that Sudama was here, Krishna stops doing whatever he was doing and runs barefoot to meet his childhood friend.

Krishna hugs Sudama welcomes him to his abode and treats him with utmost love and respect. Sudama, ashamed of the poor man's rice snacks he got for Krishna, tries to hide it. But the all-knowing Krishna asks Sudama for his gift and eats his favorite rice snacks that his friend brought for him.

Krishna and his friend spend time laughing and talking about their childhood but Sudama, overwhelmed by the kindness and compassion showed by his friend, is unable to ask Krishna for help. When he returns home, Sudama finds that his hut has been replaced by a huge mansion and his wife and kids are dressed in fine clothes.

Sudama realized how lucky he was to have a true friend like Krishna. He didn't even ask, but Krishna knew what Sudama wanted and gave it to him.

```
krishna and his friend spend time laughing
and talking about their childhood but
sudama, overwhelmed by the kindness and
compassion showed by his friend, is unable
to ask krishna for help. one day, his wife
suggested that he go and ask his friend
krishna for help. so his wife borrows some
rice from the neighbors to make some rice
snacks that krishna liked, and gave it to
sudama to take it to his friend. on hearing
that sudama was here, krishna stops doing
whatever he was doing and runs barefoot to
meet his childhood friend. sudama requested
the guards to at least inform krishna that
his friend sudama has come to meet him.
```

```
 while krishna thrived and
prospered, sudama didn't. sudama
took it and set out to dwaraka. he
was amazed at the gold that was used
to build the city. the guard,
although reluctant, goes and informs
the lord. krishna hugs sudama
welcomes him to his abode and treats
him with utmost love and. respect.
he didn't even.
```

Figure 13: Output : Summary using Textrank and TF-IDF algorithm (0.85 threshold value) together

6.3 Discussion and Comparison

The project of extractive text summarization from images is divided into three parts. This project applied two techniques for OCR which are Tesseract with OpenCV and only Tesseract, then compared result of both. Text pre-processing and cleaning has been done by regular expression and natural language toolkit. Finally, applied two unsupervised extractive text summarization algorithm which are textrank algorithm and TF-IDF algorithm and discussed about the results. The result of this project shows that Tesseract and Tesseract with OpenCV perfectly produced text from image but on the noisy image Tesseract with OpenCV act better than only Tesseract. Then this project went forward with the result of Tesseract with OpenCV result. In the middle of this project text pre-processing and cleaning successfully applied with regular expression and NLTK libraries. Textrank and TF-IDF algorithms applied on the cleaned text and produced extractive summary, here textrank algorithm is a graph based algorithm and use word embedding to generate score of the text and select only the top scored sentences to made summary, on the other side TF-IDF algorithm generate score of the sentences depends on term frequency and inverse term frequency which generate an extractive summary. This pipeline of this project successfully produce a good result with two extractive summarization algorithm. As mentioned previous research was based on PyTesseract and summarization algorithm but this project successfully implemented the PyTesseract with OpenCV which produce very effective result and applied two unsupervised models to find text summary. But, the accuracy measurement of generated summary is very challenging which need to overcome in future.

7 Conclusion and Future Work

Objective of the project is to find an extractive summary from given image using unsupervised extractive summarization algorithms and the research is successfully produce extractive summary of image using two different algorithm Textrank and TF-IDF. Equally, Tesseract with OpenCV perform very well to extract text from images. This research made a pipeline in between two different programs and produce an effective summary but the accuracy identification of produced result is an challenging task which need to be done in future. Also, total application can apply on the image without diagram to get better result which is a limitation of this research. Future work of the research is to implement this program as an application of text summarization which can be a time saver also, need to work on the measurement of the accuracy of output summary.

References

Babar, S., Tech-Cse, M. and Rit (2013). Text summarization:an overview.

Bienia, C. and Li, K. (2010). Characteristics of workloads using the pipeline programming model, *International Symposium on Computer Architecture*, Springer, pp. 161–171.

Chawla, M., Jain, R. and Nagrath, P. (2020). Implementation of tesseract algorithm to extract text from different images, *Available at SSRN 3589972* .

Christian, H., Agus, M. and Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (tf-idf), *ComTech: Computer, Mathematics and Engineering Applications* **7**: 285.

El-Refaiy, A., Abas, A. and Elhenawy, I. (2018). Review of recent techniques for extractive text summarization, *Journal of Theoretical and Applied Information Technology* **96**: 7739–7759.

Gupta, V. and Lehal, G. S. (2010). A survey of text summarization extractive techniques, *Journal of emerging technologies in web intelligence* **2**(3): 258–268.

Joshi, N. (2019). Text image extraction and summarization, *Asian Journal For Convergence In Technology (AJCT)* .

Karlsson, S. (2017). Using semantic folding with textrank for automatic summarization.

Mallick, C., Das, A. K., Dutta, M., Das, A. K. and Sarkar, A. (2019). Graph-based text summarization using modified textrank, *Soft Computing in Data Analytics*, Springer, pp. 137–146.

Palekar, R. R., Parab, S. U., Parikh, D. P. and Kamble, V. N. (2017). Real time license plate detection using opencv and tesseract, *2017 International Conference on Communication and Signal Processing (ICCSP)*, pp. 2111–2115.

Pawar, N., Shaikh, Z., Shinde, P. and Warke, Y. (2019). Image to text conversion using tesseract, *Image* **6**(02).

Petersen, K., Wohlin, C. and Baca, D. (2009). The waterfall model in large-scale development.

Ramamoorthy, C. V. and Li, H. F. (1977). Pipeline architecture, *ACM Comput. Surv.* **9**: 61–102.

ROBINSON, J. and SARAVANAN, V. (2019). An extractive based multi-document summarization using weighted tf-idf and centroid based k-means clustering (tf-idf: Cbc) for large text data, *Journal of Critical Reviews* **7**(1): 2020.

Smith, R. (2007). An overview of the tesseract ocr engine, *Ninth international conference on document analysis and recognition (ICDAR 2007)*, Vol. 2, IEEE, pp. 629–633.