

Text Summarization

On

Technical Documentation

Project Report Submitted in Partial Fulfilment of the
Requirements for the degree of
Master of Computer Application
Of
Jadavpur University
May, 2019

By
Puspendu Sarkar
Master of Computer Application – III
Examination Roll Number: MCA196003
Registration Number: 137311 of 2016 – 2017
Roll no. 001610503003

Under the guidance of

Dr. CHITRITA CHAUDHURI
Associate Professor

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University
Kolkata – 700032, India
May, 2019

**COMPUTER SCIENCE AND ENGINEERING
DEPARTMENT
FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY**

TO WHOM IT MAY CONCERN

I hereby forward the project report entitled “*Text Summarization on Technical Documentation*” prepared by **Puspendu Sarkar** under my supervision to be accepted in partial fulfilment for the degree of **Master of Computer Application** in the Faculty and Technology of Jadavpur University, Kolkata.

(Dr. Chitrita Chaudhuri)

Associate Professor

Project Supervisor

Dept. of Computer Science and Engineering

Jadavpur University

Kolkata – 700032

Countersigned:

Prof Mahantapas Kundu

Head, Dept. of Computer Science and Engineering

Jadavpur University

Kolkata – 700032

Prof. Chiranjib Bhattacharjee

Dean, Faculty of Engineering and Technology

Jadavpur University

Kolkata – 70032

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF APPROVAL *

The foregoing project report is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to the degree for which it has been submitted. It is understood that, by this approval, the undersigned do not necessary endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the project report only for the purpose for which it has been submitted.

Final Examination for
evaluation of the project

(Signatures of Examiners)

* Only in case the project report is approved

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this project report contains literature survey and original research work by undersigned candidate, as part of my Master of Computer Application studies.

All information in this document had been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

NAME : **Puspendu Sarkar**

Examination Roll Number : **MCA196003**

Registration Number : **137311 of 2016 - 2017**

Project Title : **Text Summarization On Technical
Documentation**

Signature with Date :

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of this task would be incomplete without the mention of the people who made it possible. Their constant guidance and encouragement crowned my effort with success.

It is a great pleasure to express my sincerest thanks to my project supervisor Dr. Chitrita Chaudhuri, Associate Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, for her encouragement, valuable suggestion, and constant support during the course of this project.

I would like to thank all the professors of the Department of Computer Science and Engineering, Jadavpur University, Kolkata for the guidance they provided me throughout the duration of the Master of Computer Application course.

A special note of thanks goes to Prof Mahantapas Kundu, Head, Department of Computer Science and Engineering, Jadavpur University.

I am also thankful to Prof. Chiranjib Bhattacharjee, Dean, Faculty of Engineering and Technology, for providing an excellent environment for completion of this project.

I am also indebted to my co-researchers Mr. Anupam Baidya, Mr. Rishi Dey for their seamless co-operation and help in completion of this project. I am thankful to my fellow classmates and my family for constant help and support.

Date: _____

Puspendu Sarkar
Master of Computer Application – III
Examination Roll No. – MCA196003
Registration No: 13711 of 2016 – 2017

Contents

Chapter 1	Introduction	1
Chapter 2	Related Works	2
Chapter 3	Basic Concepts	
3.1	Text Summarization	3
3.2	Automation Testing	4
3.3	Live video streaming	5
Chapter 4	Methodology	
4.1	Algorithm	6
4.2	Flowchart	7
4.3	<i>Program Code related to Algorithm</i>	8
Chapter 5	Experimental Setup	
5.1	Text Summarization Datasets	10
5.2	Machine Configuration	10
5.3	Tools	10
Chapter 6	Results	

6.1	Input dataset 1	
6.1.1	50% Summarized text of Dataset 1	12
6.1.2	30% Summarized text of Dataset 1	13
6.1.3	10% Summarized text of Dataset 1	14
6.2	Input dataset 2	
6.2.1	50% Summarized text of Dataset 2	15
6.2.2	30% Summarized text of Dataset 2	16
6.2.3	10% Summarized text of Dataset 2	16
6.3	Input dataset 3	
6.3.1	50% Summarized text of Dataset 3	17
6.3.2	30% Summarized text of Dataset 3	18
6.3.3	10% Summarized text of Dataset 3	19
Chapter 7	Conclusion and Future Scope	20
Bibliography	21
Annexure – Program Code	22

Chapter 1

Introduction

Text Summarization is one of those applications of Natural Language Processing (NLP) which is bound to have a huge impact on our lives. With growing digital media and ever growing publishing – who has the time to go through entire articles / documents / books to decide whether they are useful or not? Thankfully – this technology is already here.

Automatic Text Summarization is one of the most challenging and interesting problems in the field of NLP. It is a process of generating a concise and meaningful summary of text from several types of text resources such as books, news articles, blog posts, research papers, emails, and tweets.

The demand for automatic text summarization systems is spiking these days thanks to the availability of large amounts of textual data.

Through this work, we will explore the realms of text summarization. We will understand how the text summarization algorithm works, and will also implement our version in Python.

The researcher had the privilege to undergo a technical training session with Rebeca Technologies Pvt. Ltd as an intern during the 6th semester of the course. The topics covered included Automation Testing and Live video streaming, to mention a few. The highlights of the work are included in the report in a summarized form in relevant chapters.

The next chapter 2 also contains a survey on works related to the domain of text summarization, Automation Testing and Live video streaming. The following chapter 3 discusses the basic concepts associated with the topics. Chapter 4 next describes the methodology adopted for the summarization techniques used. Chapter 5 contains the details of data and tools used, as well as screenshots of actual summarized text obtained as output. The last chapter 6 concludes with promises of some related future research directions. The report ends with a bibliography section describing all cited major works.

Chapter 2

RELATED WORKS

On text summarization, Ani Nenkova et al [1] in their work have discussed several approaches for finding important portions of a text document to be preserved within the summarized version. Their first objective have been to derive an intermediate representation of the topic covered in the input. The sentences in the input are next scored on the basis of importance. They also discussed an alternative indicator representation approach. The sentences are selected for the summary using either a greedy approach, or an optimizing technique which chooses the best set.

On automation testing, Karuturi Sneha et al [2] discussed the processes of software testing through automation tools which decreases output errors as well as overheads such as manpower and time. They compare manual testing with automation testing applying different types of techniques.

On live video streaming in [3] Baochun Li et al describe how video streaming mechanism has changed from the design of transport protocols for streaming video, to the peer-to-peer paradigm at the application layer. Recent researchs using Dynamic Adaptive Streaming over HTTP is also probed here for building more practical and scalable systems. The work mainly focus on peer-to-peer streaming protocols, cloud computing and social media.

Chapter 3

Basic Concepts

3.1 Text summarization:

Text summarization can broadly be divided into two categories..

A. Extractive Summarization and **B. Abstractive Summarization**.

A. Extractive Summarization:

These methods rely on extracting several parts, such as phrases and sentences, from a piece of text and stack them together to create a summary. Therefore, identifying the right sentences for summarization is of utmost importance in an extractive method.

B. Abstractive Summarization:

These methods use advanced NLP techniques to generate an entirely new summary. Some parts of this summary may not even appear in the original text.

For all types of summarization the following concepts are important :-

Topic Representation:

Topic representation approaches vary tremendously in sophistication and encompass a family of methods for summarization. Here we present one of the most widely applied topic representation approaches, that have been gaining popularity because of it's recent success.

Term Frequency:

The term frequency gives us the frequency of the word in document - that is the number of times the word appears in a document.

This frequency will help to find the relevance of a word to the document. The more a word occurs in a document, the more the importance of that word increases in that document. In the present work, we are attempting to device an extractive summarization technique by scoring relevant sentences appropriately with the help of frequent terms.

3.2 Automation Testing

Manual Testing of software, API, SQL queries and data sets are all performed by humans. Usually these require immense amount of time, patience and skill. Trend is to replace manual testing by automation testing. Automation Testing means using an automation tool to execute a test case suite. [4]

The automation software is a complete tool in this respect. Not only can it enter test data into the System Under Test, but it can also perform comparisons between expected and actual results, producing comprehensive and complete test reports.

Although primarily Test Automation involves monetary investments, the consequent benefits far outweighs the initial investments in resource. For repeated execution of same test suite in successive development cycles, a test automation tool can re-play it as required. No further human intervention is required. This improves ROI feature.

However, one word of caution, Automation does not eliminate Manual Testing altogether. At best it helps reduce the number of manual test cases to be run.

Importance of Automated software testing is due to the following reasons:

- In many scenarios, Manual Testing is by itself time and money consuming.
- Multilingual sites may cause difficulty for a person.
- Automation Test can be run unattended, even overnight.
- Test execution can be speeded up by Automation.
- Manual limitations of Test Coverage are overcome through automation.
- Errors caused by human fatigue and faults are absent.

Following are benefits of automated testing:

- 1.** 70% faster than the manual testing
- 2.** Wider test coverage of application features
- 3.** Reliable in results and Ensure Consistency
- 4.** Improves accuracy, while saving Time and Cost
- 5.** Human Intervention is not required while execution
- 6.** Increases Efficiency
- 7.** Better speed in executing tests
- 8.** Re-usable test scripts and test more frequently and thoroughly
- 9.** More cycle of execution can be achieved through automation
- 10.** Early time to market

However, one word of caution, Automation does not eliminate Manual Testing altogether. At best it helps reduce the number of manual test cases to be run.

3.3 Live video streaming

Video streaming is a type of media streaming in which the data from a video file is continuously delivered via the Internet to a remote user.

Live-streaming refers to online streaming media simultaneously recorded and broadcast in real time.

Now a days, Adobe's Flash video technology has become a popular method of video streaming. But recently Adobe's Flash protocol has been replaced increasingly by video delivered using protocols like HTTP Live Streaming(HLS) played in HTML5 video players. Another advantage of these are that these can be accessed free of cost, and they are safer, more reliable, and faster than earlier technologies [5].

In content production too these new technologies are mostly advantageous. But the disadvantage lies in the work involved in replacing legacy systems and technologies with new standards that may not work with same aplomb across all platforms.

Videos are bound to occupy the majority Internet traffic in the near future [6] – and most of it will necessarily be live. From business professions to college and school level education – the demand market is high, not to mention the pioneering domains of scientific experiments, remote controlled medical attendance and video conferencing. Tech savvy operators will also be required and prior trainings in the field would be a time-saver. [7]

The actual beneficial effects of live video transmissions are manifold. One may consider the stronger emotional bonding effects experienced by the participating viewer in contrast to the passive onlooker. In some other domains, such as operating hazardous technical instruments in industry, or collecting live stream data from remote geographical locations, the benefits far exceeds the cost involved, in terms of money, safety and execution standards.

Chapter 4

Methodology

In order to better understand the operation of summarization systems and to emphasize the design choices system developers need to make, many researcher distinguish the following three relatively independent tasks performed by virtually all summarizers: creating an intermediate representation of the input which captures only the key aspects of the text, scoring sentences based on that representation, and selecting a summary consisting of several sentences.

4.1 **Algorithm:** Text Summarization

Input:

1. Text document
2. Percentage of summarization

Output:

1. Summarized text
2. Number of lines in summary

Method:

Step1. *Split the input text by the full-stop sign to get each sentence from the input text.*

Step2. *Remove stop words from the collection of the sentences.*

Step3. *Tokenize each word from the collection.*

Step4. *Create a dictionary using the collection of words to find the significance of each word and create a list of keywords.*

Step5. *Take each sentence and score them using the keywords.*

Step6. *Sort the score for each sentence and find a threshold value using the user input related to the percentage size of output text.*

Step7. *Collect all the sentences from the text which have score higher or equal to the threshold value.*

4.2 Flowchart

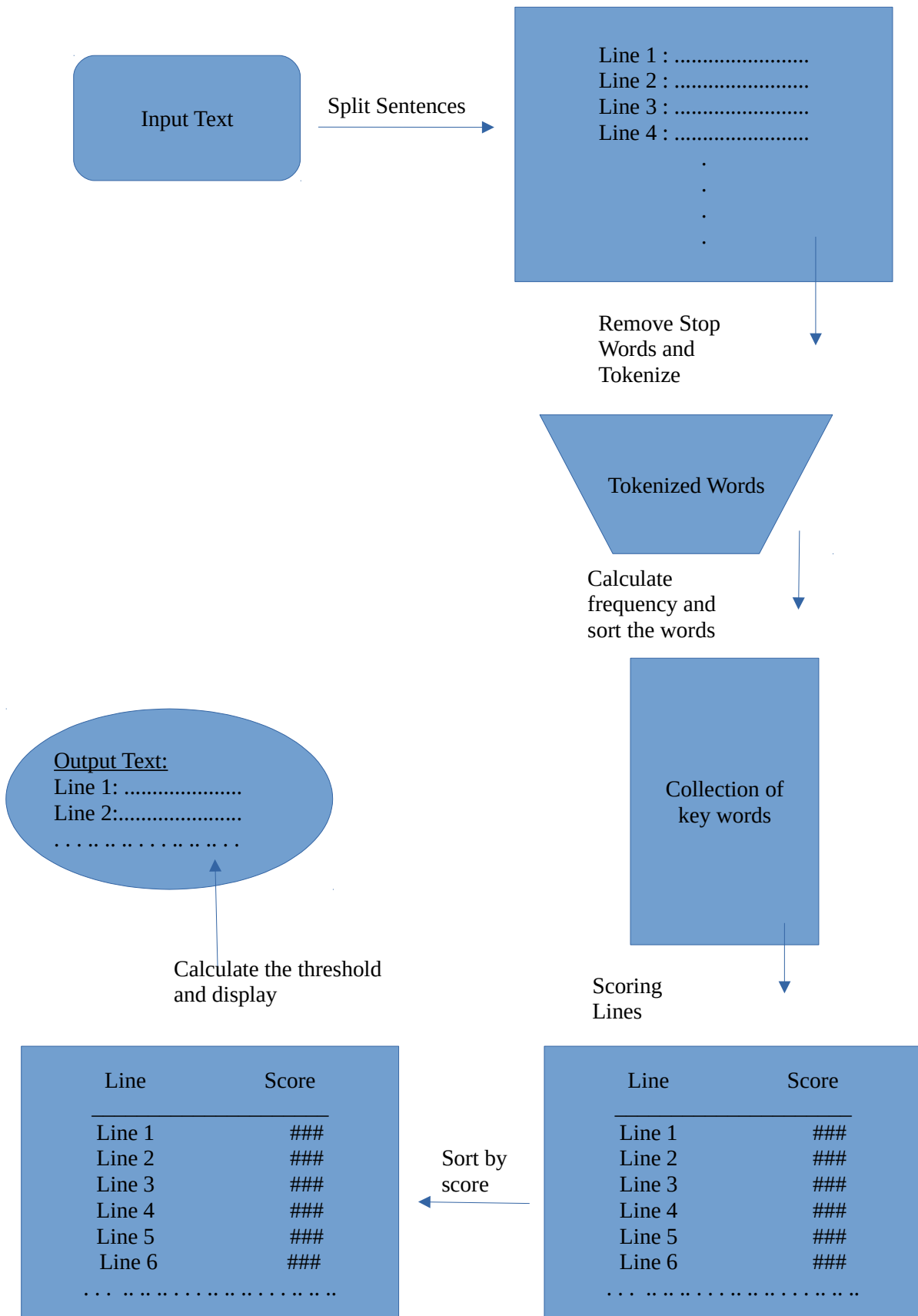


Figure 1: Flowchart for the summarization operation

4.3 Program Code related to Algorithm

Step1. `splitData = data.split('.')`

Step2. `st=set(stopwords.words('english'))`
`stopwords = nltk.corpus.stopwords.words('english')`
`newStopWords = ['n\t','\s','â€','“', '*'] # additional stop words`
`stopwords.extend(newStopWords)`

Step3. `s=sent_tokenize(data)`

Step4. `for i in s:`
 `wd=word_tokenize(i)`
 `tagged=nltk.pos_tag(wd)`
 `for j in tagged:`
 `if (j[1][0]=='N' or j[1][0]=='R' or j[1][0]=='V' or j[1][0]=='J')`
 `and j[0] not in stopwords:`
 `if j[1][0]=='v':`
 `#ROOT word`
 `k=lem.lemmatize(j[0],pos='v')`
 `else:`
 `k=j[0]`
 `if k not in word:`
 `word[k]=0`
 `word[k]+=1`

Step5. `for line in splitData:`
 `score = 0`
 `for aElement in a:`
 `index += 1`
 `if aElement[0] in line:`
 `score += aElement[1]`
 `if index == len(a):`
 `break`
 `newData = [line, score]`
 `if maxScore < score:`
 `maxScore = score`
 `if score >= 0 and`
 `((2 * len(line)) - len(line.lstrip("")) - len(line.lstrip('\n'))) <= 10 :`
 `data.append(newData)`

```

Step6. numberOfLines = ( len(data) * size ) / 100
print("Analysing required size...")
listOfScore = [0]
for collection in data:
    listOfScore.append(collection[1])
listOfScore.sort(reverse = True)
thresholdScore = listOfScore[int(numberOfLines)]

```

```

Step7. for collection in data:
    if collection[1] > thresholdScore:
        outputSize += 1
        print("_____")
        print(collection[0])
print("_____")
print("Size of output(Line number)", outputSize)

```

[All program parts coded in Python 3]

Chapter 5

Experimental Setup

5.1 Dataset

Dataset-1: Chapter 1 of [8], Page 4 (Full Page)

Dataset-2: Chapter 1 of [9], Introduction (Page 1- 2)

Dataset-3:Chapter 1 of [1], Page 1 – 4

5.2 Machine Configuration:

System – HP [Model Number :RTL8723BE]

Processor: Intel® Core™ i5-4210U CPU @ 1.70GHz × 4
RAM: 3.80 GiB
System type: UBUNTU MATE 1.20.1
UBUNTU Release 18.04.1 LTS (Bionic Beaver) 64-bit
Kernel Linux 4.15.0-50-lowlatency x86_64

5.3 Tools

- Spyder 3 to run python programming with in-built NLTK.
- Cucumber to run automation test of another software.
- Eclipse to support Cucumber software tool.
- FFmpeg handling video, audio, and other multimedia files and streams.
- HTTP Live Streaming(HLS) for HTTP-based adaptive bitrate streaming communications

Chapter 6

Results and Performance Analysis

6.1 Input dataset-1 (Actual input text-screenshot)

Since the 1960s, database and information technology has evolved systematically from primitive file processing systems to sophisticated and powerful database systems. The research and development in database systems since the 1970s progressed from early hierarchical and network database systems to relational database systems (where data are stored in relational table structures; see Section 1.3.1), data modeling tools, and indexing and accessing methods. In addition, users gained convenient and flexible data access through query languages, user interfaces, query optimization, and transaction management. Efficient methods for online transaction processing (OLTP), where a query is viewed as a read-only transaction, contributed substantially to the evolution and wide acceptance of relational technology as a major tool for efficient storage, retrieval, and management of large amounts of data.

After the establishment of database management systems, database technology moved toward the development of *advanced database systems*, *data warehousing*, and *data mining* for advanced data analysis and *web-based databases*. Advanced database systems, for example, resulted from an upsurge of research from the mid-1980s onward. These systems incorporate new and powerful data models such as extended-relational, object-oriented, object-relational, and deductive models. Application-oriented database systems have flourished, including spatial, temporal, multimedia, active, stream and sensor, scientific and engineering databases, knowledge bases, and office information bases. Issues related to the distribution, diversification, and sharing of data have been studied extensively.

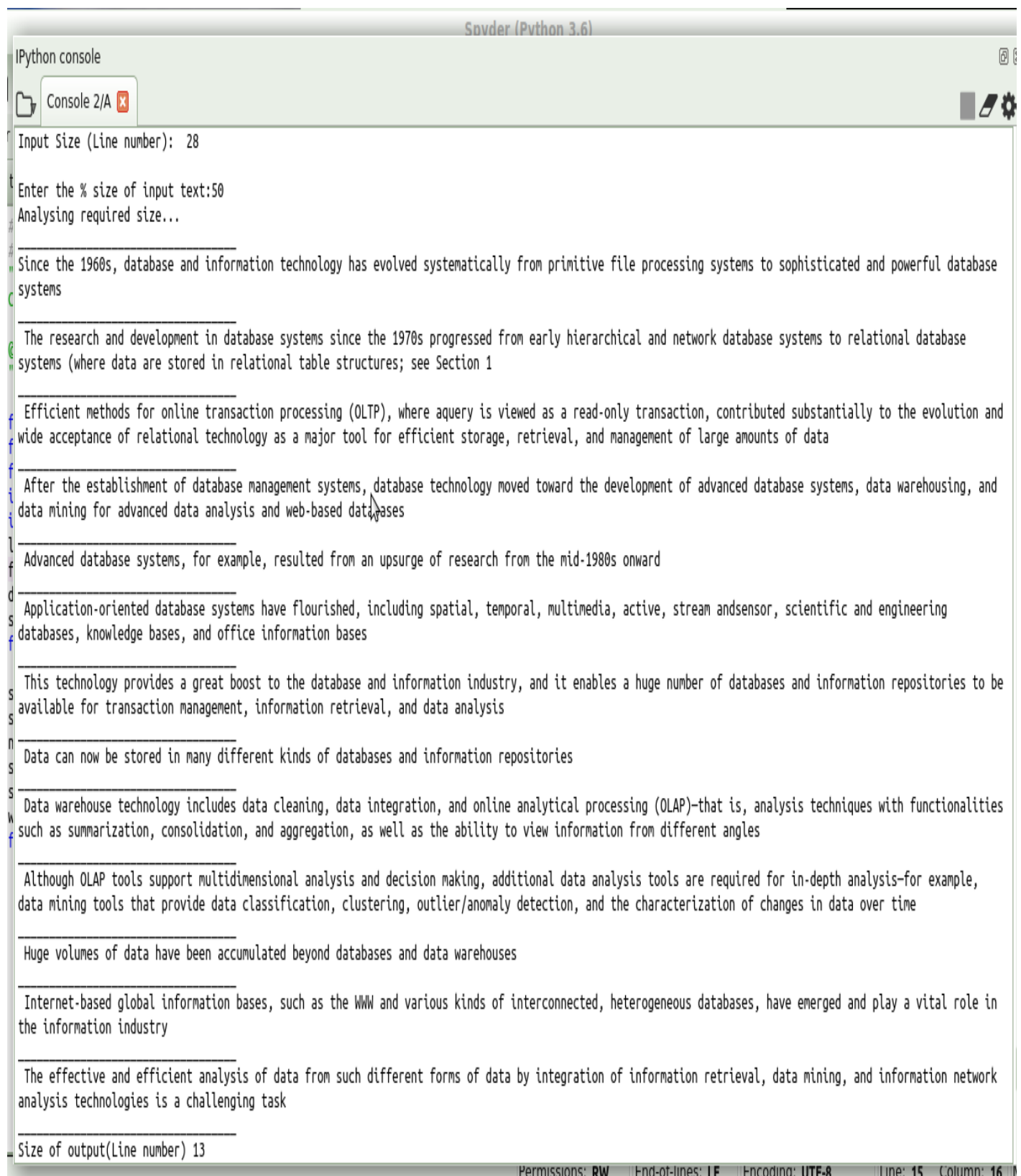
Advanced data analysis sprang up from the late 1980s onward. The steady and dazzling progress of computer hardware technology in the past three decades led to large supplies of powerful and affordable computers, data collection equipment, and storage media. This technology provides a great boost to the database and information industry, and it enables a huge number of databases and information repositories to be available for transaction management, information retrieval, and data analysis. Data can now be stored in many different kinds of databases and information repositories.

One emerging data repository architecture is the **data warehouse** (Section 1.3.2). This is a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making. Data warehouse technology includes data cleaning, data integration, and online analytical processing (OLAP)—that is, analysis techniques with functionalities such as summarization, consolidation, and aggregation, as well as the ability to view information from different angles. Although OLAP tools support multidimensional analysis and decision making, additional data analysis tools are required for in-depth analysis—for example, data mining tools that provide data classification, clustering, outlier/anomaly detection, and the characterization of changes in data over time.

Huge volumes of data have been accumulated beyond databases and data warehouses. During the 1990s, the World Wide Web and web-based databases (e.g., XML databases) began to appear. Internet-based global information bases, such as the WWW and various kinds of interconnected, heterogeneous databases, have emerged and play a vital role in the information industry. The effective and efficient analysis of data from such different forms of data by integration of information retrieval, data mining, and information network analysis technologies is a challenging task.

Figure 2: Screenshot of input dataset-1

6.1.1 50% summarized text of Dataset 1



```
Python console
Console 2/A x
Input Size (Line number): 28
Enter the % size of input text:50
Analysing required size...

Since the 1960s, database and information technology has evolved systematically from primitive file processing systems to sophisticated and powerful database systems

The research and development in database systems since the 1970s progressed from early hierarchical and network database systems to relational database systems (where data are stored in relational table structures; see Section 1

Efficient methods for online transaction processing (OLTP), where aquery is viewed as a read-only transaction, contributed substantially to the evolution and wide acceptance of relational technology as a major tool for efficient storage, retrieval, and management of large amounts of data

After the establishment of database management systems, database technology moved toward the development of advanced database systems, data warehousing, and data mining for advanced data analysis and web-based databases

Advanced database systems, for example, resulted from an upsurge of research from the mid-1980s onward

Application-oriented database systems have flourished, including spatial, temporal, multimedia, active, stream andsensor, scientific and engineering databases, knowledge bases, and office information bases

This technology provides a great boost to the database and information industry, and it enables a huge number of databases and information repositories to be available for transaction management, information retrieval, and data analysis

Data can now be stored in many different kinds of databases and information repositories

Data warehouse technology includes data cleaning, data integration, and online analytical processing (OLAP)—that is, analysis techniques with functionalities such as summarization, consolidation, and aggregation, as well as the ability to view information from different angles

Although OLAP tools support multidimensional analysis and decision making, additional data analysis tools are required for in-depth analysis—for example, data mining tools that provide data classification, clustering, outlier/anomaly detection, and the characterization of changes in data over time

Huge volumes of data have been accumulated beyond databases and data warehouses

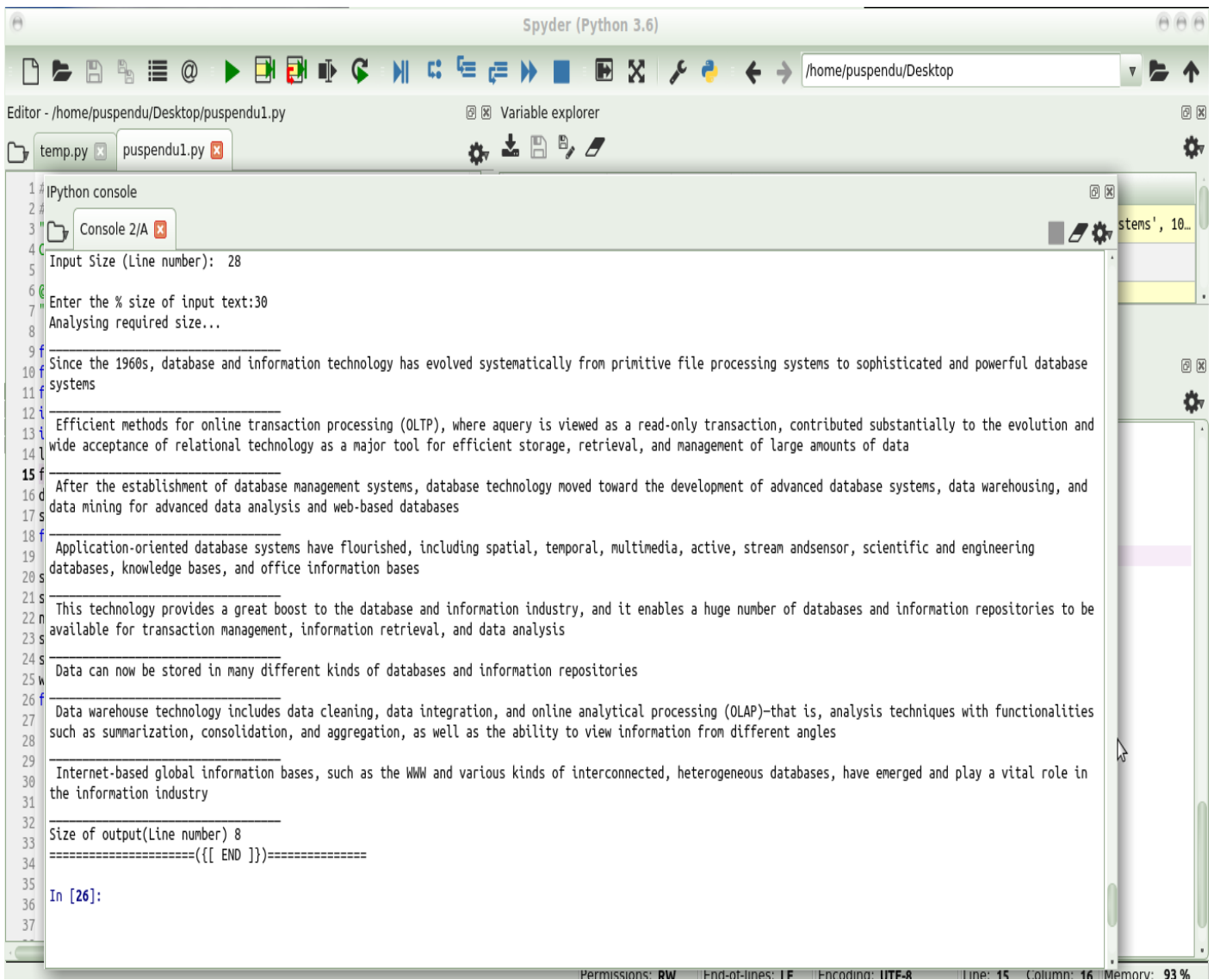
Internet-based global information bases, such as the WWW and various kinds of interconnected, heterogeneous databases, have emerged and play a vital role in the information industry

The effective and efficient analysis of data from such different forms of data by integration of information retrieval, data mining, and information network analysis technologies is a challenging task

Size of output(Line number) 13
Permissions: RW End-of-lines: LF Encoding: UTF-8 Line: 15 Column: 16
```

Figure 3: 50% summarized text of dataset-1

6.1.2 30% summarized text of Dataset 1



The screenshot shows the Spyder Python IDE interface. The main window displays a Python console with the following output:

```
1 # Python console
2 #
3 # Console 2/A
4 #
5 Input Size (Line number): 28
6 #
7 Enter the % size of input text:30
8 Analysing required size...
9 #
10 f Since the 1960s, database and information technology has evolved systematically from primitive file processing systems to sophisticated and powerful database systems
11 f
12 i Efficient methods for online transaction processing (OLTP), where aquery is viewed as a read-only transaction, contributed substantially to the evolution and wide acceptance of relational technology as a major tool for efficient storage, retrieval, and management of large amounts of data
13 i
14 f After the establishment of database management systems, database technology moved toward the development of advanced database systems, data warehousing, and data mining for advanced data analysis and web-based databases
15 f
16 d Application-oriented database systems have flourished, including spatial, temporal, multimedia, active, stream andsensor, scientific and engineering databases, knowledge bases, and office information bases
17 d
18 f This technology provides a great boost to the database and information industry, and it enables a huge number of databases and information repositories to be available for transaction management, information retrieval, and data analysis
19 f
20 s Data can now be stored in many different kinds of databases and information repositories
21 s
22 n Data warehouse technology includes data cleaning, data integration, and online analytical processing (OLAP)—that is, analysis techniques with functionalities such as summarization, consolidation, and aggregation, as well as the ability to view information from different angles
23 n
24 s Internet-based global information bases, such as the WWW and various kinds of interconnected, heterogeneous databases, have emerged and play a vital role in the information industry
25 s
26 f
27 f
28 f
29 f
30 f
31 f
32 Size of output(Line number) 8
33 =====({{ END }})=====
34 #
35 In [26]:
36 #
37 #
```

Figure 4: 30% summarized text of dataset-1

6.1.3 10% summarized text of Dataset 1

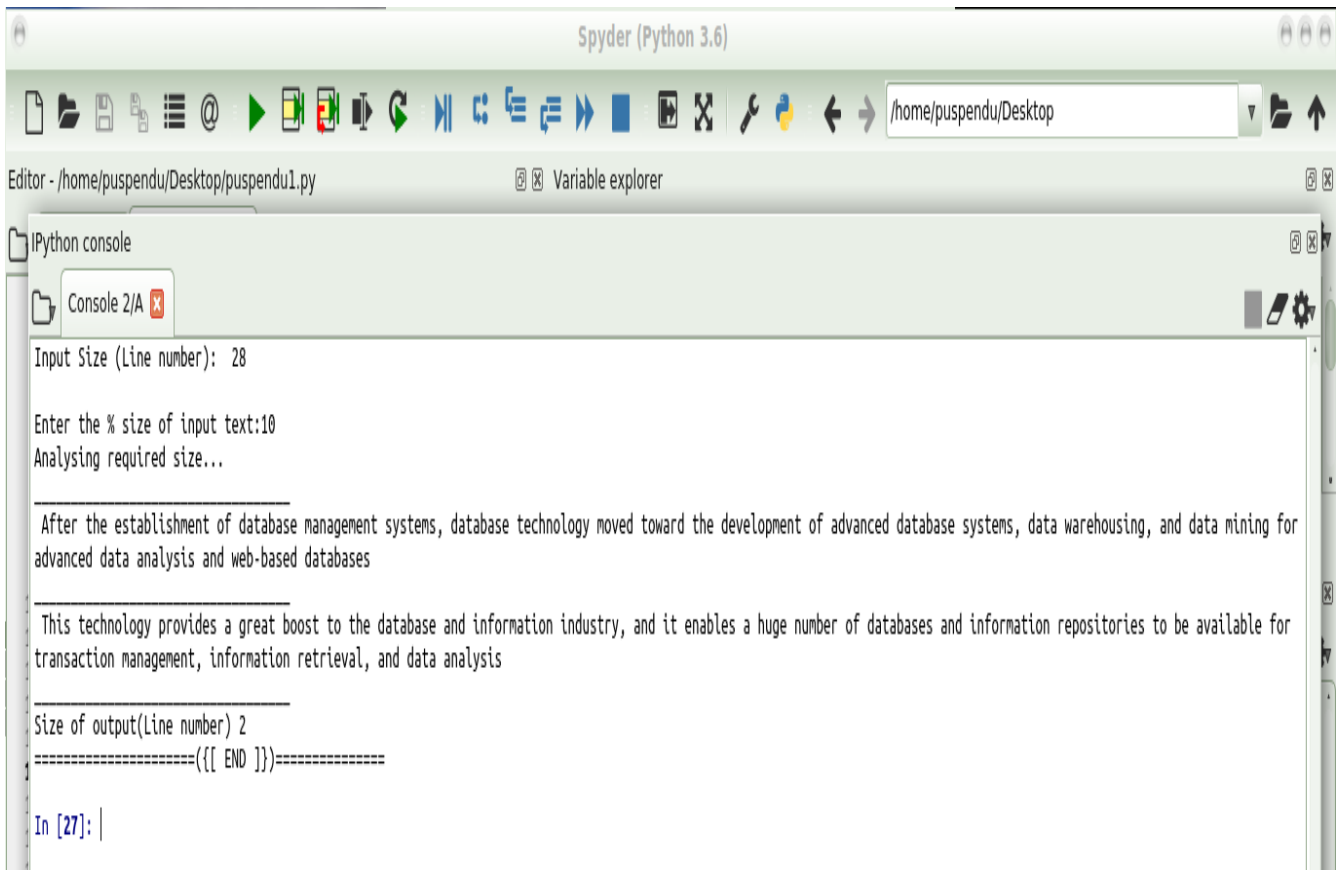


Figure 5: 10% summarized text of dataset-1

6.2 Input dataset-2 [9]

6.2.1 50% summarized text of Dataset 2

```
IPython console
Console 1/A
Input Size (Line number): 32
Enter the % size of input text:50
Analysing required size...

Imagine computers learning from medical records which treatments are most effective for new
diseases, houses learning from experience to optimize energy costs based on the
particular usage patterns of their occupants, or personal software assistants learning the evolving interests of their users in order to highlight especially relevant
stories from the online morning newspaper

A successful understanding of how to
make computers learn would open up many new uses of computers and new levels
of competence and customization

And a detailed understanding of informationprocessing algorithms for machine learning might lead to a better understanding
of human learning abilities (and disabilities) as well

However, algorithms have been invented that are effective for certain types
of learning tasks, and a theoretical understanding of learning is beginning to
emerge

Many practical computer programs have been developed to exhibit useful types of learning, and significant commercial applications have begun to appear

For problems such as speech recognition, algorithms based on machine
learning outperform all other approaches that have been attempted to date

In
the field known as data mining, machine learning algorithms are being used routinely to discover valuable knowledge from large commercial databases containing
equipment maintenance records, loan applications, financial transactions, medical
records, and the like

As our understanding of computers continues to mature, it seems inevitable that machine learning will play an increasingly central role in computer science and computer
technology

A few specific achievements provide a glimpse of the state of the art: programs have been developed that successfully learn to recognize spoken words (Waibel 1989; Lee 1989),
predict recovery rates of pneumonia patients (Cooper
et al

Theoretical results have been developed that characterize the fundamental
relationship among the number of training examples observed, the number of hypotheses under consideration, and the expected error in learned hypotheses

We
are beginning to obtain initial models of human and animal learning and to understand their relationship to learning algorithms developed for computers (e

Several recent applications of machine learning are summarized in Table 1

(1994) survey additional applications of
machine learning

This book presents the field of machine learning, describing a variety of
learning paradigms, algorithms, theoretical results, and applications

It draws on results from artificial intelligence, probability and statistics, computational complexity theory, control theory, information theory, philosophy, psychology,
neurobiology, and other
fields

2 summarizes key ideas from each of these fields that impact the
field of machine learning

Size of output(Line number) 16
```

Figure 6: 50% summarized text of dataset-2

6.2.1 30% summarized text of Dataset 2

```
IPython console
Console 2/A x
algorithms -> /
applications -> 6
Input Size (Line number): 32

Enter the % size of input text:30
Analysing required size...

Imagine computers learning from medical records which treatments are most effective for new
diseases, houses learning from experience to optimize energy costs based on the
particular usage patterns of their occupants, or personal software assistants learning the evolving interests of their users in order to highlight especially relevant
stories from the online morning newspaper

And a detailed understanding of informationprocessing algorithms for machine learning might lead to a better understanding
of human learning abilities (and disabilities) as well

However, algorithms have been invented that are effective for certain types
of learning tasks, and a theoretical understanding of learning is beginning to
emerge

Many practical computer programs have been developed to exhibit useful types of learning, and significant commercial applications have begun to appear

For problems such as speech recognition, algorithms based on machine
learning outperform all other approaches that have been attempted to date

In
the field known as data mining, machine learning algorithms are being used routinely to discover valuable knowledge from large commercial databases containing
equipment maintenance records, loan applications, financial transactions, medical
records, and the like

As our understanding of computers continues to mature, it seems inevitable that machine learning will play an increasingly central role in
computer science and computer technology

We
are beginning to obtain initial models of human and animal learning and to understand their relationship to learning algorithms developed for computers (e

This book presents the field of machine learning, describing a variety of
learning paradigms, algorithms, theoretical results, and applications

Size of output(Line number) 9
=====({[ END ]})=====

In [35]: runfile('/home/puspendu/Desktop/puspendu1.py', wdir='/home/puspendu/Desktop')
/// Ever since computers were invented, we have wondered whether they might be
made to learn ///
```

Figure 7: 30% summarized text of dataset-2

6.2.1 10% summarized text of Dataset 2

```
Input Size (Line number): 32

Enter the % size of input text:10
Analysing required size...

Imagine computers learning from medical records which treatments are most effective for new
diseases, houses learning from experience to optimize energy costs based on the
particular usage patterns of their occupants, or personal software assistants learning the evolving interests of their users in order to highlight especially relevant
stories from the online morning newspaper

In
the field known as data mining, machine learning algorithms are being used routinely to discover valuable knowledge from large commercial databases containing
equipment maintenance records, loan applications, financial transactions, medical
records, and the like

This book presents the field of machine learning, describing a variety of
learning paradigms, algorithms, theoretical results, and applications

Size of output(Line number) 3
.....
```

Figure 8: 10% summarized text of dataset-2

6.3 Input dataset-3 [1]

6.3.1 50% summarized text of Dataset 3

12

```
IPython console
Console 1/A x
Input Size (Line number): 41
Enter the % size of input text:50
Analysing required size...
Data mining is a field which has seen rapid advances in recent years because of the immense advances in hardware and software technology which has lead to the availability of different kinds of data
This is particularly true for the case of text data, where the development of hardware and software platforms for the web and social networks has enabled the rapid creation of large repositories of different kinds of data
The increasing amounts of text data available from different applications has created a need for advances in algorithmic design which can learn interesting patterns from the data in a dynamic and scalable way
While structured data is generally managed with a database system, text data is typically managed via a search engine due to the lack of structures [5]
A search engine enables a user to find useful information from a collection conveniently with a keyword query, and how to improve the effectiveness and efficiency of a search engine has been a central research topic in the field of information retrieval [13, 3], where many related topics to search such as text clustering, text categorization, summarization, and recommender systems are also studied [12, 9, 7]
However, research in information retrieval has traditionally focused more on facilitating information access [13] rather than analyzing information to discover patterns, which is the primary goal of text mining
There are also many applications of text mining where the primary goal is to analyze and discover any interesting patterns, including trends and outliers, in text data, and the notion of a query is not essential or even relevant
Technically, mining techniques focus on the primary models, algorithms and applications about what one can learn from different kinds of text data
Some examples of such questions are as follows:
What are the primary supervised and unsupervised models for learning from text data? How are traditional clustering and classification problems different for text data, as compared to the traditional database literature? What are the useful tools and techniques used for mining text data? Which are the useful mathematical techniques which one should know, and which are repeatedly used in the context of different kinds of text data? What are the key application domains in which such mining techniques are used, and how are they effectively applied?
A number of key characteristics distinguish text data from other forms of data such as relational or quantitative data
In such cases, the methods for reduction should be specifically designed while taking this characteristic of text data into account
The variation in word frequencies and document lengths also lead to a number of issues involving document representation and normalization, which are critical for text mining
Furthermore, text data can be analyzed at different levels of representation
For example, text data can easily be treated as a bag-of-words, or it can be treated as a string of words
However, in most applications, it would be desirable to represent text information semantically so that more meaningful analysis and mining can be done
For exam
ple, representing text data at the level of named entities such as people, organizations, and locations, and their relations may enable discovery of more interesting patterns than representing text as a bag of words
Thus most text
mining approaches currently still rely on the more shallow word-based representations, especially the bag-of-wrods approach, which, while los
ing the positioning information in the words, is generally much simpler to deal with from an algorithmic point of view than the string-based approach
, extraction of knowledge from the Web), natural lan
guage processing techniques, especially information extraction, are also playing an important role in obtaining a semantically more meaningful representation of text
Recently, there has been rapid growth of text data in the context of different web-based applications such as social media, which often occur in the context of multimedia or other heterogeneous data domains
Therefore, a number of techniques have recently been designed for the joint mining of text data in the context of these different kinds of data domains
For example, the Web contains text and image data which are often intimately connected to each other and these links can be used to improve the learning process from one domain to another
The next section will discuss the different kinds of algorithms and applications for text mining
Size of output(Line number) 21
```

Figure 9: 50% summarized text of dataset-3

6.3.2 30% summarized text of Dataset 3

```
IPython console
Console 1/A
Input Size (Line number): 41
Enter the % size of input text:30
Analysing required size...
This is particularly true for the case of text data, where the development of hardware and software platforms for the web and social networks has enabled the rapid creation of large repositories of different kinds of data
The increasing amounts of text data available from different applications has created a need for advances in algorithmic design which can learn interesting patterns from the data in a dynamic and scalable way
A search engine enables a user to find useful information from a collection conveniently with a keyword query, and how to improve the effectiveness and efficiency of a search engine has been a central research topic in the field of information retrieval [13, 3], where many related topics to search such as text clustering, text categorization, summarization, and recommender systems are also studied [12, 9, 7]
However, research in information retrieval has traditionally focused more on facilitating information access [13] rather than analyzing information to discover patterns, which is the primary goal of text mining
There are also many applications of text mining where the primary goal is to analyze and discover any interesting patterns, including trends and outliers, in text data, and the notion of a query is not essential or even relevant
Technically, mining techniques focus on the primary models, algorithms and applications about what one can learn from different kinds of text data
Some examples of such questions are as follows:
What are the primary supervised and unsupervised models for learning from text data? How are traditional clustering and classification problems different for text data, as compared to the traditional database literature? What are the useful tools and techniques used for mining text data? Which are the useful mathematical techniques which one should know, and which are repeatedly used in the context of different kinds of text data? What are the key application domains in which such mining techniques are used, and how are they effectively applied?
A number of key characteristics distinguish text data from other forms of data such as relational or quantitative data
Thus most text mining approaches currently still rely on the more shallow word-based representations, especially the bag-of-words approach, which, while losing the positioning information in the words, is generally much simpler to deal with from an algorithmic point of view than the string-based approach
, extraction of knowledge from the Web), natural language processing techniques, especially information extraction, are also playing an important role in obtaining a semantically more meaningful representation of text
Recently, there has been rapid growth of text data in the context of different web-based applications such as social media, which often occur in the context of multimedia or other heterogeneous data domains
Therefore, a number of techniques have recently been designed for the joint mining of text data in the context of these different kinds of data domains
For example, the Web contains text and image data which are often intimately connected to each other and these links can be used to improve the learning process from one domain to another
Size of output(Line number) 12
```

Figure 10: 30% summarized text of dataset-3

6.3.3 10% summarized text of Dataset 3

Input Size (Line number): 41

Enter the % size of input text:10

Analysing required size...

There are also many applications of text mining where the primary goal is to analyze and discover any interesting patterns, including trends and outliers, in text data, and the notion of a query is not essential or even relevant

Technically, mining techniques focus on the primary models, algorithms and applications about what one can learn from different kinds of text data

Some examples of such questions are as follows:

What are the primary supervised and unsupervised models for learning from text data? How are traditional clustering and classification problems different for text data, as compared to the traditional database literature? What are the useful tools and techniques used for mining text

data? Which are the useful mathematical techniques which one should know, and which are repeatedly used in the context of different kinds of text data? What are the key application domains in which such mining techniques are used, and how are they effectively applied?

A number of key characteristics distinguish text data from other forms of data such as relational or quantitative data

Therefore, a number of techniques have recently been designed for the joint mining of text data in the context of these different kinds of data domains

Size of output(Line number) 4

Figure 11: 10% summarized text of dataset-3

Chapter 7

Conclusion and Future Scope

As depicted in the above figures in the result section, the summarization technique adopted gives a fair output in terms of textual reduction ratio. The present technique is mainly based on extractive procedures involving no semantic analysis. Yet the quality of the summary remains competitively adequate. The procedure is simple, which is a plus point of the whole effort. The term frequency can be improved by augmenting it with inverse document frequency measures to score the selected sentences in a more comprehensive manner.

The researchers hope to better their effort by applying Abstractive Summarization techniques in their future endeavour.

Bibliography

1. Ani Nenkova, Kathleen McKeown, “A SURVEY OF TEXT SUMMARIZATION TECHNIQUES”, chapter 3 of *Mining Text Data*, ed Charu C. Aggarwal, ChengXiang Zhai, 2012.
2. Karuturi Sneha, Malle Gowda M, “Research on software Testing Techniques and Software Automation Testing Tools”, ICECDS, 2017.
3. BAOCHUN LI et al, “Two Decades of Internet Video Streaming: A Retrospective View”, *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 9, No. 1s, Article 33, 2013.
4. <https://www.guru99.com/automation-testing.html>
5. <https://www.dacast.com/blog/hls-streaming-protocol/>
6. <https://www.brand.live/assets/documents/2018-Live-Video-Benchmark-Report.pdf>
7. <http://iqa.ece.toronto.edu/papers/bli-tomccap13.pdf>
8. J. Han, M. Kamber, J. Pei, “Data Mining Concepts and Techniques”, Third Edition, Elsevier Inc., 2012.
9. T. M. Mitchell, “Machine Learning”, McGraw-Hill, 1997.

Annexure

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Fri Jan 4 14:21:17 2019

@author: Jadavpur University
"""

from nltk.corpus import stopwords
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.stem import WordNetLemmatizer
import nltk
import operator
lem=WordNetLemmatizer()
f=open('corpusMTD.txt','r')
data=f.read()
splitData = data.split('.')
for index in splitData:
    print('///', index, '///')
st=set(stopwords.words('english'))
stopwords = nltk.corpus.stopwords.words('english')
newStopWords = ['n','t','\','s','â€','"', '*'] # additional stop words
stopwords.extend(newStopWords)
s=sent_tokenize(data)
word={}
for i in s:
    wd=word_tokenize(i)
```

```

tagged=nltk.pos_tag(wd)
for j in tagged:
    if (j[1][0]=='N' or j[1][0]=='R' or j[1][0]=='V' or j[1][0]=='J') and j[0] not in
stopwords:
    if j[1][0]=='v':
        #ROOT word
        k=lem.lemmatize(j[0],pos='v')
    else:
        k=j[0]
    if k not in word:
        word[k]=0
    word[k]+=1
a=sorted(word.items(),key=operator.itemgetter(1),reverse=True)
data = ([["default" , 0]])
maxScore = 0
index = 0
for line in splitData:
    score = 0
    for aElement in a:
        index += 1
        if aElement[0] in line:
            score += aElement[1]
        if index == len(a):
            break
    newData = [line, score]
    if maxScore < score:
        maxScore = score
    if score >= 0 and ((2 * len(line)) - len(line.lstrip(' ')) - len(line.lstrip('\n'))) <= 10 :
        data.append(newData)
print("Input Size (Line number): ", len(data) - 1)
size=int(input('Enter the % size of input text:'))
numberOfLines = ( len(data) * size ) / 100
print("Analysing required size...")

```

```

listOfScore = [0]
for collection in data:
    listOfScore.append(collection[1])
listOfScore.sort(reverse = True)
thresholdScore = listOfScore[int(numberOfLines)]
outputSize = 0
for collection in data:
    if collection[1] > thresholdScore:
        outputSize += 1
        print("_____")
        print(collection[0])
print("_____")
print("Size of output(Line number)", outputSize)
print("=====[ END ]=====")

```