Research Journal of Computer and Information Technology Sciences _____________________E-ISSN 2320 – 6527
Vol. **4(11),** 16-20, November **(2016)** *Res. J. Computer and IT Sci.*

Review Paper

Machine Learning Approach upon Text from Varied Publishing Formats

Arpana Rawal[1*], Ani Thomas[2] and Saurabh Bhagvatula[1]
[1]Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg, CG, India
[2]Department of Information Technology, Bhilai Institute of Technology, Durg, CG, India
arpana.rawal@gmail.com

Available online at: www.isca.in, www.isca.me
Received 24th February 2016, revised 5th November 2016, accepted 15th November 2016

Abstract

The paper aims toward reporting the approach to challenges for conversion of documents from varied publishing formats to machine readable formats. This research objective falls in the field of information retrieval. In order to represent the documents available in machine readable format from different publishing format, there is a need to identify, access, process and finally represent the information in such a manner which makes it ready for easy machine access. The above mentioned task involves different types of challenges as discussed in detail below. These challenges are mentioned with approach as proposed to solve information retrieval task in huge text corpora.

Keywords: Document Section Extraction, Machine Readable Document, Text Extraction.

Introduction

In order to publish a document, every publisher follows one or the other publishing format for structured representation and easy access of information available in the document. There is no limit to the formats used by publishers while publishing a document. Some of which are described below:

Different level of sections: The front index of two different documents as shown in Figure-1 and Figure-2, have different level of sections. Also, Figure-1 has different level of sections within the document as the section 1.3 has no sub-sections but 1.4 has two sub-levels.

Incomplete Front Index: As shown in Figure-2 the front index is incomplete, it only shows highest level section in front index.

Varied way of section indexing: The section indexing is done in different ways as shown in Figure-1, top section is X.Y, its sub-section is X.Y.Z and its sub-sections are not indexed.

State-of-Art in Text Pre-Processing

In the field of text pre-processing, the efforts are put forward in series of research. There are proposed methodologies existing for conversion of publishing document (PDF) into editable machine readable format.

As explained by Dr. Arpana Rawal, a methodology for conversion of PDF document into machine readable format with proper tagging to each separate block of information in the corpus. This tagging process, tags the paragraphs of the corpus in a format described below:

<chapter_number.section_number / paragraph_number. page_ numbers>

Also, it extracts the front index, grown front index and back index are stores it in a file for narrowing the search space for accessing the information for search query.

Tamir Hassan explained about converting a PDF document into the HTML along with figures and images in their original style as present in the PDF, but is working only to the PDF version 1.4[1].

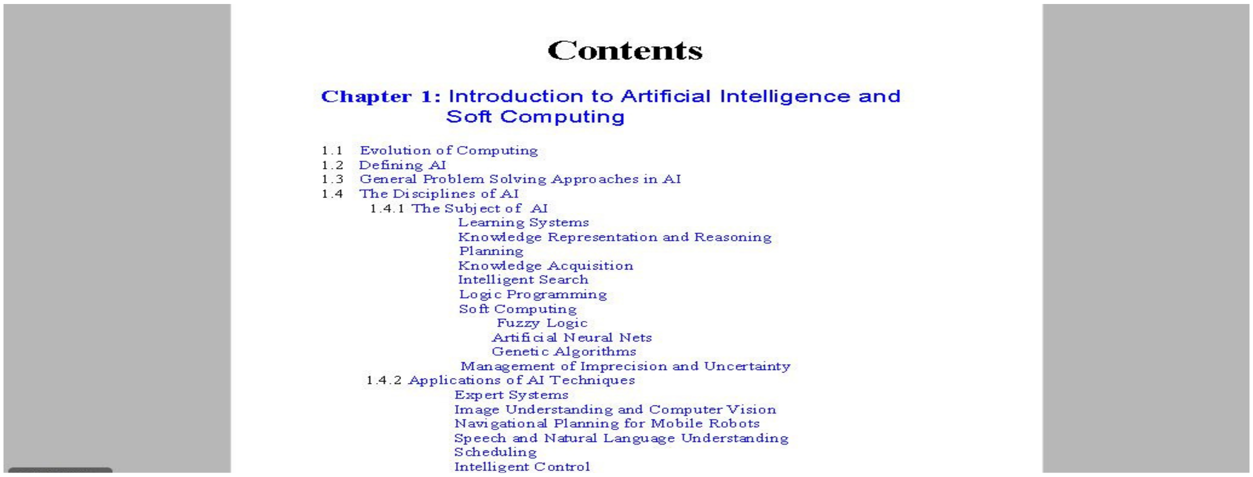
Pre-Processing Methodology (Proposed)

There are a series of steps involved in pre-processing the text of the publishing documents available in Portable Document Format (PDF). In order to execute this task, a series of process are required to be performed:

Content Extraction from PDF: In order to represent document into editable machine readable format, there is a need for extraction of content available in PDF format for processing. For the above mentioned purpose PDF text extracting tools like iText, PDF Box are utilized. The Figure-4 shows the entire text extracted from PDF, losing its original style and generating .html file of the extracting information as paragraphs.

Extracting the document Structure: Since there are numerous formats for the publishing documents, it is required to extract the sections hierarchy of the document and creation of section structure of .xml of the sections and storing them in .xml format before the creation of machine readable format for entire document, which is shown in Figure-5.

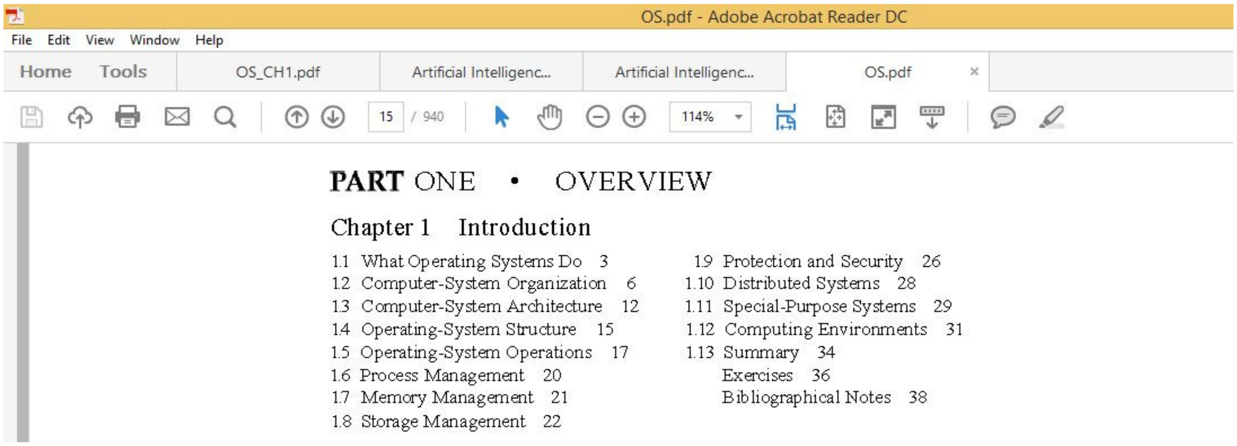
Generation of complete .xml document: After creation of .xml document containing the sections hierarchy of the document, it is required to add the content in each section and creation of final editable machine readable format document. In this the extracted text content, shown in Figure-4 is associated with the sections of document as shown in Figure-5. The Figure-6 gives the .xml with the text associated in appropriate section through inevitable human intervention.



Contents

Chapter 1: Introduction to Artificial Intelligence and Soft Computing

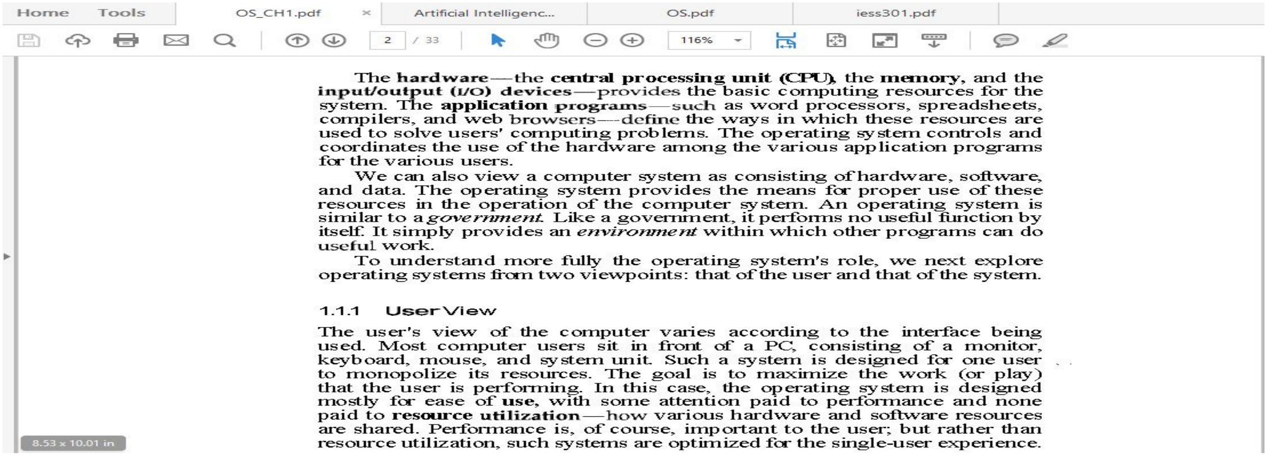
Figure-1
Three level of heading in the front index[2]


OS.pdf - Adobe Acrobat Reader DC
File Edit View Window Help
Home Tools
OS_CH1.pdf
Artificial Intelligenc...
Artificial Intelligenc...
OS.pdf
15 / 940
114%

PART ONE • OVERVIEW

Chapter 1 Introduction

Figure-2
One level of heading in the front index[3]



The **hardware**—the **central processing unit (CPU)**, the **memory**, and the **input/output (I/O) devices**—provides the basic computing resources for the system. The **application programs**—such as word processors, spreadsheets, compilers, and web browsers—define the ways in which these resources are used to solve users' computing problems. The operating system controls and coordinates the use of the hardware among the various application programs for the various users.

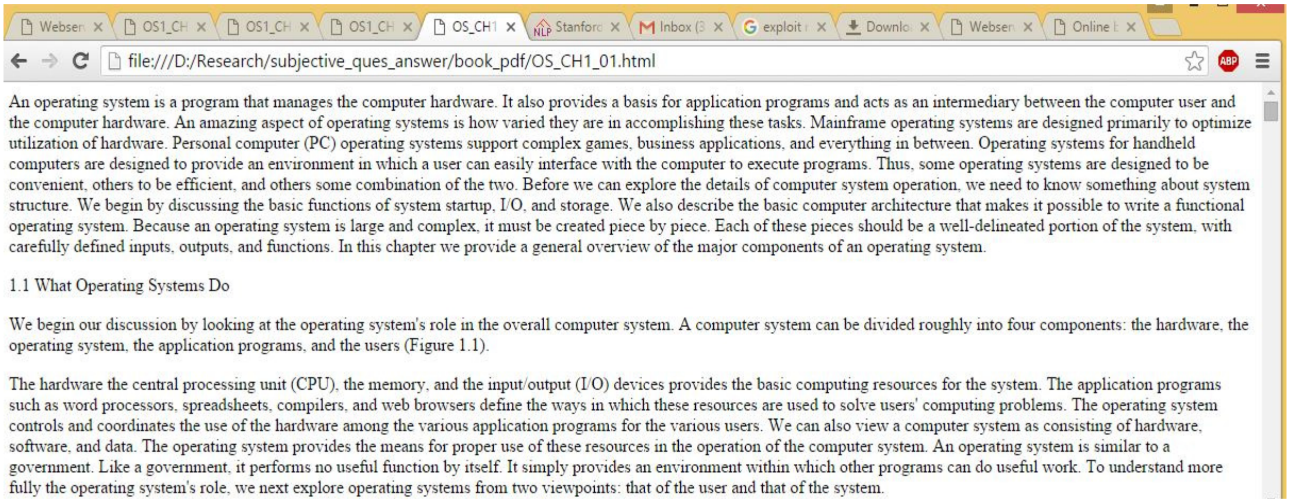
We can also view a computer system as consisting of hardware, software, and data. The operating system provides the means for proper use of these resources in the operation of the computer system. An operating system is similar to a *government*. Like a government, it performs no useful function by itself. It simply provides an *environment* within which other programs can do useful work.

To understand more fully the operating system's role, we next explore operating systems from two viewpoints: that of the user and that of the system.

1.1.1 User View

The user's view of the computer varies according to the interface being used. Most computer users sit in front of a PC, consisting of a monitor, keyboard, mouse, and system unit. Such a system is designed for one user to monopolize its resources. The goal is to maximize the work (or play) that the user is performing. In this case, the operating system is designed mostly for ease of **use,** with some attention paid to performance and none paid to **resource utilization**—how various hardware and software resources are shared. Performance is, of course, important to the user; but rather than resource utilization, such systems are optimized for the single-user experience.

Figure-3
A snapshot of PDF document used for conversion[3]



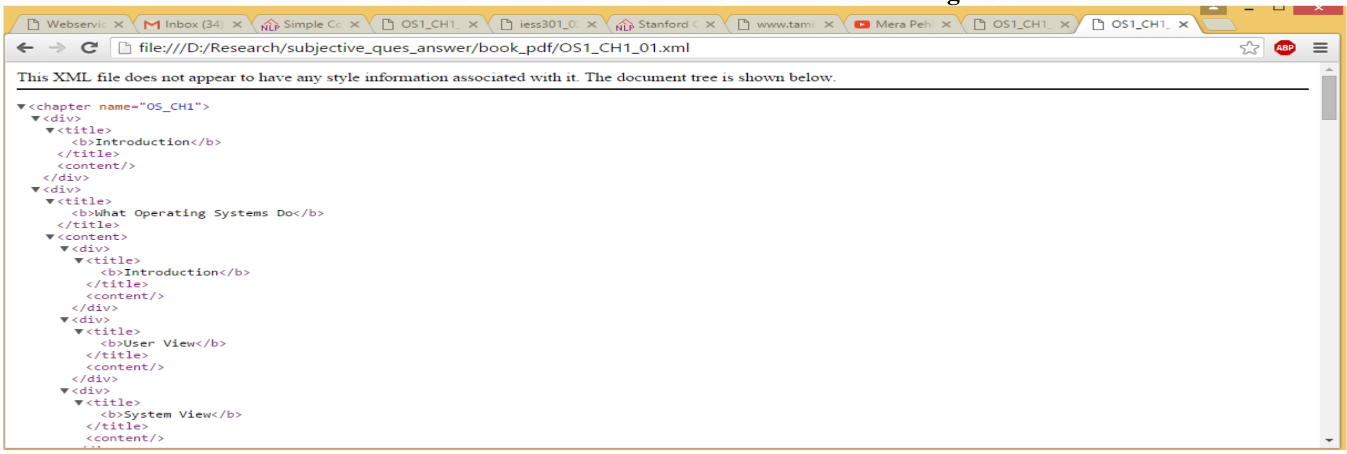
An operating system is a program that manages the computer hardware. It also provides a basis for application programs and acts as an intermediary between the computer user and the computer hardware. An amazing aspect of operating systems is how varied they are in accomplishing these tasks. Mainframe operating systems are designed primarily to optimize utilization of hardware. Personal computer (PC) operating systems support complex games, business applications, and everything in between. Operating systems for handheld computers are designed to provide an environment in which a user can easily interface with the computer to execute programs. Thus, some operating systems are designed to be convenient, others to be efficient, and others some combination of the two. Before we can explore the details of computer system operation, we need to know something about system structure. We begin by discussing the basic functions of system startup, I/O, and storage. We also describe the basic computer architecture that makes it possible to write a functional operating system. Because an operating system is large and complex, it must be created piece by piece. Each of these pieces should be a well-delineated portion of the system, with carefully defined inputs, outputs, and functions. In this chapter we provide a general overview of the major components of an operating system.

1.1 What Operating Systems Do

We begin our discussion by looking at the operating system's role in the overall computer system. A computer system can be divided roughly into four components: the hardware, the operating system, the application programs, and the users (Figure 1.1).

The hardware the central processing unit (CPU), the memory, and the input/output (I/O) devices provides the basic computing resources for the system. The application programs such as word processors, spreadsheets, compilers, and web browsers define the ways in which these resources are used to solve users' computing problems. The operating system controls and coordinates the use of the hardware among the various application programs for the various users. We can also view a computer system as consisting of hardware, software, and data. The operating system provides the means for proper use of these resources in the operation of the computer system. An operating system is similar to a government. Like a government, it performs no useful function by itself. It simply provides an environment within which other programs can do useful work. To understand more fully the operating system's role, we next explore operating systems from two viewpoints: that of the user and that of the system.

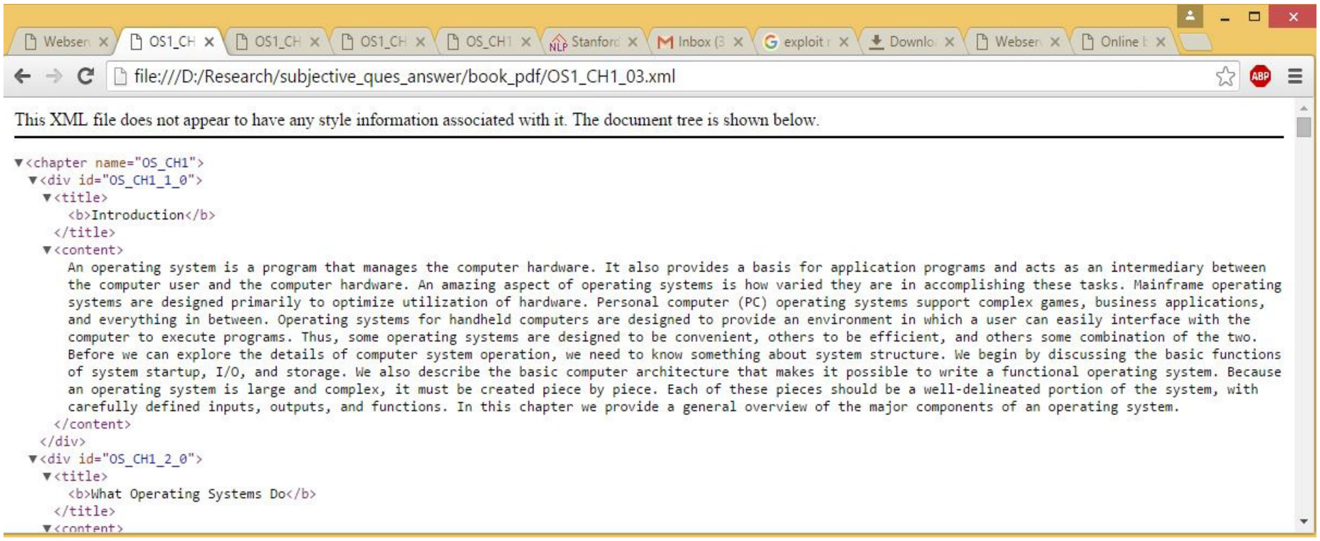
Figure-4
html document of text extracted from PDF shown in Figure-3



This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<chapter name="OS_CH1">
  ▼<div>
    ▼<title>
       <b>Introduction</b>
     </title>
     <content/>
   </div>
  ▼<div>
    ▼<title>
       <b>What Operating Systems Do</b>
     </title>
    ▼<content>
      ▼<div>
        ▼<title>
           <b>Introduction</b>
         </title>
         <content/>
       </div>
      ▼<div>
        ▼<title>
           <b>User View</b>
         </title>
         <content/>
       </div>
      ▼<div>
        ▼<title>
           <b>System View</b>
         </title>
         <content/>
```

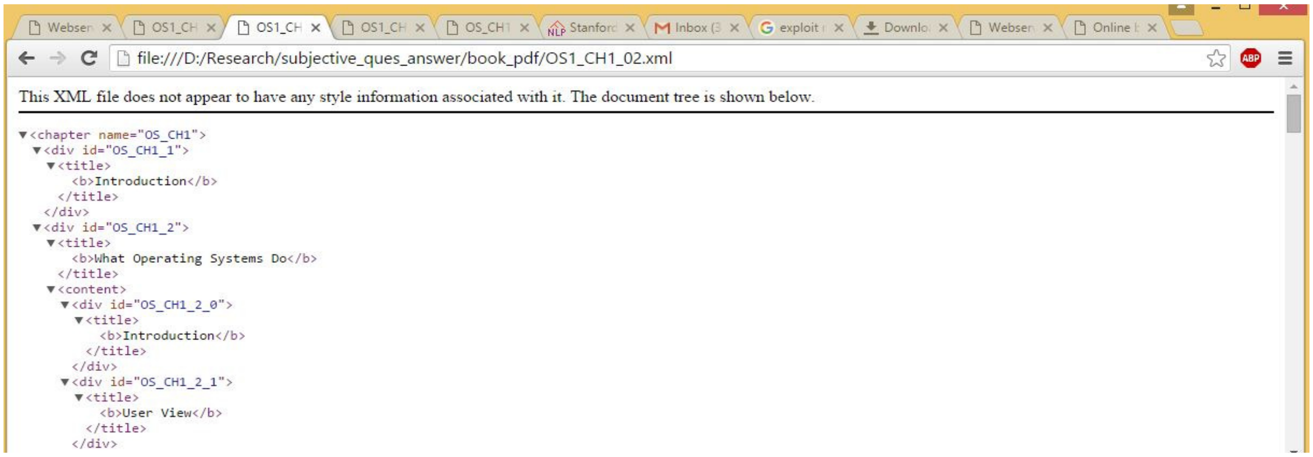
Figure-5
The .xml for all levels of section/ sub-sections of document shown in Figure-3



This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<chapter name="OS_CH1">
  ▼<div id="OS_CH1_1_0">
    ▼<title>
       <b>Introduction</b>
     </title>
    ▼<content>
       An operating system is a program that manages the computer hardware. It also provides a basis for application programs and acts as an intermediary between
       the computer user and the computer hardware. An amazing aspect of operating systems is how varied they are in accomplishing these tasks. Mainframe operating
       systems are designed primarily to optimize utilization of hardware. Personal computer (PC) operating systems support complex games, business applications,
       and everything in between. Operating systems for handheld computers are designed to provide an environment in which a user can easily interface with the
       computer to execute programs. Thus, some operating systems are designed to be convenient, others to be efficient, and others some combination of the two.
       Before we can explore the details of computer system operation, we need to know something about system structure. We begin by discussing the basic functions
       of system startup, I/O, and storage. We also describe the basic computer architecture that makes it possible to write a functional operating system. Because
       an operating system is large and complex, it must be created piece by piece. Each of these pieces should be a well-delineated portion of the system, with
       carefully defined inputs, outputs, and functions. In this chapter we provide a general overview of the major components of an operating system.
     </content>
   </div>
  ▼<div id="OS_CH1_2_0">
    ▼<title>
       <b>What Operating Systems Do</b>
     </title>
    ▼<content>
```

Figure-6
.xml with text associated with contents of document shown in Figure-3



This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<chapter name="OS_CH1">
  ▼<div id="OS_CH1_1">
    ▼<title>
        <b>Introduction</b>
      </title>
    </div>
  ▼<div id="OS_CH1_2">
    ▼<title>
        <b>What Operating Systems Do</b>
      </title>
    ▼<content>
      ▼<div id="OS_CH1_2_0">
        ▼<title>
            <b>Introduction</b>
          </title>
        </div>
      ▼<div id="OS_CH1_2_1">
        ▼<title>
            <b>User View</b>
          </title>
        </div>
```

Figure-7
The front index having all levels of section / sub-section hierarchy for front index shown in Figure-2

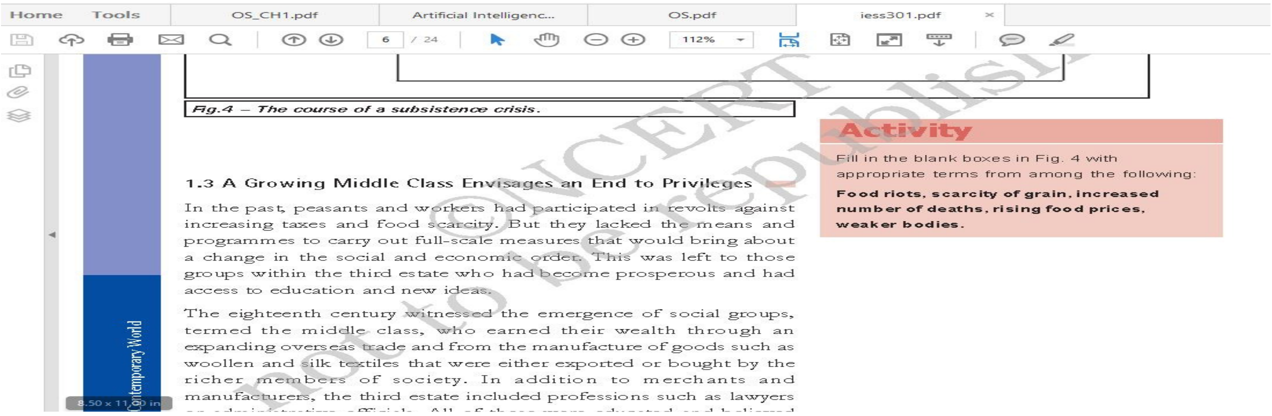


Fig.4 – The course of a subsistence crisis.

1.3 A Growing Middle Class Envisages an End to Privileges

In the past, peasants and workers had participated in revolts against increasing taxes and food scarcity. But they lacked the means and programmes to carry out full-scale measures that would bring about a change in the social and economic order. This was left to those groups within the third estate who had become prosperous and had access to education and new ideas.

The eighteenth century witnessed the emergence of social groups, termed the middle class, who earned their wealth through an expanding overseas trade and from the manufacture of goods such as woollen and silk textiles that were either exported or bought by the richer members of society. In addition to merchants and manufacturers, the third estate included professions such as lawyers

Activity

Fill in the blank boxes in Fig. 4 with appropriate terms from among the following:

Food riots, scarcity of grain, increased number of deaths, rising food prices, weaker bodies.

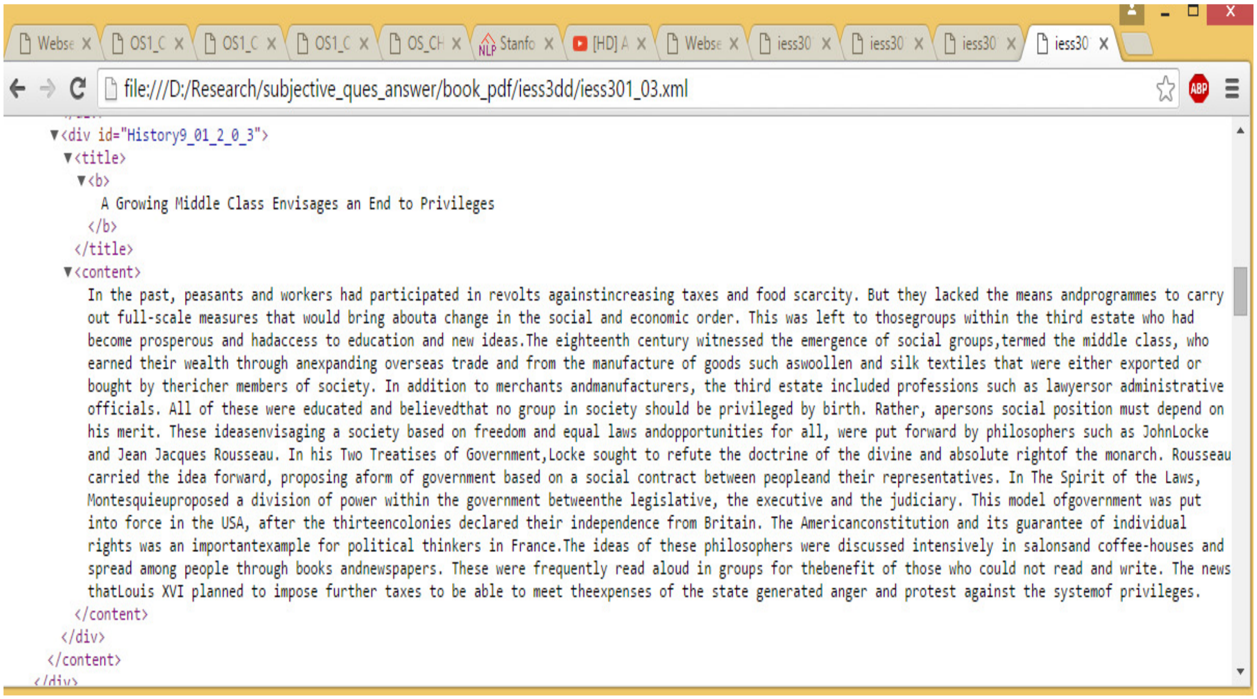
Figure-8
A snapshot of a PDF document having watermark[4]

Considerable Issues

After performing the above mentioned methodology over numerous documents, it was observed that there are few challanges which seem barrier for converting the PDF document into a machine readable format. These barriers with the approach used to tackle them are mentioned below: i. The presence of watermark in the document. Figure-8 shows a document with the watermark present in it, is required to be removed for getting the desired .xml file, which is done through human intervention over the text extracted by PDFBox and the desired .xml for the PDF is generated is shown in Figure-9. ii. The difficulty in recognizing the chapter demarcations is resolved by separating PDF file of the whole text corpus (for current reference, and e-book) into respective number of PDF corpora for all the chapters of the e-book. This trimming step can be implemented as a ready-to-use tool by taking page range of each chapter. iii. The presence of different font text, lack of line separator between section title and content, page headers coming as part of text are dealt with overriding the methods of PDFBox library. iv. The different hierarchical levels of sections in document are dealt with creation of document section's machine readable document and then inserting the text to each section by inevitable stages of human intervention. As described in Figure-5 which is all levels of section/ sub-section hierarchy in .xml format and the Figure-6, .xml after text associated with all section/ sub-section. v. With the baseline text corpus for all chapters which is now available in .xml format, the front index given at the beginning of document, now can be grown into front index covering all levels of section / sub-section hierarchies. This can be generated automatically by removing textual para-phrased content and tagging each section with its indexes. This is illustrated in Figure-7, where the all level of section / sub-section hierarchy is generated in .xml format for the front index shown in Figure-2.

Conclusion

The generation of machine readable document from the PDF can assist in variety of purposes related to text mining. It gives the utility to extract the searched information and application of processing tools over the content of a PDF document. The document converted in .xml format so it is also possible to make it portable over the network or access the information of the over the network.



file:///D:/Research/subjective_ques_answer/book_pdf/iess3dd/iess301_03.xml

```
▼<div id="History9_01_2_0_3">
  ▼<title>
    ▼<b>
       A Growing Middle Class Envisages an End to Privileges
     </b>
   </title>
  ▼<content>
     In the past, peasants and workers had participated in revolts againstincreasing taxes and food scarcity. But they lacked the means andprogrammes to carry
     out full-scale measures that would bring abouta change in the social and economic order. This was left to thosegroups within the third estate who had
     become prosperous and hadaccess to education and new ideas.The eighteenth century witnessed the emergence of social groups,termed the middle class, who
     earned their wealth through anexpanding overseas trade and from the manufacture of goods such aswoollen and silk textiles that were either exported or
     bought by thericher members of society. In addition to merchants andmanufacturers, the third estate included professions such as lawyersor administrative
     officials. All of these were educated and believedthat no group in society should be privileged by birth. Rather, apersons social position must depend on
     his merit. These ideasenvisaging a society based on freedom and equal laws andopportunities for all, were put forward by philosophers such as JohnLocke
     and Jean Jacques Rousseau. In his Two Treatises of Government,Locke sought to refute the doctrine of the divine and absolute rightof the monarch. Rousseau
     carried the idea forward, proposing aform of government based on a social contract between peopleand their representatives. In The Spirit of the Laws,
     Montesquieuproposed a division of power within the government betweenthe legislative, the executive and the judiciary. This model ofgovernment was put
     into force in the USA, after the thirteencolonies declared their independence from Britain. The Americanconstitution and its guarantee of individual
     rights was an importantexample for political thinkers in France.The ideas of these philosophers were discussed intensively in salonsand coffee-houses and
     spread among people through books andnewspapers. These were frequently read aloud in groups for thebenefit of those who could not read and write. The news
     thatLouis XVI planned to impose further taxes to be able to meet theexpenses of the state generated anger and protest against the systemof privileges.
   </content>
 </div>
</content>
</div>
```

Figure-9
A snapshot of .xml document for document in Figure-7

References

1. Hassan Tamir (2003). A PDF TO HTML conversion. Third Year Project, University of Warwick, Coventry, West Midlands, U.K.
2. Konar Amit (2000). Artificial Intelligence and Soft Computing: Behavioral and Cognitive Modeling of Human Brain. CRC press, USA, ISBN:0-8493-1385-6.
3. Silberschatz Abharam, Galvin Peter Baer and Gagne Greg (2005). Operating System and Concepts. 7th Edition, John Wiley and Sons. Inc., USA, p:xvii, ISBN:0-471-69466-5
4. NCERT (2014) India and the Contemporary World-I: A Textbook in History for class IX. Nation Council of Educational Research and Training, India, 6, ISBN:8-7450-536-9.