Reading/Extracting data from PDF file

Agenda of this document: To discuss about the approach of how to read/extract data from PDF files.

In selenium we need to fetch and validate data from different format like Properties file, Excel file, text files, Csv files, PDF files.

To get the data from a PDF file we need to use couple of jar files i.e:

1. Pdfbox jar
2. Fontbox jar
3. Commons-logging jar

After extracting/reading the data from PDF file we need to store the data in string and can perform the desired actions like validations and all.

Here is the code snippet which performs the extraction/reading of the data from PDF file:

```
package com.dataManager;

import java.io.File;
import java.io.FileInputStream;

import org.apache.pdfbox.cos.COSDocument;
import org.apache.pdfbox.pdfparser.PDFParser;
import org.apache.pdfbox.pdmodel.PDDocument;
import org.apache.pdfbox.util.PDFTextStripper;

public class PDFReader {

	public static void main(String[] args){

		PDFParser parser = null;
		PDDocument pdDocument = null;
		COSDocument cosDocument = null;
		PDFTextStripper textStripper;
```

```
                String parsedText;
                String fileName = "..\\AdvanceConcepts\\src\\DataStorage\\blog.PDF";
                File file = new File(fileName);
                try {
                    parser = new PDFParser(new FileInputStream(file));
                    parser.parse();
                    cosDocument = parser.getDocument();
                    textStripper = new PDFTextStripper();
                    pdDocument = new PDDocument(cosDocument);
                    parsedText = textStripper.getText(pdDocument);
                    System.out.println(parsedText.replaceAll("[^A-Za-z0-9. ]+", ""));
                } catch (Exception e) {
                    e.printStackTrace();
                    try {
                        if (cosDocument != null)
                            cosDocument.close();
                        if (pdDocument != null)
                            pdDocument.close();
                    } catch (Exception ex) {
                        ex.printStackTrace();
                    }

                }
            }
}
```