**Web Scraping and PDF Reading in Java**

December 03, 2023

**Task 1**

import org.jsoup.Jsoup;

import org.jsoup.nodes.Document;

import org.jsoup.nodes.Element;

import org.jsoup.select.Elements;

import java.io.IOException;

import java.util.Scanner;

public class QuoteScraper {

public static void main(String[] args) {

Scanner scanner = new Scanner(System.in);

System.out.print("Enter the page number (1-10): ");

int pageNumber = scanner.nextInt();

if (pageNumber < 1 || pageNumber > 10) {

System.out.println("Invalid page number. Please enter a number between 1 and 10.");

return;

}

String url = "https://quotes.toscrape.com/page/" + pageNumber;

try {

Document document = Jsoup.connect(url).get();

Elements quotes = document.select(".quote .text");

System.out.println("Quotes from Page " + pageNumber + ":");

for (Element quote : quotes) {

System.out.println("- " + quote.text());

}

} catch (IOException e) {

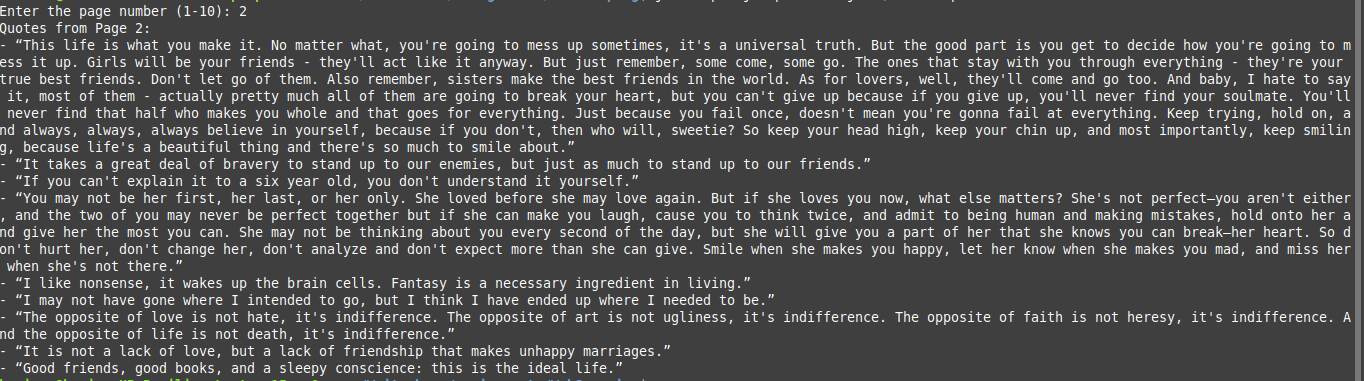
e.printStackTrace();

}

}

}

**Screenshot task 1:**



**Task 2:**

import org.apache.pdfbox.pdmodel.PDDocument;

import org.apache.pdfbox.text.PDFTextStripper;

import org.apache.pdfbox.Loader;

import java.io.File;

import java.io.FileWriter;

import java.io.IOException;

import java.io.Writer;

public class PDFReader {

public static void main(String[] args) {

try {

String file = "USCensus.pdf";

PDDocument document = Loader.loadPDF(new File(file));

PDFTextStripper pdfStripper = new PDFTextStripper();

String text = pdfStripper.getText(document);

String myTxt = "\_USCensus.txt";

Writer writer = new FileWriter(myTxt);

pdfStripper.writeText(document, writer);

System.out.println("Text extracted from PDF:\n" + text);

writer.close();

document.close();

} catch (IOException e) {

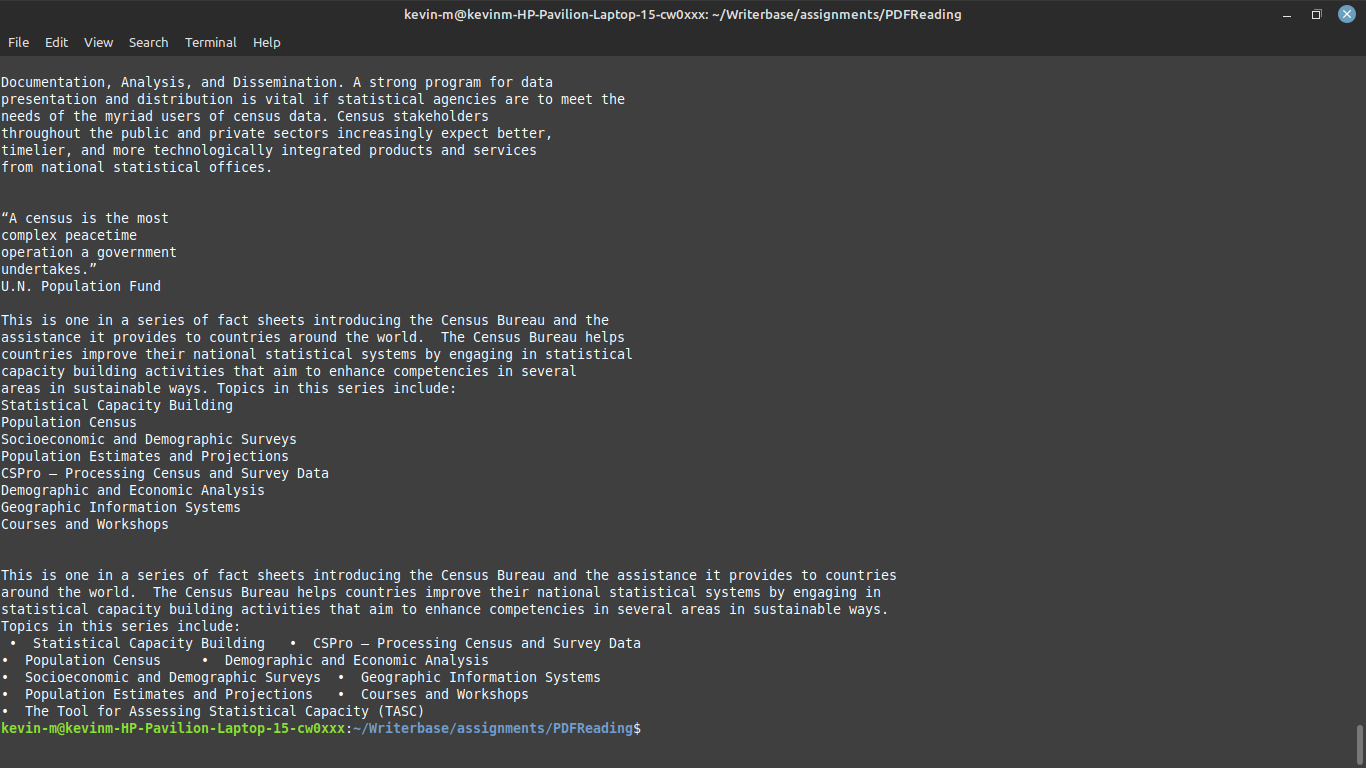
e.printStackTrace();

}

}

}

**Screenshot task 2:**



**Notes**

### ensure Java is installed, then download jsoup, and pdfbox libraries.

Copy the programs and save them in .java files.

Navigate to the location where the java files have been stored:

## To Run the WebScraping program:

Run on your terminal:

'javac -cp .:jsoup-1.17.1.jar QuoteScraper.java' then

'java -cp .:jsoup-1.17.1.jar QuoteScraper'

# This will run the java file and generate a some quote data

## To Run the PDFReader program:

Run on your terminal:

'javac -cp .:pdfbox-app-3.0.0.jar PDFReader.java'

'java -cp .:pdfbox-app-3.0.0.jar PDFReader'

**References**

baeldung. (2023, February 22). *Reading PDF File Using Java* (baeldung, Ed.)

[Review of *Reading PDF File Using Java*]. Baeldung.com; baeldung.com.

https://www.baeldung.com/java-pdf-file-read#using-apache-pdfbox

*com.itextpdf.text.pdf.PdfReader java code examples | Tabnine*. (n.d.).

Www.tabnine.com. Retrieved December 1, 2023, from

https://www.tabnine.com/code/java/classes/com.itextpdf.text.pdf.PdfReader

‌‌Mykong. (2017, July 24). PDFBox – How to read PDF file in Java (Mykong, Ed.)

[Review of PDFBox – How to read PDF file in Java]. Mkyong.com.

https://mkyong.com/java/pdfbox-how-to-read-pdf-file-in-java/

*PDFBox - Reading Text*. (n.d.). Www.tutorialspoint.com.

https://www.tutorialspoint.com/pdfbox/pdfbox\_reading\_text.htm

*Web Scraping with Java in 2023 | Nanonets*. (2023, March 16). Nanonets Intelligent

Automation, and Business Process AI Blog.

https://nanonets.com/blog/java-web-scraping-tutorial/

ZenBrows. (2022, December 19). Web Scraping in Java in 2023: The Complete Guide

[Review of Web Scraping in Java in 2023: The Complete Guide]. ZenRows

Blog. https://www.zenrows.com/blog/web-scraping-java